8-9-2014

A Multiple Imputation Approach For Semiparametric Cure Model With Interval Censored Data

Jie Zhou
University of South Carolina - Columbia

A Multiple Imputation Approach For Semiparametric Cure Model With Interval Censored Data

by

Jie Zhou

Bachelor of Science
East China Normal University 2012

Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Public Health in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2014

Accepted by:

Jiajia Zhang, Director of Thesis

Bo Cai, Reader

Alexander C. McLain, Reader

Lacy Ford, Vice Provost and Dean of Graduate Studies

Acknowledgments

Many thanks to my thesis advisor, Dr. Jiajia Zhang, for her continual support and advice throughout this work. Without her, this project could not have been completed. She has been a tremendous mentor for me. I would like to thank her for encouraging my research and for allowing me to grow as a research scientist. Her advice on both research as well as on my career have been priceless.

I would also like to thank Dr. Cai Bo and Dr. Alex McLain for serving as my committee members even at hardship. I would also like to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, appreciations for you.

My thanks also go to my parients, without their support and love, I will not have courage to complete my study in United States. I also thank my fiance, Weinan Xu, for supporting me all the time.

Abstract

Interval censored survival data, where the exact event time is only known to lie in an observed time interval, is commonly encountered in practice. Such data analysis may be conducted under the setting where a fraction of patients can be considered as fully recovered and will not experience the event of interest in the future; while the other patients who did not recover totally will have the outcome of interest. We proposed a semiparametric estimation method for the proportional hazard mixture cure model, which is easy to implement and computationally efficient. A multiple imputation approach based on the asymptotic normal data augmentation (ANDA) is used to obtain parameter and variance estimates for both the cure probability and survival probability for uncured patients. A simulation study is performed to evaluate the proposed method and the results are compared with a fully parametric approach. The proposed method is applied to 2000-2010 Greater Georgia breast cancer dataset from the Surveillance, Epidemiology, and End Results (SEER) Program.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival analysis is a subject studying statistical methods for analyzing and modeling lifetime data or failure time data. Two main purposes in survival analysis are: estimating the survival probabilities and building regression models between survival time and predictors of interest. One of difficulties lies in analyzing survival data is that it is very likely to be right or interval censored.

1.1 Censoring Data

Time-to-event data present themselves in different ways which create special problems in analyzing such data. One particular feature, often present in time-to-event data, is known as censoring, which, broadly speaking, occurs when some lifetimes are known to have occurred only within certain intervals. There are various categories of censoring, such as right censoring, left censoring, and interval censoring.

Right censored data

In right censoring, the event is observed only if it occurs prior to some prespecified time. For a specific individual under study, we assume that there is a lifetime X and a censoring time, C. The X's are assumed to be independent and identically distributed with probability density function $f(x)$ and survival function $S(x)$. The exact lifetime X of an individual will be known if, and only if, X is less than or equal to C. If X is

greater than C, the individual is a survivor, and his or her event time is censored at C. The data can be conveniently represented by pairs of random variables (T, δ), where δ indicates whether the lifetime X corresponds to an event ($\delta = 1$) or censored ($\delta = 0$), and T is equal to X, if the lifetime is observed, and C if censored, i.e., $T = min(X, C)$.

For right censored data, the commonly used nonparametric technique for estimating survival probabilities is Kaplan-Meier method, also known as the product limit estimator:

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \hat{Pr}(T > t_{(i)} | T \geq t_{(i)}) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i}$$

where $t_{(i)}$ is the ith ordered event times with $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)}$, if we have total of k events. d_i and n_i are number of events and number of cases at risk at time $t_{(i)}$.

Interval censored data

Interval censored survival data is also commonly encountered in practice. For example, in medical research, animal carcinogenicity and epidemiological study ([4], [6]) when patients in a clinical trial or longitudinal study have periodic follow-ups or in industrial experiments where there is periodic inspection for proper functioning of equipment items. In most studies, we know the event happened or will happen in a certain time interval based on a sequence of examination times instead of exact event time. That is, each patient have an observed time interval $(L, R]$ which includes the true event time of interest. When $L = 0$ it is left censored, and if $R = \infty$ it is right censored.

There are many discussions on the estimation methods for survival probabilities under such data. Turnbull [24] developed a self-consistency algorithm for computing nonparametric maximum likelihood estimator (NPMLE); and Groeneboom and

Wellner [5] proposed an Iterative Convex Minorant (ICM) algorithm for NPMLE, which is considerably faster than Turnbull's method when the sample size is large.

1.2 Proportional Hazards Model

In the proportional hazrds (PH) model [3], the hazard function $h(.)$ is assumed to be the product of a baseline hazard function $h_0(.)$ and an exponential linear combination of covariates $\boldsymbol{x}$, that is

$$h(t|\boldsymbol{x}) = h_0(t) \times e^{\beta'\boldsymbol{x}}$$

Under the PH model, the hazard ratio of two subjects is invariant over time. For example, the hazard ratio of two patients with covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is

$$HR = \frac{h_0(t) \times e^{\beta'\boldsymbol{x}_1}}{h_0(t) \times e^{\beta'\boldsymbol{x}_2}} = e^{\beta'(\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

which is a constant over time. An appealing property of the PH model is that, even though the baseline hazard function $h_0(.)$ is unspecified, it is still possible to estimate β's through maximizing the partial likelihood function. The partial likelihood function is:

$$L_p(\beta) = \prod_{j=1}^{n} \frac{exp(\beta'\boldsymbol{x}_j)}{\sum_{k \in R(t_{(j)})} exp(\beta'\boldsymbol{x}_k)}$$

where $R(t_{(j)})$ is the risk set at time $t_{(j)}$ which includes all the subjects that are alive at time $t_{(j)}$. Estimates based on the partial likelihood function are efficient and consistent. It also satisfies the asymptotic normality.

The interpretation of the PH model will be mainly focused on hazard ratios. The survival probabilities can be estimated, for example, using Breslow's approach, which is

$$\hat{S}_{u0}(t|\hat{\beta}) = \prod_{t_i < t} (1 - \frac{d_i}{\sum_{j \in R(t_i)} e^{\hat{\beta}'\boldsymbol{x}_j}})$$

where d_i is the number of events at time t_i.

Because we can obtain the corresponding survival curves $S(t|\boldsymbol{x})$ without specifying the baseline hazard function, the PH model is most popular in survival analysis. The existing package, such as “phreg" in SAS and “coxph" in R, make the PH model very easy to use in practice.

Therefore, we still consider the PH model due to its good properties and straightforward interpretation based on hazard ratios in the interval censored data. However, the partial likelihood does not work under the PH model when the data is interval censored. There are several ways to estimate coefficients and the baseline hazard function for PH model with interval censored data. Basically, it is a maximization problem subject to monotonicity constraints on the part of cumulative hazard function. For example, a two-step algorithm, or sometimes called generalized Gauss-Seidel algorithm [7], is usually recommended to get the maximum likelihood estimators. However, the algorithm could be very slow to compute the profile likelihood curve when sample size is large.

Pan [13] extended the ICM algorithm to the PH model with interval censored survival data. However, the variance for the estimated coefficients needs to be obtained separately through resampling. Later, Pan [12] proposed a multiple imputation approach for the PH model under the interval censored data based on the *Poor Man's data augmentation* (PMDA) and *Asymptotic Normal data augmentation* (ANDA) [25]. These two approaches are approximately equivalent except that PMDA will underestimate the variance when the data have heavy right censoring. The advantage of the multiple imputation is that one can obtain the estimates and their variances simultaneously. The estimation in the multiple imputation used the partial likelihood

estimates for the PH model, the implementation is straightforward and efficient.

1.3 Imputation Approach

Interval-censored data are usually regarded as incomplete data, and imputation or multiple imputation is a general approach for handling missing data or incomplete data problems. The general idea of imputation for observed data $\boldsymbol{O} = \{L_j, R_j, \boldsymbol{x}_j; j = 1, 2, \cdots, n\}$ is to generate one or multiple sets of right censored failure time for T_j's using the observed data. Then use these new data $\boldsymbol{D}_I = \{T_j, \delta_j = 1_{(R_j<\infty)}, \boldsymbol{x}_j; j = 1, 2, \cdots, n\}$ to do estimation and inference about unknown parameters.

For single point imputation approaches, we can use left end point ($T_j = L_j$) or middle point ($T_j = (L_j + R_j)/2$ for $\delta_{R_j} = 0$ and $T_j = L_j$ for $\delta_{R_j} = 1$) for event times. For multiple imputation, we need to use some data augmentation algorithms to impute values for T_j several times and get estimates iteratively.

To be specific, say our interest is to make inference about unknown parameter θ and survival probabilities $S(t)$ based on the observed interval censored data $\boldsymbol{O} = \{L_j, R_j, \boldsymbol{x}_j; j = 1, 2, \cdots, n\}$. The general steps for multiple imputation are as follows:

Step0: Give initial values $\hat{\theta}^{(0)}$ and $\hat{S}^{(0)}(t)$

Step1: At ith iteration and kth imputation: let $T^{(i)}_{(k),j} = L_j$ and $\delta^{(i)}_{(k),j} = 0$ if $R_j = \infty$, otherwise, sample $T^{(i)}_{(k),j}$ from $\hat{S}^{(i-1)}(t)$ conditional on $T^{(i)}_{(k),j} \epsilon (L_j, R_j]$ and let $\delta^{(i)}_{(k),j} = 1$. This gives m sets of right-censored data $\boldsymbol{D}^{(i)}_{I(k)} = \{T^{(i)}_{k,j}, \delta^{(i)}_{(k),j}; j = 1, \cdots, n\}, k = 1, \cdots, m$.

Step2: Obtain estimates $\hat{\theta}^{(i)}_{(k)}$, $\hat{\Sigma}^{(i)}_{\theta(k)}$ and $\hat{S}^{(i)}_{(k)}(t)$ for each $\boldsymbol{D}^{(i)}_{I(k)}$, $k = 1, \cdots, m$.

Step3: Update $\hat{\theta}^{(i)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_{(k)}^{(i)}$ and $\hat{S}^{(i)}(t) = \frac{1}{m}\sum_{k=1}^{m}\hat{S}_{(k)}^{(i)}(t)$, and variance estimate

$$\hat{\Sigma}_{\theta}^{(i)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\Sigma}_{\theta(k)}^{(i)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\theta}_{(k)}^{(i)} - \hat{\theta}^{(i)})(\hat{\theta}_{(k)}^{(i)} - \hat{\theta}^{(i)})'}{m-1}$$

Step4: Repeat Steps 1-3 until convergence.

The initial values in Step 0 can be obtained through some single point imputation approach, for example, mid-point imputation. The advantage of multiple imputation is that the variance estimate includes both the within imputation variance and the between imputation variance. The weight 1 and $1+1/m$, where the additional weight for between imputation variance is used to account for the finite number of imputations.

1.4 Mixture Cure Model

As we all know that, regular survival models assume that all the subjects will have the event of interest eventually, but nowadays we have many diseases such as prostate cancer and breast cancer that are curable. Where a fraction of patients can be considered as fully recovered after some treatment and will not experience the event of interest in the future; while the other patients who did not recover totally will have a survival probability. In this case, the regular survival models are not appropriate and we need to use mixture cure models.

The most popular approach in analyzing survival data with potentially cured patients is the mixture cure model proposed by Boag [1], which has been commonly used in the last decade. Let T be a nonnegative random variable denoting the failure time of a patient, and $S(t)$ be the survival functions of T, The mixture cure model is

given by

$$S(t) = 1 - \pi + \pi S_u(t),$$

where π is the proportion of uncured patients and $S_u(t)$ denotes the survival probability of uncured patients.

