Titre: Title:	Cure Rate Models
Auteur: Author:	Mahroo Vahidpour
Date:	2016
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Vahidpour, M. (2016). Cure Rate Models [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/2454/

UNIVERSITÉ DE MONTRÉAL

CURE RATE MODELS

MAHROO VAHIDPOUR
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(MATHÉMATIQUES APPLIQUÉES)
DÉCEMBRE 2016

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

CURE RATE MODELS

présenté par : VAHIDPOUR Mahroo
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. LEFEBVRE Mario, Ph. D., président
M. ADJENGUE Luc-Désiré, Ph. D., membre et directeur de recherche
M. ASGHARIAN-DASTENAEI Masoud, Ph. D., membre et codirecteur de recherche
Mme ATHERTON Juli , Ph. D., membre

DEDICATION

To my lovely family,
you are always in my heart...

ACKNOWLEDGMENTS

It has been a period of intense learning during my research for masters, not only in the scientific arena, but also on a personal level. Writing this thesis has had a big impact on my training. I would like to reflect on the people who have supported and helped me throughout this period.

First, I would like to thank my thesis advisers Prof. Luc Adjengue from Mathematics and Industrial Engineering Department at Polythechnique University, and Prof. Masoud Asgharian from faculty of Mathematics and Statistics Department at McGill University. My supervisors consistently allowed this thesis to be my own work, but steered me in the correct direction whenever they thought I needed it.

Besides my advisers, I would particularly like to thank Prof. Vahid Partovi Nia for dedication of his time to revise this thesis and give me many critical advices. He supported me greatly and was always willing to help. I would also, here by give my many thanks to the rest of my thesis committee Prof. Mario Lefebvre from Mathematics and Industrial Engineering Department at Polythechnique University and Prof. Juli Atherton from Mathematics and Statistics Department at UQAM who were involved in the validation survey of this thesis. Without their passionate participation and input, the validation survey could not have been successfully conducted.

My sincere thanks also goes to Mr. Eric Jobin and Mr. Jean-Francois Ross From Pratt and Whitney Canada for offering me the internship opportunities in their groups and leading me working on diverse exciting statistical projects.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of writing this thesis. This accomplishment would not have been possible without their support. Thank you all.

RÉSUMÉ

Les modèles de survie avec taux de guérison ont une vaste gamme d'applications dans de nombreux domaines tels que la médecine et la santé publique, en particulier dans des études sur le cancer. Dans ces études, les chercheurs sont intéressés par le temps d'attente jusqu'à l'apparition d'un événement d'intérêt, ainsi qu'à la proportion des cas où cet événement ne survient jamais, qualifiée dans ce contexte de taux de guérison. D'une manière générale, il existe deux types de modèles pour l'estimation du taux de guérison. Le premier type est le modèle de mélange avec taux de guérison. Ce type de modèles suppose que l'ensemble de la population se compose de deux groupes d'individus : les individus susceptibles de subir l'événement d'intérêt et les individus non susceptibles de subir l'événement d'intérêt, ou immunisés. Le deuxième modèle est un modèle de non-mélange avec taux de guérison qui se base, par exemple, sur le nombre de cellules cancéreuses qui restent après le traitement. Ce mémoire conçoit et présente une revue des modèles de survie avec taux de guérison depuis les premières études jusqu'aux articles récents. Puisqu'il n'y a pas d'étude exhaustive des modèles de taux de guérison, ma mission se limitera au regroupement de tous les modèles dans une notation unique et cohérente.