There are two components in the mixture cure model: the cure rate component, sometimes called the incidence part and a survival model for the uncured patients, sometimes called the latency part. After 1950, there are many discussions on this model. The patients' characteristic and other possible risk factor have been incorporated into the incidence and latency part.

For right censored data, an EM algorithm was proposed by Peng [14] for fitting the mixture cure model, and an R package "smcure" was contributed by Cai [2] to realize this method.

There is sparse research that focuses on the mixture cure model with interval censored data. For current status data, Ma [11] used the penalized maximum likelihood method, where the baseline hazard is updated through the *pool-adjacent violator algorithm* separately, and the weighted bootstrap was adopted for variance estimation. For interval censored data, Kim [8] proposed the EM algorithm to estimate the mixture cure model by assuming a piecewise exponential distribution for the baseline hazard with multiple imputation to estimate the variance. Later, a multiple imputation method based on ANDA was discussed in Lam [10] to deal with a frailty cox PH model under right censoring and interval censoring case. The frailty term was assumed to follow a noncentral chi-square distribution with zero degree of freedom and was updated in the algorithm using the posterior gamma distribution.

1.5 Outline of Thesis

The main aim of this thesis is to develop a semiparametric estimation method for the mixture cure model for interval censored data, which is easy to implement and efficient in computation. We adopt the multiple imputation approach based on the PMDA and ANDA in the proportional hazard mixture cure (PHMC) model for the interval censored data. The attractive property of the proposed method is that it can be easily implemented using existing functions in statistical software, such as “glm" and “coxph" in R [16].

In Chapter 2, we introduce the multiple imputation algorithms. We outline the notation, model description of the proportional hazards mixture cure model in 2.1. Then the algorithm based on PMDA and ANDA will be discussed in Section 2.3 separately. We conduct extensive simulation studies in Chapter 3. We evaluate the performance of the multiple imputation approach based on the PMDA, ANDA and the parametric methods in Section 3.1. We also carry out a sensitivity analysis regarding to the misspecification of cure rate component in Section 3.2. And a sensitivity analysis to compare the rounding down survival time of right censored data and interval censored data in Section 3.3. In Chapter 4, we apply the proposed methods to the 2000-2010 Greater Georgia breast cancer SEER dataset. Both of the crude model and adjusted model are fitted and discussed. Finally, we finish with a discussion and conclusions in Chapter 5.

Chapter 2

Multiple imputation Algorithm for PHMC model with Interval censored data

2.1 Proportional Hazards Mixture Cure Model

Let T be a nonnegative random variable denoting the failure time of a patient, and $S(t|\boldsymbol{x}, \boldsymbol{z})$ be the survival functions of T, where $\boldsymbol{x}$ and $\boldsymbol{z}$ are observed values of two covariate vectors on which the distribution of T may depend. The mixture cure model is given by

$$S(t|\boldsymbol{x}, \boldsymbol{z}) = 1 - \pi(\boldsymbol{z}) + \pi(\boldsymbol{z}) S_u(t|\boldsymbol{x}),$$

where $\pi(\boldsymbol{z})$ is the proportion of uncured patients depending on $\boldsymbol{z}$ via the inverse logistic function

$$\pi(\boldsymbol{z}) = \frac{e^{\gamma' \boldsymbol{z}}}{1 + e^{\gamma' \boldsymbol{z}}},$$

and $S_u(t|\boldsymbol{x})$ denotes the survival probability of uncured patients depending on $\boldsymbol{x}$. Note, as mentioned in Price [15], other link functions can also be specified for $\pi(\boldsymbol{z})$, such as the probit link ($\pi(\boldsymbol{z}) = \Phi(\boldsymbol{\gamma}'\boldsymbol{z})$) and complementary log-log link ($\pi(\boldsymbol{z}) = 1 - e^{-e^{\gamma' \boldsymbol{z}}}$). We can model the survival probability of uncured patients by the PH model. That is,

$$S_u(t|\boldsymbol{x}) = S_{u0}(t)^{e^{\boldsymbol{\beta}' \boldsymbol{x}}},$$

where $S_{u0}(\cdot)$ is the survival function of baseline distribution. We refer to this as the proportional hazard mixture cure (PHMC) model.

Let $\boldsymbol{O} = (L_j,\ R_j, \delta_{L_j}, \delta_{R_j}, \delta_{I_j}, \boldsymbol{x}_j, \boldsymbol{z}_j, j = 1, 2, \ldots, n)$ denote the observed data, where $(L_j,\ R_j]$ is the observed time interval including the exact event time T_j for the jth subject, and $\delta_{L_j}, \delta_{R_j}, \delta_{I_j}$ are the censoring indicators. $\delta_{L_j} = 1$ when it is left censored ($L_j = 0$); $\delta_{R_j} = 1$ when it is right censored ($R_j = \infty$); and $\delta_{I_j} = 1$ when it is interval censored ($0 < L_j < R_j < \infty$) with the convention that $\delta_{L_j} + \delta_{I_j} + \delta_{R_j} = 1$ for all j.

2.2 Likelihood Function

Assuming the censoring is independent and non-informative, the observed likelihood function of interval censored data is

$$\prod_{j=1}^{n} [1 - S(R_j|\boldsymbol{x}_j, \boldsymbol{z}_j)]^{\delta_{L_j}} [S(L_j|\boldsymbol{x}_j, \boldsymbol{z}_j) - S(R_j|\boldsymbol{x}_j, \boldsymbol{z}_j)]^{\delta_{I_j}} S(L_j|\boldsymbol{x}_j, \boldsymbol{z}_j)^{\delta_{R_j}}$$

Specifically, for the PHMC model, it is

$$\prod_{j=1}^{n} \pi(\boldsymbol{z}_j)^{1-\delta_{R_j}} [1 - S_{u0}(R_j)^{e^{\beta' \boldsymbol{x}_j}}]^{\delta_{L_j}} [S_{u0}(L_j)^{e^{\beta' \boldsymbol{x}_j}} - S_{u0}(R_j)^{e^{\beta' \boldsymbol{x}_j}}]^{\delta_{I_j}} \times [1 - \pi(\boldsymbol{z}_j) + \pi(\boldsymbol{z}_j) S_{u0}(L_j)^{e^{\beta' \boldsymbol{x}_j}}]^{\delta_{R_j}} \quad (2.1)$$

Once $\pi(\boldsymbol{z}_j)$ and $S_{u0}(\cdot)$ are fully specified, the unknown parameters can be obtained directly from maximum likelihood estimation. The disadvantage of the fully parametric approach is that it is hard to verify the appropriate parametric form or there may not exist a specific parametric form for a particular data. When this is the case, semiparametric estimation is preferred. However, it is a challenge task to achieve the semiparametric estimation method under the mixture cure model with interval censored data.

For the purpose of computation, it is convenient to define a latent variable $u_j, j = 1, \cdots, n$, which is the uncure indicator where $u_j = 1$ if the jth subject is uncured and

0 if cured. Here, we have $u_j = 1$ if $\delta_{L_j} = 1$ or $\delta_{I_j} = 1$. There is the probability of being cured for the right censored observations, and its contribution to the likelihood function is $[1 - \pi(\boldsymbol{z}_j)]$ when $u_j = 0$ and $[\pi(\boldsymbol{z}_j)S_{u0}(L_j)^{e^{\beta'\boldsymbol{x}_j}}]$ when $u_j = 1$. Thus, the likelihood function conditional on the $u_j's$ can be written as

$$\prod_{j=1}^{n} \pi(\boldsymbol{z}_j)^{1-\delta_{R_j}}[1 - S_{u0}(R_j)^{e^{\beta'\boldsymbol{x}_j}}]^{\delta_{L_j}}[S_{u0}(L_j)^{e^{\beta'\boldsymbol{x}_j}} - S_{u0}(R_j)^{e^{\beta'\boldsymbol{x}_j}}]^{\delta_{I_j}} \times [1 - \pi(\boldsymbol{z}_j)]^{(1-u_j)\delta_{R_j}}[\pi(\boldsymbol{z}_j)S_{u0}(L_j)^{e^{\beta'\boldsymbol{x}_j}}]^{u_j\delta_{R_j}} \quad (2.2)$$

Using the fact that $(1-u_j)\delta_{R_j} = 1-u_j$, we can simplify the conditional likelihood function as

$$L(\beta, \gamma, S_{u0}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{O}) = L_1(\gamma) \times L_2(\beta, S_{u0})$$

where

$$L_1(\gamma) = \prod_{j=1}^{n} \pi(\boldsymbol{z}_j)^{u_j}(1 - \pi(\boldsymbol{z}_j))^{1-u_j}$$

$$L_2(\beta, S_{u0}) = \prod_{u_j=1} [1 - S_{u0}(R_j)^{e^{\beta'\boldsymbol{x}_j}}]^{\delta_{L_j}}[S_{u0}(L_j)^{e^{\beta'\boldsymbol{x}_j}} - S_{u0}(R_j)^{e^{\beta'\boldsymbol{x}_j}}]^{\delta_{I_j}}[S_{u0}(L_j)^{e^{\beta'\boldsymbol{x}_j}}]^{\delta_{R_j}}$$

If u can be observed, the likelihood function can be estimated by maximizing $L_1(\gamma)$ and $L_2(\beta, S_{u0})$ separately regarding to β, γ and S_{u0}.

Using a similar idea as Lam [9], we impute the uncure rate indicator u_j through the binomial distribution with probability being conditioned on the censoring indicators. The conditional expectation of u_j given $(\beta,\gamma, S_{u0}, \boldsymbol{O})$ is

$$\begin{aligned} w_j &= E(u_j|\beta,\gamma, S_{u0}, \boldsymbol{O}) \\ &= 1 - \delta_{R_j} + \delta_{R_j}\frac{\pi(\boldsymbol{z}_j)(S_{u0})^{e^{\beta'\boldsymbol{x}_j}}}{1 - \pi(\boldsymbol{z}_j) + \pi(\boldsymbol{z}_j)(S_{u0})^{e^{\beta'\boldsymbol{x}_j}}} \end{aligned} \quad (2.3)$$

Where w_j is 1 when time is left or interval censored, which indicates the patients are uncured, and a probability of being cured for right censored observations. Then we sample $\boldsymbol{u} = (u_1, u_2, \cdots, u_n)$ from a Bernoulli distribution with the probability

$\boldsymbol{w} = (w_1, w_2, \cdots, w_n)$. Conditional on $\boldsymbol{u}$, we can estimate γ by maximize $L_1(\gamma)$ through generalized linear models using "glm" in R, and β by maximizing $L_2(\beta, S_{u0})$ through semiparametric approach using "intcox" package in R. However, there exist some issues related with "intcox" in this approach: sometimes the algorithm cannot converge; there is no variance estimation directly from "intcox"; and the computational speed is slow. Therefore, instead of estimating β directly from the interval censored data, we adopt the method in Pan [12], where β is estimated through the multiple imputation with right censored data through "coxph" function in R.

2.3 Multiple Imputation Algorithms

Follow Wei and Tanner [25] and Pan [12], we develop multiple imputation algorithms based on the *Poor man's data augmentation* (PMDA) and the *asymptotic normal data augmentation* (ANDA) under the PHMC model for interval censored data.

We use the superscript i and subscript k to denote the ith iteration and the kth imputed data set $\boldsymbol{O}_{(k)}^{(i)} = \{L, R, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}_{(k)}^{(i)}, T_{(k)}^{(i)}\}$. Where $\boldsymbol{u}_{(k)}^{(i)}$ and $T_{(k)}^{(i)}$ are the vectors of imputed cure indicators and event times.

Algorithm based on PMDA

The details of our algorithm based on PMDA are as follows:

Step 1: Initial values: let $\boldsymbol{u}^{(0)} = 1 - \boldsymbol{\delta}_R$, and $T^{(0)}$ be the midpoints of time intervals (L, R) (see Sun [19]) for uncured patients only ($\boldsymbol{u}^{(0)} = 1$). We fit a logistic regression with $\boldsymbol{u}^{(0)}$ and $\boldsymbol{z}$ to get $\hat{\gamma}^{(0)}$. We also fit a Cox PH model for $\{T^{(0)}, 1 - \boldsymbol{\delta}_R, \boldsymbol{x}\}$ and get $\hat{\beta}^{(0)}$ and estimate for baseline survival function $\hat{S}_{u0}^{(0)}$.

Step 2: Multiple imputation: Assume that the estimates from the ith iteration are denoted as $(\hat{\gamma}^{(i)}, \hat{\beta}^{(i)}, \hat{S}_{u0}^{(i)})$. For the $(i+1)$th iteration, we impute the cure indicator $\boldsymbol{u}^{(i+1)}$ and the survival time $T^{(i+1)}$ for the interval censored observation m times. Specifically, in the kth $(k = 1, \cdots, m)$ imputation,

$(s1)$ Update the posterior probability of being uncured

$$\boldsymbol{w}_{(k)}^{(i+1)} = E(\boldsymbol{u}^{(i+1)} | \hat{\beta}_{(k)}^{(i)}, \hat{\gamma}_{(k)}^{(i)}, \hat{S}_{u0}^{(i)}, \boldsymbol{O}_{(k)}^{(i)})$$

.

$(s2a)$ Based on $\boldsymbol{w}_{(k)}^{(i+1)}$, we sample $\boldsymbol{u}_{(k)}^{(i+1)} \sim Ber(\boldsymbol{w}_{(k)}^{(i+1)})$, where $Ber(.)$ is the Bernoulli distribution.

$(s2b)$ Sample $T_{(k),j}^{(i+1)}$ from the time intervals $(L_j, R_j]$, $(j = 1, \cdots, n)$ in the following way: if subject j is right censored, we keep it and let $T_{(k),j}^{(i+1)} = L_j$ and $\delta_{(k),j}^{(i+1)} = 0$; otherwise, we sample $T_{(k),j}^{(i+1)}$ from $\hat{S}_{u(k),j}^{(i)}$, $\hat{S}_{u(k),j}^{(i)} = (\hat{S}_{u0}^{(i)})^{e^{\hat{\beta}_{(k)}^{(i)\prime} \boldsymbol{x}_j}}$, conditional on $(L_j, R_j]$ and let $\delta_{(k),j}^{(i+1)} = 1$.

$(s3)$ Use the imputed cure rate indicator in $(s2a)$ as outcome to update $\hat{\gamma}_{(k)}^{(i+1)}$ using the logistic regression.

$(s4)$ Fit the PH model for uncured patients only ($\boldsymbol{u}_{(k)}^{(i+1)} = 1$ from $(s2a)$) with the imputed data set $\{T_{(k)}^{(i+1)}, \delta_{(k)}^{(i+1)}, \boldsymbol{x}\}$ and get estimates $\hat{\beta}_{(k)}^{(i+1)}$, and $\hat{S}_{u0(k)}^{(i+1)}$.

Step 3: Calculate updated estimates: at the end of each iteration, we can obtain the updated estimates for coefficients:

$$\hat{\beta}^{(i+1)} = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_{(k)}^{(i+1)} \tag{2.4}$$

$$\hat{\gamma}^{(i+1)} = \frac{1}{m} \sum_{k=1}^{m} \hat{\gamma}_{(k)}^{(i+1)} \tag{2.5}$$

As well as the estimated baseline survival function:

$$\hat{S}_{u0}^{(i+1)} = \frac{1}{m} \sum_{k=1}^{m} \hat{S}_{u0(k)}^{(i+1)} \tag{2.6}$$

Step 4: Repeat $Step2 - Step3$ until convergence. For the convergence criterion, we use $max\{(\hat{\beta}^{(i)} - \hat{\beta}^{(i-1)})^2, (\hat{\gamma}^{(i)} - \hat{\gamma}^{(i-1)})^2\} < .001$ or $i > 100$.

After convergence, the variance of the coefficients can be obtained using estimates from the last iteration, say g, $\{\hat{\beta}^{(g)}_{(k)}, \hat{\Sigma}^{(g)}_{\beta(k)}, \hat{\gamma}^{(g)}_{(k)}, \hat{\Sigma}^{(g)}_{\gamma(k)}\}$:

$$\hat{\Sigma}_\beta = \frac{1}{m}\sum_{k=1}^{m}\hat{\Sigma}^{(g)}_{\beta(k)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\beta}^{(g)}_{(k)} - \hat{\beta}^{(g)})(\hat{\beta}^{(g)}_{(k)} - \hat{\beta}^{(g)})'}{m-1} \tag{2.7}$$

$$\hat{\Sigma}_\gamma = \frac{1}{m}\sum_{k=1}^{m}\hat{\Sigma}^{(g)}_{\gamma(k)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\gamma}^{(g)}_{(k)} - \hat{\gamma}^{(g)})(\hat{\gamma}^{(g)}_{(k)} - \hat{\gamma}^{(g)})'}{m-1} \tag{2.8}$$

Algorithm based on ANDA

The details of our algorithm based on ANDA are as follows:

Step 1: Initial values: let $\boldsymbol{u}^{(0)} = 1 - \boldsymbol{\delta}_R$, and $T^{(0)}$ be the midpoints of time intervals (L, R) (see Sun [19]) for uncured patients only ($\boldsymbol{u}^{(0)} = 1$). We fit a logistic regression with $\boldsymbol{u}^{(0)}$ and $\boldsymbol{z}$ to get $\hat{\gamma}^{(0)}$ and the covariance matrix estimate $\hat{\Sigma}^{(0)}_\gamma$. We also fit a Cox PH model for $\{T^{(0)}, 1 - \boldsymbol{\delta}_R, \boldsymbol{x}\}$ and get $\hat{\beta}^{(0)}$, covariance matrix $\hat{\Sigma}^{(0)}_\beta$ and estimates for baseline survival $\hat{S}^{(0)}_{u0}$.

Step 2: Multiple imputation: assume that the estimates from the ith iteration are denoted as ($\hat{\gamma}^{(i)}$, $\hat{\beta}^{(i)}$, $\hat{\Sigma}^{(i)}_\gamma$, $\hat{\Sigma}^{(i)}_\beta$, $\hat{S}^{(i)}_{u0}$). For the $(i+1)$th iteration, we impute the cure indicator $\boldsymbol{u}^{(i+1)}$ and the survival time $T^{(i)}$ for the interval censored observation m times. Specifically, in the kth ($k = 1, \cdots, m$) imputation,

($s1$) Sample β and γ from the normal distributions: $\beta \sim N(\hat{\beta}^{(i)}, \hat{\Sigma}^{(i)}_\beta)$ and $\gamma \sim N(\hat{\gamma}^{(i)}, \hat{\Sigma}^{(i)}_\gamma)$. Say we get $\tilde{\beta}^{(i+1)}_{(k)}$ and $\tilde{\gamma}^{(i+1)}_{(k)}$.

($s2$) Update the posterior probability of being uncured

$$\boldsymbol{w}_{(k)}^{(i+1)} = E(\boldsymbol{u}^{(i+1)}|\tilde{\beta}_{(k)}^{(i+1)}, \tilde{\gamma}_{(k)}^{(i+1)}, \hat{S}_{u0}^{(i)}, \boldsymbol{O}_{(k)}^{(i)})$$

.

($s3a$) Based on $\boldsymbol{w}_{(k)}^{(i+1)}$, we sample $\boldsymbol{u}_{(k)}^{(i+1)} \sim Ber(\boldsymbol{w}_{(k)}^{(i+1)})$, where $Ber(.)$ is the Bernoulli distribution.

($s3b$) Sample $T_{(k),j}^{(i+1)}$ from the time intervals $(L_j, R_j]$, $(j = 1, \cdots, n)$ in the following way: if subject j is right censored, we keep it and let $T_{(k),j}^{(i+1)} = L_j$ and $\delta_{(k),j}^{(i+1)} = 0$; otherwise, we sample $T_{(k),j}^{(i+1)}$ from $\hat{S}_{u(k),j}^{(i)}$, $\hat{S}_{u(k),j}^{(i)} = (\hat{S}_{u0}^{(i)})^{e^{\tilde{\beta}_{(k)}^{(i)\prime}\boldsymbol{x}_j}}$, conditional on $(L_j, R_j]$ and let $\delta_{(k),j}^{(i+1)} = 1$.

($s4$) Use the imputed cure rate indicator in ($s3a$) as outcome to update $\hat{\gamma}_{(k)}^{(i+1)}$ and $\hat{\Sigma}_{\gamma(k)}^{(i+1)}$ using the logistic regression.

($s5$) Fit the PH model for uncured patients only ($\boldsymbol{u}_{(k)}^{(i+1)} = 1$ from ($s3a$)) with the imputed data set $\{T_{(k)}^{(i+1)}, \delta_{(k)}^{(i+1)}, \boldsymbol{x}\}$ and get estimates $\hat{\beta}_{(k)}^{(i+1)}$, $\hat{\Sigma}_{\beta(k)}^{(i+1)}$ and $\hat{S}_{u0(k)}^{(i+1)}$.

Step 3: Calculate updated estimates: at the end of each iteration, we can obtain the updated estimates for coefficients and their variance:

$$\hat{\beta}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_{(k)}^{(i+1)} \tag{2.9}$$

$$\hat{\gamma}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\gamma}_{(k)}^{(i+1)} \tag{2.10}$$

$$\hat{\Sigma}_{\beta}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\Sigma}_{\beta(k)}^{(i+1)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\beta}_{(k)}^{(i+1)} - \hat{\beta}^{(i+1)})(\hat{\beta}_{(k)}^{(i+1)} - \hat{\beta}^{(i+1)})'}{m-1} \tag{2.11}$$

$$\hat{\Sigma}_{\gamma}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\Sigma}_{\gamma(k)}^{(i+1)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\gamma}_{(k)}^{(i+1)} - \hat{\gamma}^{(i+1)})(\hat{\gamma}_{(k)}^{(i+1)} - \hat{\gamma}^{(i+1)})'}{m-1} \tag{2.12}$$

As well as the estimated baseline survival function:

$$\hat{S}_{u0}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{S}_{u0(k)}^{(i+1)} \tag{2.13}$$

Step 4: Repeat $Step2 - Step3$ until convergence. For the convergence criterion in $Step$ 4, we use $max\{(\hat{\beta}^{(i)} - \hat{\beta}^{(i-1)})^2, (\hat{\gamma}^{(i)} - \hat{\gamma}^{(i-1)})^2\} < .001$ or $i > 100$.

The coefficient estimates are the average of the estimates from the m imputations, and the variance estimates are the weighted average of the within imputation variances and between imputation variance. The second term of the variance components in equations 2.11 and 2.12 has an additional weight $1/m$ which is an inflation factor used to take account of a finite number of imputations (Rubin [17]; Tanner and Wong [22]; Schenker and Welsh [18]). As the size of the multiple imputation m does not necessarily to be large, we choose $m = 10$ in our study as suggested by Pan [12].