Les modèles de taux de guérison font partie de l'analyse de survie et incluent les cas de censure. Par conséquent, pour une analyse convenable des modèles de taux de guérison, une bonne connaissance de l'analyse de survie est nécessaire. Ainsi, au chapitre 2, les définitions relatives aux sujets censurés et tronqués sont données. En outre, les concepts et formulations de l'analyse de survie de base sont expliqués. Des modèles de survie courants, paramétriques, non paramétriques et semi-paramétriques sont également expliqués. Les chapitres 3 et 4 forment la partie principale de ce mémoire ; ils portent essentiellement sur l'explication des modèles de taux de guérison. Au chapitre 3, des tests préliminaires pour l'existence d'un taux de guérison sont expliqués. Les premiers travaux dans les modèles de mélange avec taux de guérison et certains modèles paramétriques et non paramétriques avec taux de guérison y sont discutés. L'un des modèles de non-mélange avec taux de guérison et deux modèles de mélange semi-paramétriques qui ont été adaptés des travaux plus récents dans les modèles avec taux de guérison sont expliqués au chapitre 4.

Au chapitre 5, une étude de simulation a été effectuée pour chaque modèle introduit précédemment afin d'en tester la précision. Les méthodes des études de simulation sont les mêmes que celles utilisées dans les articles originaux. Au chapitre 6, les modèles avancés avec taux de guérison sont mis en œuvre pour des bases de données de transplantation de moelle osseuse et

les résultats sont discutés. Le chapitre 7 présente la conclusion de l'étude ainsi qu'un aperçu d'une future recherche.

ABSTRACT

Cure rate models have a broad range of application in many fields, such as medicine, public health, and especially in cancer studies. Researchers in these studies are interested in waiting time until occurrence of the event of interest, as well as the proportion of instances where the event never occurs, known in this case as the cure fraction. In general, there are two types of models for estimation of the cure fraction. The first one is the mixture cure rate model. This type of models assumes that the whole population is composed of two groups of subjects, susceptible subjects and insusceptible (or cured) subjects. The second model is non-mixture cure model, based on number of cancer cells which remain after treatment. This thesis devises and presents a review for cure rate models form early studies to recent articles. Since there is not a comprehensive review on cure rate models, my mission was to put all models together in a single and coherent notation.

Cure rate models are a part of survival analysis and involve censored subjects. Therefore, analyses cure rate models a sufficient knowledge in survival analysis is required. In Chapter 2, the definitions of censored and truncated subjects are given. In addition, the basic concepts and formulations of survival analysis are explained. Some common parametric and non-parametric, and semi-parametric survival models are explained as well.

Chapter 3 and Chapter 4 are the main body of this dissertation, and focus on explaining cure rate models. In Chapter 3, some preliminary tests for existence of cured fraction are described. Early works in mixture cure rate models and some parametric and non-parametric cure models are discussed. In Chapter 4 non-mixture cure rate models and two semi-parametric mixture cure rate models which are adopted from more recent works in cure models are explained.

An extensive simulation study has been done for each model type and is discussed in Chapter 5. Simulation setup is the same as in original article and we were able to reproduce their results. In Chapter 6, advanced cure models are implemented for bone marrow transplantation dataset. Chapter 7 presents the conclusion of the study and an overview of future research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LISTE OF APPENDICES

CHAPTER 1 INTRODUCTION

In many scientific studies, researchers are interested in waiting time until occurrence of an event of interest. For instance, in many clinical trials, the study involves following patients for a period of time and monitoring patient's survival to assess the efficacy of new treatment regimes. The event of interest in such studies could be death, hearth attack, relapse from remission, or adverse reaction. If the event of interest is the heart attack, the *waiting time* would be the time (in years or months) until the heart attack occurs. In survival analysis, such event times are often the outcome of interest. Another example is in reliability studies, where, for example, the repair history of manufactured items might be examined. The question of interest is how long it takes for a manufactured item to be returned by the customer for repair. The outcome variable in such studies is the length of time that a manufactured item functions properly.