We use Breslow's method to estimate the baseline survival function

$$\hat{S}_{u0}(t|\hat{\beta}) = \prod_{t_i<t} \left(1 - \frac{d_i}{\sum_{j\in R(t_i)} e^{\hat{\beta}'\boldsymbol{x}_j}}\right)$$

As suggested by Taylor [23], we apply zero constraint to the survival function. That is, the survival probability for uncured patient after the maximum value of finite R_j will be restricted to zero. When we update $\boldsymbol{w}$, we calculate the survival probability at the left time points for right censored patients, based on the estimated step function of baseline survival from previous iteration.

In $(s2b)$ of PMDA (or $(s3b)$ of ANDA), we sample $T^{(i+1)}_{(k),j}$ from $\hat{S}^{(i)}_{u(k),j}$ conditional on $(L_j, R_j]$ for non-right-censored patients. That is, for the jth patient with $R_j < \infty$, we have $\{t_1, t_2, ..., t_{n_j}\}$ discrete time points in interval $(L_j, R_j]$. The corresponding probability mass is $\{p_1, p_2, ..., p_{n_j}\}$, which can be calculated as $p_{n_s} = \hat{S}^{(i)}_{u(k),j}(t_{n_s-1}) - \hat{S}^{(i)}_{u(k),j}(t_{n_s})$. Thus we sample $T^{(i+1)}_{(k),j}$ from $\{t_1, t_2, ..., t_{n_j}\}$ with probability proportional to $\{p_1, p_2, ..., p_{n_j}\}$.

Pan [12] mentioned that the results from PMDA is asymptotically equivalent to that based on ANDA, except that PMDA will underestimate the variance when the data has relatively large proportion of right censoring observations. Because we usually have high proportion of censoring when there exists cure rate, we expect that the performance of the algorithm based on ANDA will be over PMDA in variance estimation. We compare these two approaches in the simulation study using different censoring proportions. Based on the results, we recommend to use ANDA for the mixture cure model.

Chapter 3

Simulation Study

3.1 Performance of PMDA and ANDA

We design a simulation study to examine the performance of the proposed method. In our design, we have two predictors in the latency part, $\boldsymbol{x}_1$ follows uniform distribution $U(0,2)$ and $\boldsymbol{x}_2$ follows bernoulli distribution $Ber(.5)$. For the incidence part, we have $\boldsymbol{z}_1$ follows $U(0,2)$ and $\boldsymbol{z}_2$ follows $Ber(.5)$. The coefficients are set to $\beta = (1,-1)$ and $\gamma = (0,1,-1)$, yielding an average cure probability of 40% and $\gamma = (-1,1,-1)$ yielding an average cure probability of 60%. For the baseline distribution, we use a Weibull baseline distribution with shape parameter 2 and scale parameter 1 and a Lognormal baseline distribution with log mean of 0 and log standard deviation of 0.2. A sample size of 500 is used.

We generate the interval censored survival time in the following way: each subject has p equally spaced random intervals of length τ that their events could fall into, defined as the time points $0 < Y_j < Y_j + \tau < ... < Y_j + p \times \tau < \infty$ where $Y_j \sim U(0,\nu)$, $j = 1, 2, \ldots, n$. We have left censoring if the generated survival time $T_j < Y_j$ and the patient is not cured $u_j = 1$. We have right censoring if $T_j > Y_j + p \times \tau$ or the subject is cured $u_j = 0$. Otherwise, if the patient is not cured $u_j = 1$, and $T_j \in (Y_j, Y_j + p \times \tau]$, we have interval censoring and we choose the smallest interval which covers the true event time. We fix $\nu = 1$ and $\tau = 0.2$, and adjust p to have different right censoring rate such as 45% or 60% for 40% cure rate.

Table 3.1 Simulation Results for 1000 Replications for 40% Cure Rate

		Fully Parametric				**MI PMDA**				**MI ANDA**			
cens%	par.	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP
Lognormal Baseline													
45	γ_0	-0.008	0.218	0.226	0.956	0.011	0.225	0.227	0.948	-0.002	0.223	0.229	0.960
	γ_1	0.007	0.191	0.191	0.952	0.021	0.197	0.192	0.953	0.010	0.193	0.193	0.952
	γ_2	0.000	0.205	0.214	0.963	-0.015	0.212	0.215	0.961	-0.005	0.210	0.216	0.956
	β_1	0.021	0.126	0.136	0.973	0.032	0.132	0.132	0.945	0.015	0.134	0.143	0.964
	β_1	-0.024	0.148	0.150	0.959	-0.043	0.156	0.149	0.941	-0.026	0.157	0.155	0.949
60	γ_0	0.004	0.272	0.285	0.965	0.036	0.289	0.261	0.931	0.033	0.288	0.300	0.954
	γ_1	0.024	0.244	0.242	0.950	0.050	0.260	0.222	0.904	0.037	0.259	0.254	0.938
	γ_2	-0.020	0.271	0.271	0.948	-0.045	0.286	0.248	0.909	-0.032	0.285	0.284	0.941
	β_1	0.018	0.171	0.169	0.944	0.032	0.180	0.160	0.910	0.004	0.180	0.180	0.942
	β_1	-0.002	0.176	0.186	0.960	-0.022	0.186	0.180	0.941	0.008	0.184	0.188	0.954
Weibull Baseline													
45	γ_0	0.006	0.240	0.228	0.938	0.013	0.247	0.226	0.926	-0.006	0.244	0.227	0.933
	γ_1	0.000	0.200	0.192	0.940	0.004	0.208	0.191	0.935	-0.013	0.203	0.191	0.938
	γ_1	-0.006	0.220	0.216	0.940	-0.011	0.227	0.214	0.935	0.006	0.222	0.214	0.941
	β_1	0.026	0.132	0.136	0.952	0.022	0.139	0.130	0.924	-0.006	0.140	0.138	0.949
	β_1	-0.028	0.148	0.151	0.950	-0.028	0.152	0.146	0.936	-0.006	0.154	0.151	0.942
60	γ_0	-0.012	0.295	0.309	0.957	-0.014	0.320	0.262	0.890	-0.037	0.313	0.310	0.939
	γ_2	0.031	0.265	0.265	0.957	0.038	0.293	0.224	0.879	0.001	0.265	0.262	0.944
	γ_2	-0.012	0.274	0.293	0.964	-0.017	0.310	0.251	0.895	0.014	0.283	0.292	0.945
	β_1	0.015	0.173	0.175	0.954	0.008	0.183	0.159	0.910	-0.020	0.183	0.190	0.949
	β_1	-0.009	0.196	0.197	0.953	-0.009	0.205	0.179	0.897	0.036	0.203	0.195	0.940

We did 1000 replications for each simulation setting, and report the bias, average estimated standard deviation (StErr), empirical standard deviation (StDev) and empirical coverage probability (CP). Our multiple imputation algorithms based on both ANDA and PMDA are reported, and all the results are compared with the fully parametric maximum likelihood estimates. Table 3.1 is for the 40% cure rate probability and Table 3.2 is for the 60% case. Basically, all the three approaches are comparable with respect to biases under the Weibull and Lognormal baseline distributions. The StErr and StDev for our proposed algorithm based on ANDA are similar and the coverage probabilities are close to 95%, the overall performance of this approach is

Table 3.2 Simulation Results for 1000 Replications for 60% Cure Rate

		Fully Parametric				**MI PMDA**				**MI ANDA**			
cens%	par.	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP
Lognormal Baseline													
65	γ_0	-0.005	0.240	0.249	0.955	0.017	0.247	0.246	0.947	0.007	0.246	0.254	0.957
	γ_1	0.022	0.197	0.201	0.950	0.037	0.202	0.200	0.939	0.030	0.200	0.205	0.947
	γ_2	-0.023	0.227	0.225	0.950	-0.036	0.233	0.224	0.939	-0.029	0.234	0.230	0.948
	β_1	0.019	0.185	0.189	0.954	0.045	0.197	0.180	0.929	0.015	0.198	0.202	0.951
	β_2	0.001	0.202	0.207	0.957	-0.031	0.214	0.201	0.934	-0.007	0.211	0.214	0.962
80	γ_0	-0.011	0.364	0.354	0.950	0.066	0.402	0.286	0.822	0.068	0.399	0.386	0.929
	γ_1	0.028	0.296	0.286	0.942	0.090	0.331	0.235	0.822	0.069	0.328	0.315	0.939
	γ_2	-0.018	0.311	0.318	0.952	-0.077	0.352	0.261	0.859	-0.052	0.341	0.349	0.935
	β_1	0.036	0.266	0.273	0.958	0.007	0.276	0.242	0.916	-0.077	0.262	0.270	0.941
	β_2	-0.025	0.310	0.304	0.942	-0.009	0.322	0.273	0.894	0.086	0.307	0.289	0.917
Weibull Baseline													
65	γ_0	-0.008	0.250	0.252	0.949	0.002	0.256	0.245	0.931	-0.016	0.257	0.253	0.950
	γ_1	0.026	0.204	0.202	0.946	0.035	0.211	0.198	0.936	0.023	0.206	0.203	0.952
	γ_2	-0.025	0.232	0.226	0.938	-0.033	0.237	0.222	0.932	-0.021	0.234	0.228	0.936
	β_1	0.016	0.185	0.193	0.953	0.014	0.198	0.177	0.917	-0.036	0.196	0.198	0.940
	β_2	-0.007	0.215	0.217	0.948	-0.009	0.223	0.198	0.923	0.034	0.216	0.213	0.946
80	γ_0	0.000	0.361	0.380	0.975	0.019	0.399	0.286	0.857	-0.046	0.376	0.361	0.929
	γ_1	0.049	0.346	0.339	0.951	0.083	0.366	0.235	0.842	-0.001	0.292	0.289	0.945
	γ_2	-0.042	0.341	0.349	0.964	-0.070	0.388	0.261	0.858	0.001	0.323	0.320	0.935
	β_1	0.057	0.280	0.277	0.939	0.040	0.297	0.234	0.871	-0.012	0.276	0.305	0.957
	β_2	-0.052	0.306	0.310	0.953	-0.033	0.313	0.262	0.908	0.050	0.304	0.305	0.934

similar to the fully parametric method. The StErr and StDev for PMDA are similar when the censoring rate is small, for example 45% right censoring under 40% cure rate and 65% right censoring under 60% cure rate. However, PMDA underestimates variance when censoring rate is large, which results in a low coverage probability.

We further investigate the multiple imputation methods with regard to their estimated survival curves. Curves under both multiple imputation methods are similar in different settings. For illustration purpose, we present the estimated baseline survival curves along with 2.5% and 97.5% quartiles under the lognormal and weibull settings with 40% cure rate in Figures 3.1 to 3.4. These were obtained from the multiple

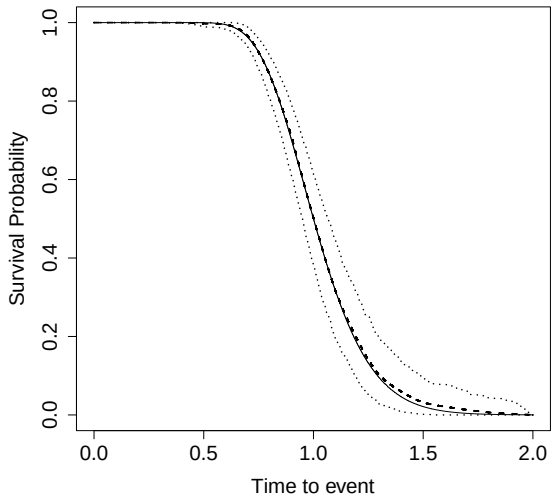

Figure 3.1 Estimated Baseline Survival Probability for Lognormal with 45% Censoring

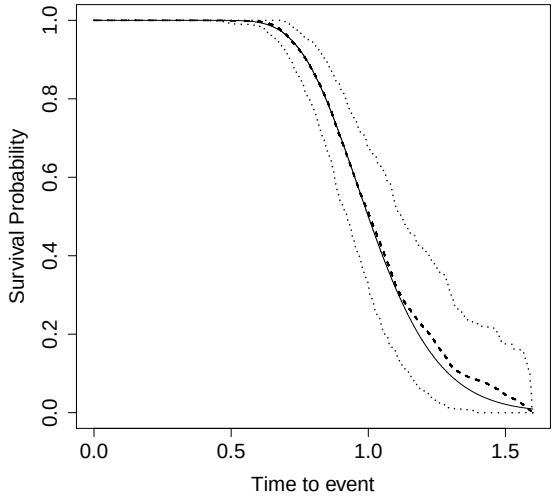

Figure 3.2 Estimated Baseline Survival Probability for Lognormal with 60% Censoring

imputation approach with ANDA where the solid lines are the truth. Plots for the 60% cure rate are very similar to these results. It turns out that the estimated curves are close to the truth for most of the time points, and the small differences at the tail of the survival functions for large censoring cases may due to the zero-tail restriction. The true survival curves lie in the 95% confidence interval, which demonstrates the

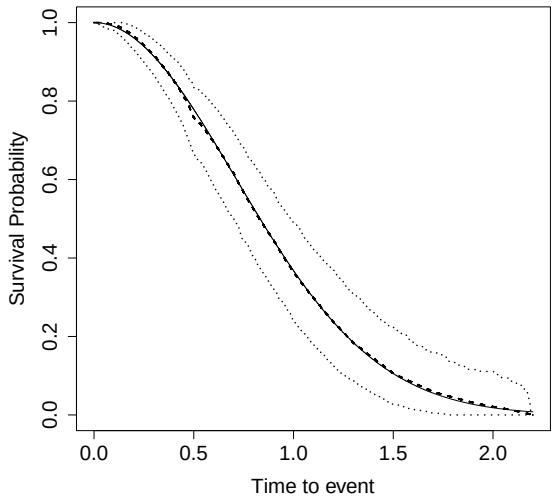

Figure 3.3 Estimated Baseline Survival Probability for Weibull with 45% Censoring

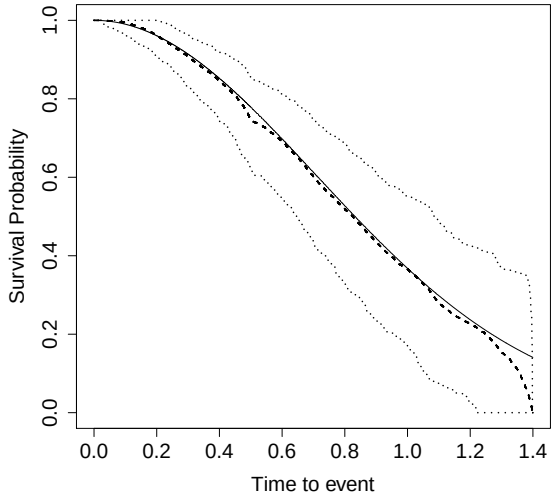

Figure 3.4 Estimated Baseline Survival Probability for Weibull with 60% Censoring

validity of the proposed method.

3.2 Sensitivity Analysis: link functions for uncured rate

Furthermore, we conducted a sensitivity analysis for our proposed algorithm based on ANDA with respect to different link functions, such as logit link, probit link and complimentary log-log link, for the cure probability. Similar to the setting in the simulation study, we have two predictors in the latency part, $\boldsymbol{x}_1$ follows uniform distribution $U(0,2)$ and $\boldsymbol{x}_2$ follows bernoulli distribution $Ber(.5)$. For the incidence part, we have $\boldsymbol{z}_1$ follows $Ber(.5)$ and $\boldsymbol{z}_2$ follows $U(0,2)$. The coefficients are set to $\beta = (1,-1)$ and $\gamma = (0,1,-1)$. Weibull baseline distribution with shape parameter 2 and scale parameter 1 is used and the sample size is set to 500.

For the purpose of easy interpretation of cure probability, we fix $\boldsymbol{z}_2$ at the global mean value of 1, and calculate the uncure rate for $\boldsymbol{z}_1 = 0$ and $\boldsymbol{z}_1 = 1$ separately. The true values for the uncure rate $\pi(\boldsymbol{z}_1 = 0)$ and $\pi(\boldsymbol{z}_1 = 1)$ and right censoring rate are listed in Table 3.3 for each link function.

Table 3.3 Setting: Cure Rate and Right Censoring Rate

Link	$\pi(\boldsymbol{z}_1 = 0)$	$\pi(\boldsymbol{z}_1 = 1)$	Right Cen.%
Logit	0.269	0.500	0.682
Probit	0.308	0.632	0.728
Cloglog	0.159	0.500	0.582

We generate data with one of the link functions, for example logit link, and fit the data using all the three links separately. We calculate the uncure rate $\pi(\boldsymbol{z}_1 = 0)$ and $\pi(\boldsymbol{z}_1 = 1)$, as well as their variance estimates based on delta method. The bias, average estimated standard deviation (StErr), empirical standard deviation (StDev) and empirical coverage probability (CP) for uncure rates and coefficients for the latency part are reported in Table 3.4. From the results, we find the proposed multiple

imputation approaches are pretty stable in estimating the uncure rate when we specify a different link function for the model regarding to biases. Variances for the latency part are close to 95%.

Table 3.4 Sensitivity Analysis for Different Link Functions

	Logit				**Probit**				**Cloglog**			
par.	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP	Bias	StDev	StErr	CP
Logit												
$\pi(\boldsymbol{z}_1 = 0)$	0.000	0.034	0.034	0.946	0.001	0.033	0.033	0.950	-0.001	0.032	0.032	0.948
$\pi(\boldsymbol{z}_1 = 1)$	-0.002	0.042	0.041	0.938	-0.005	0.042	0.040	0.934	-0.014	0.041	0.040	0.918
β_1	-0.021	0.203	0.200	0.920	-0.028	0.206	0.199	0.924	-0.021	0.202	0.200	0.938
β_2	0.035	0.232	0.215	0.938	0.038	0.235	0.215	0.914	0.038	0.235	0.215	0.918
Probit												
$\pi(\boldsymbol{z}_1 = 0)$	-0.003	0.028	0.027	0.922	-0.002	0.029	0.028	0.918	-0.003	0.035	0.034	0.934
$\pi(\boldsymbol{z}_1 = 1)$	0.000	0.048	0.044	0.928	-0.005	0.043	0.041	0.936	-0.007	0.043	0.041	0.936
β_1	-0.034	0.212	0.211	0.946	-0.039	0.216	0.209	0.936	-0.026	0.166	0.173	0.942
β_2	0.022	0.255	0.225	0.916	0.031	0.254	0.227	0.914	0.024	0.194	0.187	0.936
Cloglog												
$\pi(\boldsymbol{z}_1 = 0)$	0.000	0.038	0.036	0.942	0.004	0.037	0.035	0.932	-0.003	0.035	0.034	0.934
$\pi(\boldsymbol{z}_1 = 1)$	0.008	0.041	0.041	0.934	0.003	0.041	0.039	0.934	-0.007	0.043	0.041	0.936
β_1	-0.030	0.167	0.173	0.948	-0.032	0.168	0.174	0.950	-0.026	0.166	0.173	0.942
β_2	0.025	0.195	0.188	0.934	0.028	0.197	0.188	0.928	0.024	0.194	0.187	0.936

3.3 Sensitivity Analysis: interval and rounded right censoring

Since day is a confidential variable to some U.S. registries, only the processed survival month is available for analysis in the SEER data [21]. The definition of survival months after diagnosis (T) is

$$T = \text{floor}\Big(\frac{\text{last contact date} - \text{diagnosis date}}{\text{average days in a month}}\Big),$$

where floor$(\cdot)$ denotes a function rounding down to the integer. Therefore, even for the complete observations, the exact survival month cannot be read directly from the SEER data. The survival month after diagnosis T indicates that the exact event

time happens in the interval $[T, T+1)$, which can be viewed as an interval censored survival data.

Here, we perform a simulation study to mimic the rounding down procedure and compare the continuous approach for right censored data and our proposed method for interval censored data. Covariates and coefficients for the cure rate part and the latency part are set to be the same as that in Section 3.1, baseline distribution for survival time X is Weibull with shape parameter 1 and scale parameter 200. Based on these, the cure rate is about 60% and the mean survival time is around 430. Censoring time C follows Uniform distribution from 0 to 2000, leading to an average right censoring rate of 68%. Sample size of 500 is used. We denote this data set as the Original Dataset $\boldsymbol{D} = \{T, \delta, \boldsymbol{x}, \boldsymbol{z}\}$, where $T = min(X, C)$ and $\delta = 1_{(X \leq C)}$. We divide T by a unit l and round down to the integer like Equation 3.3, that is, let $T^{(1)} = floor(T/l)$, we have the data set with rounded down survival times $\boldsymbol{D}^{(1)} = \{T^{(1)}, \delta, \boldsymbol{x}, \boldsymbol{z}\}$. As we claimed that the interval censoring data set is $\boldsymbol{D}^{(2)} = \{L, R, \boldsymbol{x}, \boldsymbol{z}\}$, where $L = T^{(1)}$ and $R = T^{(1)} + 1$ if $\delta = 1$, otherwise $R = \infty$.

The unit is chosen to be weekly ($l = 7$), monthly ($l = 30$) and quarterly ($l = 90$). Right censored data sets $\boldsymbol{D}$ and $\boldsymbol{D}^{(1)}$ are fitted using the “smcure" package in R, and the interval censored data $\boldsymbol{D}^{(2)}$ is fitted using our proposed algorithm based on ANDA. Biases and Coverage probabilities (CP) are reported. The estimated baseline survival curves are compared with the true distribution from Figure 3.5 to 3.7.

Based on the results, the “smcure" function for the rounded down right censoring data set $\boldsymbol{D}^{(1)}$ is sensitive to the unit l. As the unit l increases, biases of the two β coefficients in the latency part increases, and coverage probabilities decreases. But if we treat the data set as interval censored like $\boldsymbol{D}^{(2)}$, and fit with our proposed method,

Table 3.5 Estimates for Rounded Down Survival Times

	smcure for $\boldsymbol{D}$		**smcure for $\boldsymbol{D}^{(1)}$**		**ANDA for $\boldsymbol{D}^{(2)}$**	
par.	Bias	CP	Bias	CP	Bias	CP
$l =$ **7**						
γ_0	-0.038	0.940	-0.036	0.935	-0.029	0.940
γ_1	-0.030	0.890	-0.029	0.905	-0.028	0.905
γ_2	0.004	0.950	0.003	0.940	0.002	0.950
β_1	-0.060	0.965	-0.076	0.950	-0.072	0.935
β_2	0.031	0.955	0.046	0.950	0.033	0.960
$l =$ **30**						
γ_0	-0.029	0.945	-0.019	0.935	-0.012	0.970
γ_1	-0.032	0.950	-0.027	0.960	-0.028	0.960
γ_2	0.017	0.970	0.013	0.955	0.013	0.955
β_1	-0.036	0.950	-0.101	0.910	-0.050	0.925
β_2	0.045	0.950	0.112	0.890	0.049	0.950
$l =$ **90**						
γ_0	-0.018	0.930	0.040	0.925	0.034	0.945
γ_1	0.005	0.945	0.028	0.930	0.019	0.940
γ_2	-0.010	0.930	-0.034	0.935	-0.021	0.925
β_1	-0.031	0.950	-0.227	0.740	-0.056	0.935
β_2	0.017	0.975	0.221	0.685	0.045	0.940

we have smaller biases and the coverage probabilities are close to 95%. We also compared the estimated baseline curves from these three approaches, the red line is the true distribution, the black line is from "smcure" for $\boldsymbol{D}$, the blue line is from ANDA for $\boldsymbol{D}^{(2)}$. Both of these two step functions are close to the truth under all settings. However, the estimated baseline curves for rounded down right censored times (green lines) get far away from the truth as the unit l goes up.