When the outcome variable is the waiting time until occurrence of an event the data are called time-to-event data. A particular feature of time-to-event data is censorship. Censoring happens when information about the outcome variable is incomplete. Consider a clinical trial where the event times of some patients are missing due to loss-to-follow-up. It is clear that ignoring such incomplete data can lead to incorrect inference. This is an example of right censoring, one of the most common type of censoring among various censoring mechanisms. Censoring is divided into different types, which are discussed in more details in Chapter 2. Incomplete information is studied in the context of survival analysis. Survival analysis focuses on analyzing time-to-event data and often includes censored observations.

In addition to covering censored individuals in survival models, the response variable is always a positive real number, convenient for time-to-event data. Unlike ordinary statistical models, for example regression models, survival models have two components for response variable: one is the time-to-event and another is an indicator for each individual, whether or not that individual is censored. An essential assumption in survival data analysis is that every individual in the study will eventually experience the event of interest if they are followed long enough. However, the event of interest may not occur for some individuals, even after a very long period of follow-up time. The question is, what should been done for these cases?

Consider the heart attack example, where some of the patients do not die or relapse by the end of the study. Many people experience a heart attack, but recover from this attack and have a normal life for a long while before dying due to other causes is matter of interest. How can we describe this fraction of patients? There are many follow-up studies that include such cases where the standard survival models cannot accurately describe the behaviour of all individuals. This flaw in modelling data with survival methods leads us to look for other models to fit on time-to-event data with long term survivors. The fraction of individuals who do not meet the event of interest even after a long period of follow-up is called *cured fraction*. In the heart attack example, cured fraction consists of those patients who survived by the end of the study and did not show any further sign of heart conduction. In the manufacturing example, those items that did not fail nor malfunction during the examination comprise the cured fraction. Thus, cured fraction is used to refer to any fraction of individuals who never meet the event of interest, regardless of the nature of the study.

Since in real experiments there is always time restriction to follow-up the individuals, usually cured fraction appears in the dataset with censorship at the end point of the study period. Cure rate models focus on modelling cured proportion who survived long enough to be considered as the cured individuals. Additionally, cure rate models concentrate on the probability of survival up to a certain time for those individuals who are not assumed as cured. To model cured fraction we need to modify or extend existing survival models in which those could include another set of parameters to explain nonzero limiting survival probability. Most of the present cure models are modified survival models that include the probability of being cured. There is the possibility of inferring the effect of covariates on cured fraction by assuming a link function to connect the covariates to the probability of being cured. Further discussion about this link function will be given in Chapter 3.

Although cure rate models have a shorter history compared to survival models, they have been an area of active research in statistics since the early 50's. Cure rate models can be divided into two main categories, mixture cure rate models and nonmixture cure rate models. The most widely-used cure rate model is the mixture cure model which is also known as the standard cure rate model. This model was first introduced by Boag (1949). He introduced the traditional definition of cured fraction for the patients who have a specific illness, by five-years survival rate. Although the definition of cured fraction has changed and improved over the years, many authors developed and improved the original mixture model further. Berkson and Gage (1952) divided the population into two groups: susceptible individuals to the event of interest and insusceptible individuals. They suggested that a group of treated

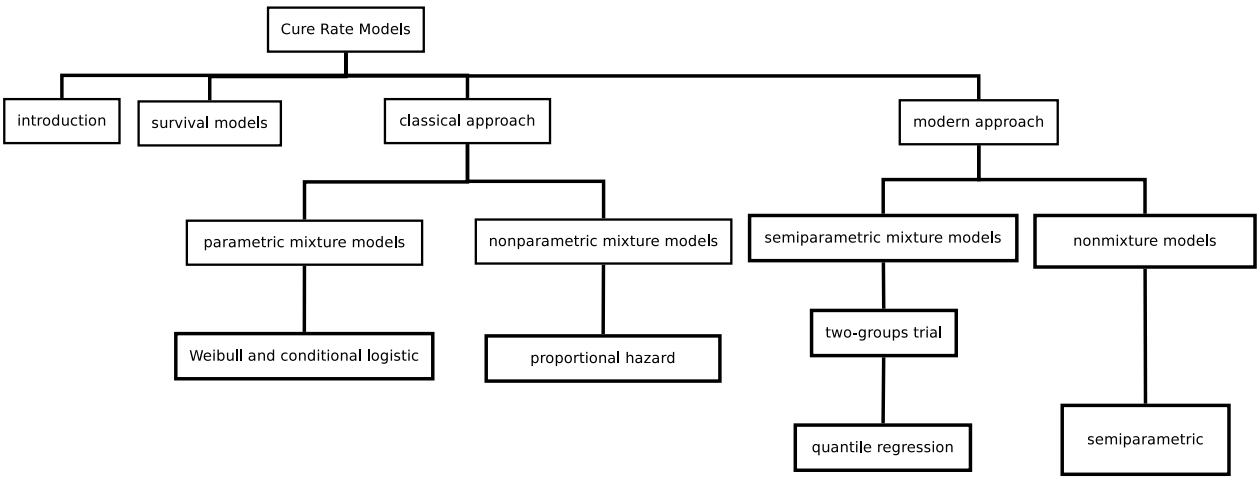

Figure 1.1 The concept diagram of cure rate models.

patients is considered to be cured if they have approximately the same survival distribution as the general population who have never had the disease of interest. Farewell (1982) used mixture model as a combination of logistic model and Weibull distribution to model the toxicant and stress level for laboratory animals. KUK and CHEN (1992) proposed a semi-parametric mixture cure model consisting of the logistic model for the probability of occurrence of the event of interest, and the proportional hazard model to predict time-to-event of interest. Maller and Zhou (1996) have collected a comprehensive account of mixture cure rate models with various survival functions. Goldman (1984), Taylor (1995), and Peng and Dear (2000), among others, have also investigated parametric, semi-parametric, and non-parametric mixture cure rate models. More recently, Wu and Yin (2013) suggested quantile regression methods and martingale estimation equation for a better assessment of covariate effects on quantiles of event of interest.

The non-mixture cure model is another method for modelling time-to-event data with a cure fraction. Non-mixture cure rate models were first introduced by Yakovlev et al. (1993) and then discussed by Ming-Hui Chen (1999), Ibrahim et al. (2001), Chen et al. (2002), and Tsodikov et al. (2003). These models were based on the underlying biological mechanism and the assumption that the number of cancer cells that remain active and may grow after cancer treatment follows a Poisson distribution. Most of the current investigations on the non-mixture cure models are in the Bayesian context due to its special form. Moreover, these two modelling methods, mixture and non-mixture cure rate models, are related, and have meaningful connection together. The non-mixture cure rate models can be transformed into

the mixture cure rate models, if the cured fraction is determined. In the future chapters we will discuss these two models comprehensively.

We present some recent studies to demonstrate the progress in this area over the past decades. The layout of this dissertation is as follow, see Figure 1.1. Chapter 2 introduces basic concepts and primary definitions in survival analysis. In Chapter 3, we introduce statistical inferences to test the presence of cured fraction in a sample of censored time-to-event data. We also explain some selected early cure rate models from mixture. In Chapter 4 three of the more recent mixture and non-mixture cure rate models are discussed. In Chapter 5 we carry out several simulation studies for the models introduced in Chapter 4. In Chapter 6 we apply all three different approaches to a real dataset. The chosen dataset is *bone marrow transplant* dataset used to illustrate survival analysis methods in Klein and Moeschberger (1997). Further discussion about the dataset is given in Chapter 6. In chapter 7 a conclusion of the study and an overview of future research have been presented.

CHAPTER 2 SURVIVAL MODELS

2.1 Introduction

Application of survival analysis abounds in medical and biological studies, among others. In a considerably large portion of studies conducted in such areas, the outcome of interest is time to an event. These types of dataset arise when some subjects are followed during the study period under controlled conditions, to see whether the specific event of interest happens or not. This event could be death or recurrence of a tumor, or discharge from a hospital, or cessation of breastfeeding. This is why survival data are also being referred to as time-to-event, or failure time data.