Therefore, we can increase the accuracy of estimates by treating the rounded down right censored data as interval censored data in the mixture cure model. And the method will be applied in the SEER Breast Cancer data in Chapter 4.

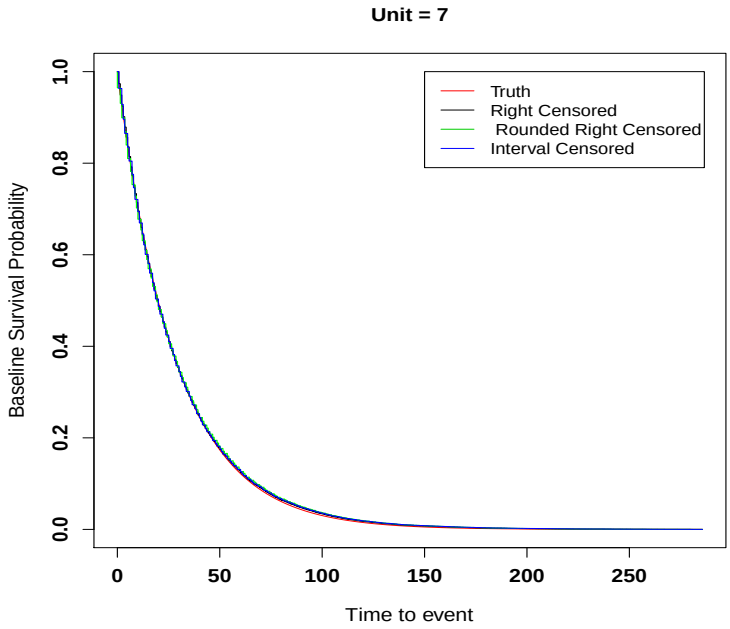

Figure 3.5 Estimated Baseline Survival Probability for Sensitivity Analysis (Weekly)

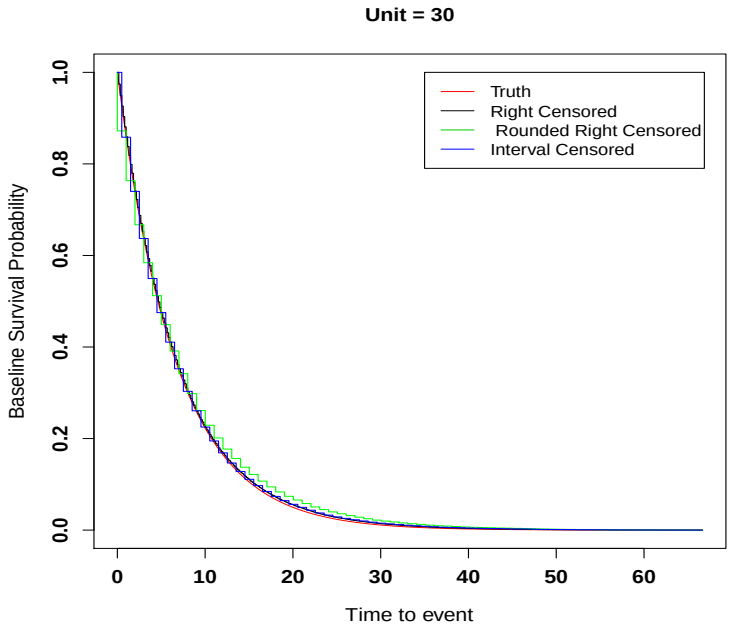

Figure 3.6 Estimated Baseline Survival Probability for Sensitivity Analysis (Monthly)

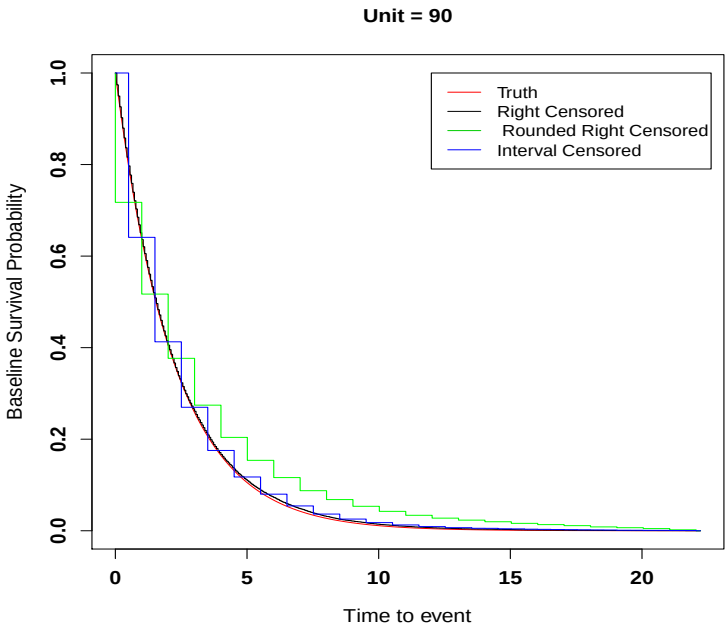

Figure 3.7 Estimated Baseline Survival Probability for Sensitivity Analysis (Quarterly)

Chapter 4

Breast Cancer Data

Breast cancer is the most common cancer in women worldwide, and approximately 12.3 percent of women will be diagnosed with breast cancer at some point during their lifetime, based on 2008-2010 Surveillance, Epidemiology, and End Results (SEER) data [20].

Cancer stage at diagnosis, which characterizes the extent of cancer in the body, determines treatment options and has a strong influence on the length of survival. It has five main categories for cancer cases. In situ refers to the case where abnormal cells are present only in the layer of cells in which they developed. Localized means the cancer is limited to the organ in which it began, without evidence of spread. In the regional case, cancer has spread beyond the primary site to nearby lymph nodes or tissues and organs. For distant stage, cancer has spread from the primary site to distant tissues or organs or to distant lymph nodes. And if There is not enough information to determine the stage, the cancer is unstaged. For example, in breast cancer cases, the 5-year survival probability for localized breast cancer is as high as 98.5%.

4.1 Dataset Description

For the purpose of illustration, we extracted 2000-2010 Greater Georgia breast cancer dataset from the SEER cancer incidence public-use database [21]. We choose

Greater Georgia because we want to adjust the effect of race on survival probabilities, and we have relatively high proportion for African Americans in this state, approximately 17.6% of the population.

Variables of interest include: age at diagnosis, marital status (single, married or other), SEER summary stage (in situ, local, reginal or distant), race (white, black or other) and survival months after diagnosis. Observations with missing values for any of these variables are excluded from this study. Among these variables, the main covariate of interest is the stages of cancer.

Since there are only 54 subjects with other race category, we also exclude these observations. Subjects with recurrent events are not under our consideration and were excluded as well. Subjects who lost to follow-up after diagnosis or with incomplete or unknown survival time were also deleted.

As a result, we have a total of 7,249 patients; 394 of them have tumor in situ, 4,182 of them have local cancer stage, 2,458 of them have reginal cancer stage and 215 have distant cancer stage. We have 86.3% of the observations that are subject to right censoring, such a high right censoring proportion suggests that a proportion of these patients may be cured.

4.2 Preliminary Data Analysis

As we mentioned in 3.3, the survival month T was calculated based on the last contact date and the diagnosis date, and rounded down to the integer. Thus we modify the derived follow-up survival month T into time intervals. For example, the true survival time for subject j is within $[T_j, T_j + 1)$ months, where T_j was the derived survival month in the dataset. We draw the Turnbull [24] NPMLE fitted survival

curves with respect to the four breast cancer stages (Fig. 4.1). The curves show that the survival rate decreases when the stage becomes severe. Furthermore, there is a leveling off of survival curves at the end of study, which indicates a possible cure rate in breast cancer. As we claimed before, by treating the data as interval censored, we

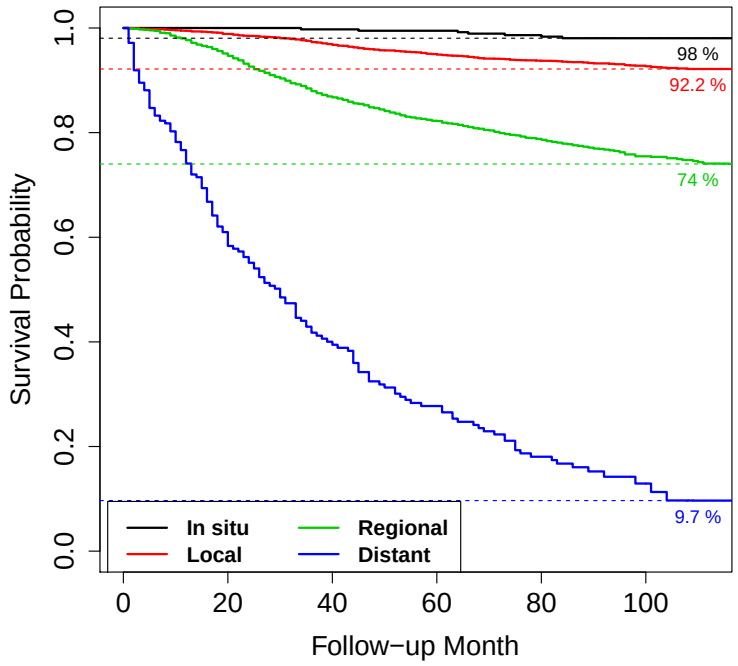

Figure 4.1 Turnbull NPMLE for Different Cancer Stages

can increase the accuracy of the estimates in the mixture cure model. Therefore, we model the interval censored survival times with the stage variable using the proposed multiple imputation algorithm based on ANDA, and the results are listed in Table 4.1. Cure rate and survival probability of uncured patients are compared among the four cancer stages: in situ, local, regional and distant. Survival probabilities for the four stages are calculated and compared with the results from Turnbull [24] NPMLE in figure 4.2. From the plot, we can see that our results are pretty close to the curves based on the NPMLE.

Table 4.1 Crude Model for SEER Breast Cancer Study

	Crude Model		
Parameters	Est	Std	P
Incidence part			
Intercept	-3.169	0.347	<.001
Stage : *Local*	0.798	0.353	0.024
Stage : *Regional*	2.199	0.349	<.001
Stage : *Distant*	5.366	0.542	<.001
Latency part			
Stage : *Local*	0.945	0.535	0.077
Stage : *Regional*	1.150	0.525	0.029
Stage : *Distant*	1.762	0.470	<.001

The stage in situ is treated as baselines in the model. Based on the results, we can derive cure probabilities for different stages and hazard ratios between two of the stages for uncured patients. For example, the cure rate for in situ, local, regional and distant stages are 96.0%, 91.5%, 72.5% and 10.0% respectively, and the hazard ratio between regional stage and stage in situ for uncured patients is 3.16 with 95% confidence interval (1.128, 8.837). Therefore, as the cancer progresses from stage in situ to distant, cure probabilities decreases and hazard risk for uncured patients increases.

4.3 Adjusted Model

We apply our multiple imputation method based on ANDA to the breast cancer dataset described above. Cure rate and survival probability of uncured patients are compared among the four cancer stages after adjusted for age at diagnosis, race and marital status at diagnosis.