Survival, failure time or time-to-event data comprise an initiating event, say event of a disease, followed by a terminating event, say death. In most studies where the aim is to collect such data, there are missing or partial information about either the initiating or the terminating event or both. In retrospective studies, for instance, ascertainment of the initiating event may not be possible with the desired accuracy. In prospective studies, the terminating event may not be observed for some subjects. In cross-sectional sampling with follow-up studies both cases can happen. In this later type of studies a further complication known as biased sampling where the collected data do not form a representative sample from the target population, may also happen. Such complications fall into two general categories: censoring and truncation. Censoring is generally reserved for situation where only partial information on some subjects under study is available, while truncation refers to cases where some subjects in the population have no chance to be recruited to the study. There are different types of censoring and truncation for survival data. We explain different types of censoring and truncation with more details in the next section.

2.2 Censoring types and truncation

Censoring and truncation are two common features of time to event data. There are various categories of censoring and truncation like right censoring, left censoring, and interval censoring. The same versions exist for truncation, i.e. right truncation, left truncation and interval truncation. Each category leads to a different likelihood function which will be the basis for making statistical inference.

2.2.1 Right censoring

Right censoring, in general, means that the actual event time happens after the observation is seized on a subject. Right censoring is perhaps the most common type of censoring and has been extensively studied in the literature. There are several types of right censoring. Below we describe the most common types of right censoring.

Type I right censoring: This type of censoring happens when the event occurs after some prespecified time. In other words, there is a fixed follow-up time on each subject. In this case, the study begins at a specific time and ends after passing a predetermined period of time. The individuals could enter into the study once at the beginning of the study or they could enter one by one with a random distance from the start point of the study. In the later case, we can shift each individual's starting time to 0, in order to have a convenient representation for data. Those individuals who experience the event of interest before the end of the study period are uncensored, others are considered to be censored. A typical animal study or clinical trials are examples of this kind of data.

Type II right censoring: The type II right censoring happens when the starting time of the study is predetermined but the ending time depends on the time when the first r individuals, where r is some prespecified integer, experience the event of interest. Obviously r should be an integer less than or equal to the total number of individuals recruited for the study. An example of this kind of censoring can be seen in testing of equipment life; where produced items are put on test at the same time, and the test is terminated when r of the items fail.

Random right censoring: This type of censoring occurs when there are some other factors except the event of interest which could remove some of the individuals from a trial during the study period. Those factors are named competing events, and the individuals who have been removed from the study are considered to be random right censored. In this case, the event time and censoring time are often assumed to be independent of each other. In many studies, the censoring scheme is a combination of type I censoring and random right censoring.

2.2.2 Left censoring

Left censoring occurs when the event of interest happens some time before the beginning of the study, the failure times are left censored.
An example of this kind of censoring is when we collect data by distributing a questionnaire among students in a college to ask them about their first time they used marijuana if they have used any. In this case, one of the answers could be that " I used it but I do not know when was the first time. " This answer is an example of having left censoring. This type of censoring commonly happens in retrospective studies.

2.2.3 Interval censoring

A more general kind of censoring happens when the event time is only known to have occurred within an interval. Interval censoring happens when the individuals, involved in a study, have a periodic follow-up.

2.2.4 Truncation

The difference between truncation and censoring is that we could have some censored individuals with partial information about their event time, while truncation is a feature that limits our observation to only a part of the target population; subjects whose event time meet some criteria. The criteria can be, for example, surviving beyond some age.
Like censoring, we have random and fixed left or right truncation, and interval truncation. Fixed left (right) truncation occurs when only subjects whose event time is greater (smaller) than a fixed age can be observed. In random left (right) truncation, the fixed age is replaced by another positive random variable. For interval truncation, the event of interest should happen within some prespecified interval for a subject to be observable. It is then clear that fixed left and right truncations are special cases of interval truncation, where one side of the interval is ∞ or $-\infty$, for left and right truncation respectively. Figure 2.1 illustrates some types of censoring and truncation.