We use $m = 10$ imputations in our MI-ANDA algorithm. The estimated coeffi-

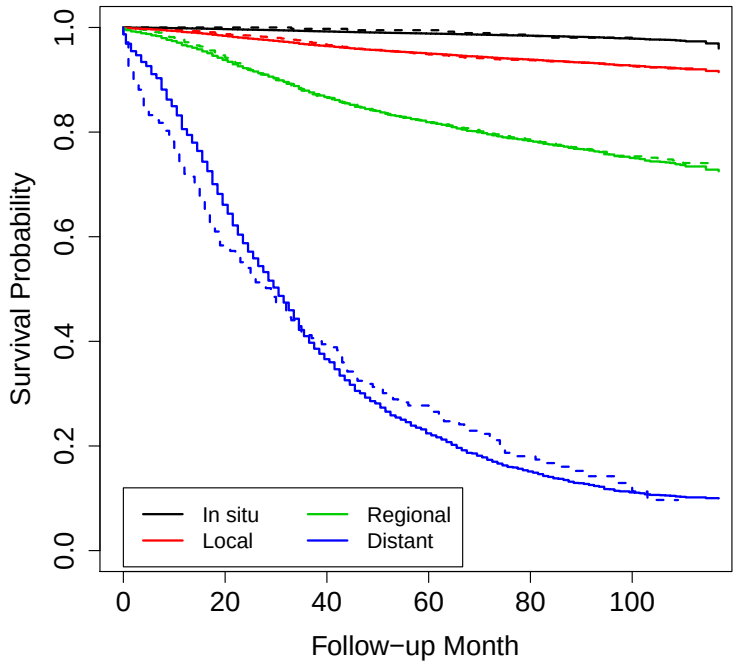

Figure 4.2 Estimated Survival Curves in Crude Model(dotted line for Turnbull and solid line for MI-ANDA)

cients and standard deviations are listed in Table 4.2. The stage in situ, other marital status and race of black were treated as baselines in the model. Based on the results, we can derive cure probabilities for different stages and hazard ratios between two of the stages for uncured patients adjusting for the other covariates. For example, the cure rate for in situ, local, regional and distant stages are 95.5%, 89.9%, 69.1% and 8.5% respectively, and the hazard ratio between regional stage and stage in situ for uncured patients is 3.09 with 95% confidence interval (0.970, 9.839).

We also have some interesting findings for these predictor variables. In the adjusted model, as the cancer progresses from stage in situ to distant, cure probabilities decreases and hazard risk for uncured patients increases. Cure rate is not significantly different for people at different age but the survival probability of uncured patients

Table 4.2 Adjusted Model for SEER Breast Cancer Study

	Adjusted Model		
Parameters	Est	Std	P
Incidence part			
Intercept	-3.050	0.438	<.001
Stage : *Local*	0.866	0.459	0.059
Stage : *Regional*	2.245	0.466	<.001
Stage : *Distant*	5.426	0.640	<.001
Age	0.003	0.003	0.391
Marital : *Single*	0.238	0.139	0.086
Marital : *Married*	-0.284	0.108	0.008
Race : *White*	-0.388	0.094	<.001
Latency part			
Stage : *Local*	0.914	0.584	0.117
Stage : *Regional*	1.128	0.591	0.056
Stage : *Distant*	1.606	0.580	0.006
Age	0.007	0.003	0.013
Marital : *Single*	0.035	0.134	0.795
Marital : *Married*	-0.066	0.099	0.504
Race : *White*	-0.146	0.089	0.100

is actually affected by age. Women with different marital status have significantly different cure rate, but survival probabilities for uncured patients are similar. White women have significantly higher cure rate compared with black women, however, for uncured patients, the difference between survival probabilities for different races are not significant.

Predicted survival curves for the three stages in the adjusted model are in Figure 4.3, where age, marital status and race are fixed at median values. Based on the estimated overall survival curves, we can conclude that the local stage have a highest survival probability, while the distant stage actually have the lowest survival probability, and the regional stage is in between them.

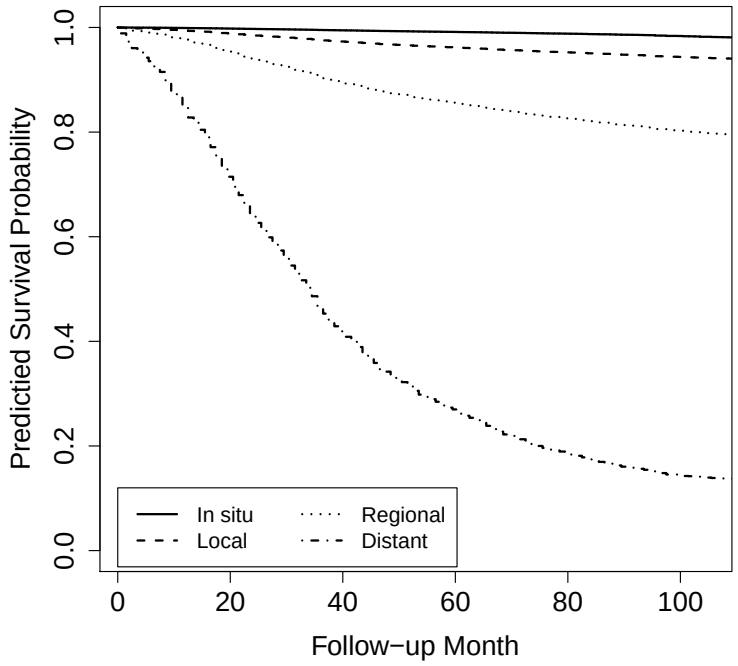

Figure 4.3 Predicted Survival Curves in Adjusted Model

Chapter 5

Discussion

In our paper, we developed a multiple imputation algorithm for interval censored data under the proportional hazards mixture cure model. Our algorithm is based on the *Asymptotic normal data augmentation* (ANDA) or the *Poor Man's data augmentation* (PMDA). Based on the simulation results, we find these two approaches are close to each other when we have relatively small right censoring rate for uncured patients. However, when we have large proportion of right censoring that could not be explained by the incidence part, the algorithm based on PMDA will underestimate the variance.

From the sensitivity analysis for link functions in the cure rate part, we find the estimates for uncure rate and coefficients in latency part are quite stable. Therefore, we do not need to worry about which link function to choose for the incidence part when we apply the model to our data in practice. We also proved in another sensitivity analysis that by treating the rounded right censored survival months as interval censored data, and fit with the proposed method, we can increase the accuracy of estimates in the model.

The method was further applied to the 2000-2010 Greater Georgia SEER Breast cancer study to compare different cancer stages under the mixture cure model. The results are easily interpreted based on the cure probability for different cancer stages and hazard ratio in different stages.

One of the greatest advantages of this method is the easy application in softwares. We can use some well established softwares such as R and SAS to implement the algorithm. In R, the “glm" function can be used for the incidence part and the “coxph" function in the “survival" package can be used for the latency part. In SAS, the “GLM" and “PHREG" procedures can be used. Since these existing softwares are efficient in calculating both of the estimates and the variance needed in the algorithm, and the imputation number m usually is small, our proposed algorithm is relatively fast to obtain the estimates.

Bibliography

[1] John W Boag. "Maximum likelihood estimates of the proportion of patients cured by cancer therapy". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 11.1 (1949), pp. 15–53.

[2] Chao Cai et al. "smcure: An R-Package for estimating semiparametric mixture cure models". In: *Computer methods and programs in biomedicine* 108.3 (2012), pp. 1255–1260.

[3] David R Cox et al. "Regression models and life tables". In: *JR stat soc B* 34.2 (1972), pp. 187–220.

[4] Dianne M Finkelstein and Robert A Wolfe. "A semiparametric model for regression analysis of interval-censored failure time data." In: *Biometrics* 41.4 (1985), pp. 933–945.

[5] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*. Vol. 19. Springer, 1992.

[6] David G Hoel and HE Walburg. "Statistical analysis of survival experiments". In: *Journal of the National Cancer Institute* 49.2 (1972), pp. 361–372.

[7] Jong S Kim. "Efficient estimation for the proportional hazards model with left-truncated and" case 1" interval-censored data". In: *Statistica Sinica* 13.2 (2003), pp. 519–538.

[8] Yang-Jin Kim and Myoungshic Jhun. "Cure rate model with interval censored data". In: *Statistics in medicine* 27.1 (2008), pp. 3–14.

[9] KF Lam, Daniel YT Fong, and OY Tang. "Estimating the proportion of cured patients in a censored sample". In: *Statistics in medicine* 24.12 (2005), pp. 1865–1879.

[10] Kwok Fai Lam, Kin Yau Wong, and Feifei Zhou. "A semiparametric cure model for interval-censored data". In: *Biometrical Journal* 55.5 (2013), pp. 771–788.

[11] Shuangge Ma. "Cure model with current status data". In: *Statistica Sinica* 19.1 (2009), p. 233.

[12] Wei Pan. "A Multiple Imputation Approach to Cox Regression with Interval-Censored Data". In: *Biometrics* 56.1 (2000), pp. 199–203.

[13] Wei Pan. "Extending the iterative convex minorant algorithm to the Cox model for interval-censored data". In: *Journal of Computational and Graphical Statistics* 8.1 (1999), pp. 109–120.

[14] Yingwei Peng. "Fitting semiparametric cure models". In: *Computational statistics & data analysis* 41.3 (2003), pp. 481–490.

[15] Dionne L Price and Amita K Manatunga. "Modelling survival data with a cured fraction using frailty models". In: *Statistics in medicine* 20.9-10 (2001), pp. 1515–1527.

[16] R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2008. URL: `http://www.R-project.org`.

[17] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. Vol. 307. John Wiley & Sons, 2009.

[18] Nathaniel Schenker and AH Welsh. "Asymptotic results for multiple imputation". In: *The Annals of Statistics* (1988), pp. 1550–1566.

[19] Jianguo Sun. *The statistical analysis of interval-censored failure time data*. Vol. 2. Springer, 2006.

[20] Surveillance, Epidemiology, and End Results (SEER) Program. *SEER Stat Fact Sheets:Breast Cancer*. `http://seer.cancer.gov/statfacts/html/breast.html`. Online; accessed: 2014-04-29. 2008-2010.

[21] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2010). *National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2013, based on the November 2012 submission.* 1973-2010.

[22] Martin A Tanner and Wing Hung Wong. "An application of imputation to an estimation problem in grouped lifetime analysis". In: *Technometrics* 29.1 (1987), pp. 23–32.

[23] Jeremy MG Taylor. "Semi-parametric estimation in failure time mixture models". In: *Biometrics* (1995), pp. 899–907.

[24] Bruce W Turnbull. "The empirical distribution function with arbitrarily grouped, censored and truncated data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1976), pp. 290–295.

[25] GC Wei and Martin A Tanner. "Applications of multiple imputation to the analysis of censored regression data." In: *Biometrics* 47.4 (1991), pp. 1297–1309.

Appendix A

Source Codes

```
# Generating Interval censored data in PHMC model
iccure.data<-function(N,beta0,gamma0,my=1,len=.5,nk=1,a=2,b=1,xx,zz,
baseline="weibull",link="logit"){
if(link=="logit") temp<-exp(cbind(1,zz)%*%gamma0)/
(1+exp(cbind(1,zz)%*%gamma0))
if(link=="probit") temp<-pnorm(cbind(1,zz)%*%gamma0)
if(link=="cloglog") temp<-1-exp(-exp(cbind(1,zz)%*%gamma0))
ui<-rbinom(N,1,temp)

# Time from PH model with weibull baseline dist.
xb<-xx%*%beta0
ft<-1-exp(log(runif(N))/exp(xb))
if(baseline=="weibull") t.true<-qweibull(ft,shape=a,scale=b)
if(baseline=="lognormal") t.true<-qlnorm(ft,meanlog=a,sdlog=b)
#if(baseline=="llogistic") t.true<-qllog(ft,shape=a,scale=b)
t.obs<-runif(N,0,my)

L<-t.true<t.obs & ui==1 # left censor indicator
M<-!(t.true<t.obs | t.true>t.obs+nk*len) & ui==1
R<-t.true>t.obs+nk*len | ui==0 # right censored indicator
```

```
t.left<-t.right<-t.obs
t.left[L]<-0

t.left[M]<-t.obs[M]+floor((t.true[M]-t.obs[M])/len)*len
t.right[M]<-t.obs[M]+ceiling((t.true[M]-t.obs[M])/len)*len

t.left[R]<-t.obs[R]+nk*len
t.right[R]<-NA

status<-as.numeric(!R)
t.time<-t.left+(t.right-t.left)/2
t.time[R]<-t.left[R]
mydata<-data.frame(left=t.left,right=t.right,deltaL=L,deltaI=M,
    deltaR=R,time=t.time,true=t.true,status=status,xx,zz,ui)
return(mydata)
}

# MI-ANDA function
library(splines)
library(survival)
library(MASS)
ANDA.initial<-function(formula,data,xx,zz,linkfun){
# M-step1: glm
nz<-ncol(zz)
tempf<-paste("z",1:nz,sep="",collapse="+")
formu<-paste("status~",tempf)
fit1<-glm(formu,data,family=binomial(link = linkfun))
```

```
est.gam0<-fit1$coefficients
est.varg0<-vcov(fit1)

# M-step2: Cox PH model
nx<-ncol(xx)
sdata<-subset(data,data$status==1)
      formula.covar <- formula[[3]]
      formula <- as.formula(Surv(time,status) ~ .)
      formula[[3]] <- formula.covar
fit2<-coxph(formula, sdata)
est.beta0<-coef(fit2)
est.varb0<-vcov(fit2)

newdf<-data.frame(matrix(rep(0,nx),ncol=nx))
colnames(newdf)<-all.vars(formula)[-c(1:2)]
surv0<-summary(survfit(fit2,newdata=newdf))
BaseSurv<-cbind(c(0,surv0$time),c(1,surv0$surv))

list(Beta=est.beta0,Var.b=est.varb0,Gamma=est.gam0,Var.g=est.varg0,
BaseSurv=BaseSurv)
}

add <- function(x) Reduce("+", x)

ANDAstep<-function(formula,data,Beta,Gamma,Var.b,Var.g,BaseSurv,xx,
zz,linkfun,m){
  nz<-ncol(zz)
```

```
  nx<-ncol(xx)
  temp.gam<-matrix(ncol=nz+1,nrow=m)
  temp.beta<-matrix(ncol=nx,nrow=m)
  temp.varg<-temp.varb<-sfun<-list()
  temp.t<-numeric()

  data$Ti<-data$time

  gam0<-mvrnorm(n=m,mu=Gamma,Sigma=Var.g)
  beta0<-mvrnorm(n=m,mu=Beta,Sigma=Var.b)

for(k in 1:m){
  # E-step
  if(linkfun=="logit") pi_z<-exp(cbind(1,zz)%*%gam0[k,])/
      (1+exp(cbind(1,zz)%*%gam0[k,]))
  if(linkfun=="probit") pi_z<-pnorm(cbind(1,zz)%*%gam0[k,])
  if(linkfun=="cloglog") pi_z<-1-exp(-exp(cbind(1,zz)%*%gam0[k,]))

  estb<-beta0[k,]
  exb<-exp(xx%*%estb)
  su0<-stepfun(BaseSurv[-1,1],BaseSurv[,2])(data$time)
  surv<-su0^exb
  wi<-data$status+((1-data$status)*pi_z*surv)/(1-pi_z+pi_z*surv)

  # M-step1: glm
  data$y<-rbinom(length(wi),1,wi)
  tempf<-paste("z",1:nz,sep="",collapse="+")
```

```
formu<-paste("y~",tempf)
fit1<-glm(formu,data,family=binomial(link = linkfun))
temp.gam[k,]<-fit1$coefficients
temp.varg[[k]]<-vcov(fit1)

# M-step2: Cox PH model
sdata<-subset(data,data$y==1)
sxx<-subset(xx,data$y==1)

# conditional sampling Ti from Sj
for(j in 1:nrow(sdata)){
  if(sdata$status[j]==1){
  Sj<-cbind(BaseSurv[,1],BaseSurv[,2]^exp(sxx[j,]%*%estb))
  sub.s0<-Sj[Sj[,1]>=sdata$left[j] & Sj[,1]<=sdata$right[j],]
  if(length(dim(sub.s0))>1){
  prob<-c(0,diff(1-sub.s0[,2]))/sum(diff(1-sub.s0[,2]))
  sdata$Ti[j]<-sample(sub.s0[,1],1,prob=prob)
  } else sdata$Ti[j]<-sub.s0[1]
  }
}

      formula.covar <- formula[[3]]
      formula <- as.formula(Surv(Ti,status) ~ .)
      formula[[3]] <- formula.covar
fit2<-coxph(formula, sdata)
temp.beta[k,]<-coef(fit2)
temp.varb[[k]]<-vcov(fit2)
```

```
  newdf<-data.frame(matrix(rep(0,nx),ncol=nx))
  colnames(newdf)<-all.vars(formula)[-c(1:2)]
  surv0<-summary(survfit(fit2,newdata=newdf))
  sfun[[k]]<-stepfun(surv0$time,c(1,surv0$surv))
  temp.t<-c(temp.t,surv0$time)
}

estgam<-apply(na.omit(temp.gam),2,mean) # Estimated gamma coef.
estb<-apply(na.omit(temp.beta),2,mean) # Estimated beta coef.

mgam<-na.omit(temp.gam)-matrix(estgam,nrow=m,ncol=nz+1,byrow=TRUE)
est.varg<-add(temp.varg) / length(temp.varg)+(1+1/m)*t(mgam)
%*%mgam/(m-1)
mbeta<-na.omit(temp.beta)-matrix(estb,nrow=m,ncol=nx,byrow=TRUE)
est.varb<-add(temp.varb) / length(temp.varb)+(1+1/m)*t(mbeta)
%*%mbeta/(m-1)

uniq.t<-sort(unique(c(0,temp.t)))
temp.s0<-matrix(ncol=0,nrow=length(uniq.t))
for(kk in 1:m){
  temp.s0<-cbind(temp.s0,sfun[[kk]](uniq.t))
}
est.s0<-apply(temp.s0,1,mean)
BaseSurv<-rbind(cbind(uniq.t,est.s0),c(max(na.omit(data$right))
,0))
```

```
return(list(Gamma=estgam,Beta=estb,Var.g=est.varg,Var.b=est.varb,
BaseSurv=BaseSurv))
}

MI_ANDA<-function(formula = formula(data), curefun= formula(data),
data = parent.frame(),linkfun="logit",max.iter=100,cov.rate=.01,
m=10){
    copy.data <- data
    call <- match.call()
    mf <- match.call(expand.dots = FALSE)
    temp <- c("", "formula", "data", "copy.data", "na.action")
    mf <- mf[match(temp, names(mf), nomatch = 0)]
    mf[[1]] <- as.name("model.frame")
    temp.x<-terms(formula, data = copy.data)
    temp.z<-terms(curefun, data = copy.data)
    mf$formula <- temp.x
    mf <- eval(mf, parent.frame())
    Y <- model.extract(mf, "response")

    attr(temp.x, "intercept") <- 0
    attr(temp.z, "intercept") <- 0

    xx <- model.matrix(temp.x,copy.data)
    colnames(xx)<-paste("x",1:ncol(xx),sep="")
    zz <- model.matrix(temp.z,copy.data)
    colnames(zz)<-paste("z",1:ncol(zz),sep="")
```

```
    new.time<-new.left<-Y[,1]
    new.right<-Y[,2]
    new.status<-Y[,3]/3
    DEAD<-new.status==1
    new.right[!DEAD]<-NA
    new.time[DEAD]<-new.left[DEAD]+(new.right[DEAD]-new.left[DEAD])/2
    new.data<-data.frame(left=new.left,right=new.right,time=new.time,
    status=new.status,xx,zz)
    formula1<-as.formula(paste("Surv(",formula[[2]][2],",",
    formula[[2]][3],")~",
    paste(colnames(xx),
    sep="",collapse="+"),sep=""))

  step0<-ANDA.initial(formula1,data=new.data,xx=xx,zz=zz,
  linkfun=linkfun)
  Beta0<-step0$Beta
  Gamma0<-step0$Gamma
  Var.b0<-step0$Var.b
  Var.g0<-step0$Var.g
  BaseSurv0<-step0$BaseSurv

  dif<-numeric()
  est.par<-est.var<-matrix(ncol=ncol(xx)+ncol(zz)+1,nrow=max.iter)
  dd<-1
  ii<-0
  while(dd>cov.rate & ii<max.iter){
ii<-ii+1
```

```
temp<-try(ANDAstep(formula1,data=new.data,Beta=Beta0,Gamma=Gamma0,
    Var.g=Var.g0,Var.b=Var.b0, BaseSurv=BaseSurv0,xx=xx,zz=zz,linkfun=
    linkfun,m=m),silent=TRUE)
while(is.character(temp)){
temp<-try(ANDAstep(formula1,data=new.data,Beta=Beta0,Gamma=Gamma0,
    Var.g=Var.g0,Var.b=Var.b0, BaseSurv=BaseSurv0,xx=xx,zz=zz,linkfun=
    linkfun,m=m),silent=TRUE)
}


if(!is.character(temp)){
  est.par[ii,]<-c(temp$Gamma,temp$Beta)
  est.var[ii,]<-c(diag(temp$Var.g),diag(temp$Var.b))
  Beta0<-temp$Beta
  Gamma0<-temp$Gamma
  Var.b0<-temp$Var.b
    Var.g0<-temp$Var.g
  BaseSurv0<-temp$BaseSurv
} else break
if(ii>1){
  dd<-max((est.par[ii,]-est.par[(ii-1),])^2)
  dif<-c(dif,dd)
}
  }
  Gamma<-temp$Gamma
  Var.g<-diag(temp$Var.g)
  names(Gamma)<-names(Var.g)<-c("Intercept",all.vars(curefun)[-1])
  Beta<-temp$Beta
```

```
  Var.b<-diag(temp$Var.b)
  names(Beta)<-names(Var.b)<-all.vars(formula)[-c(1:2)]
  covg<-ii<max.iter
return(list(Gamma=Gamma,Beta=Beta,Var.g=Var.g,Var.b=Var.b,
Cov.g=temp$Var.g,Cov.b=temp$Var.b,BaseSurv0=BaseSurv0,converge=covg,
n.iter=ii))

# Example
N<-500
m<-10
beta0<-c(1,-1) #parameters for cox model
gamma0<-c(0,1,-1) #parameters for cure prob.
n.simu=200
xx<-cbind(runif(N,0,2),rbinom(N,1,.5))
colnames(xx)<-c("x1","x2")
zz<-cbind(runif(N,0,2),rbinom(N,1,.5))
colnames(zz)<-c("z1","z2")
sdata<-iccure.data(N,beta0,gamma0,my=1,len=.5,nk=1,a=0,b=1,xx,zz,
baseline="lognormal",link="logit")
temp.em1<-MI_ANDA(formula = Surv(left,right,type="interval2")~ x1+x2,
curefun= status~ z1, data =sdata, linkfun="logit",max.iter=100,
cov.rate=1e-3,m=m),silent=TRUE)
}
```