2.3 Analysis of survival data

In this section we introduce some basic concepts in survival analysis.
Let T be the time to some specific event. We treat T as a random variable taking non-

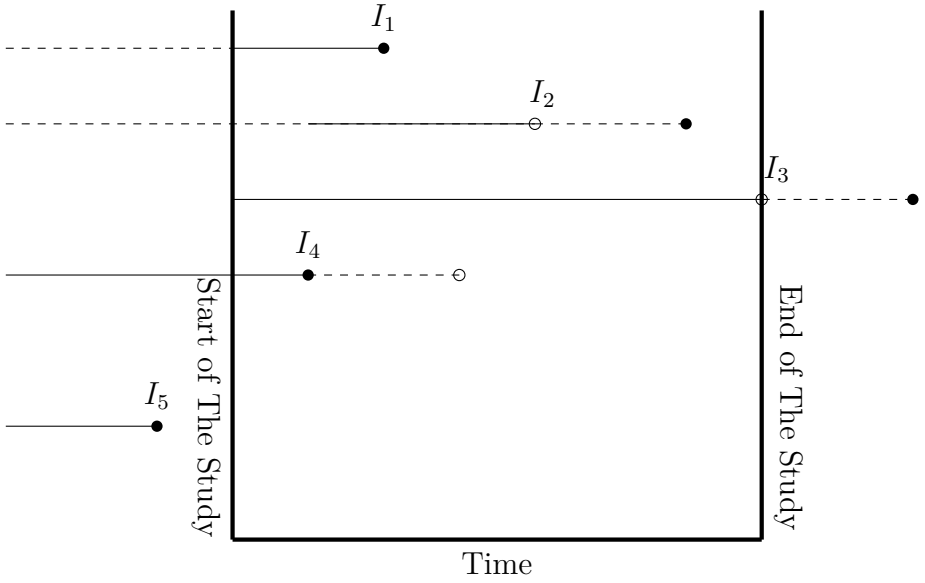

Figure 2.1 Dashed line for each individual is the time the individual is not observed. Straight lines are the time that each individual is observed. Empty circles are the censoring time, and solid circles are the time that event of interest happened. I_1 is a complete observation, I_2 is random right censoring, and I_3 is type I right censored individuals. I_4 is random left truncation, and I_5 is an individual who are not observed because of truncation.

negative values. In survival analysis, instead of the cumulative distribution function of T, i.e. $F(t) = P(T \leq t)$, we mostly use survival function, which is the probability of an individual surviving beyond time t. Survival function is denoted by $S(t) = P(T > t)$. Another fundamental concept in survival analysis is the hazard rate function (or risk function). This function represents the instantaneous density of failure, i.e. the chance for an individual who has survived until time t, to experience the event of interest in the next instant in time. Mathematically the hazard rate function is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

There is, of course, one-to-one correspondence between F, $S(t) = (1 - F(t))$ and $h(t)$. Below we present these relationships.

$$S(t) = P(T \geq t) = \int_t^\infty f(s)ds \tag{2.1}$$

$$F(t) = 1 - S(t) \tag{2.2}$$

$$f(t) = -\frac{dS(t)}{dt} \tag{2.3}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d\log\{S(t)\}}{dt} \tag{2.4}$$

All the above formulas can be proved easily. You can refer to Klein and Moeschberger (1997). In addition to all these definitions, another quantity is the cumulative hazard function $H(t)$, defined by:

$$H(t) = \int_0^t h(x)dx = -\ln\{S(t)\}. \tag{2.5}$$

Thus, we can have the following formula for continuous lifetime variable:

$$S(t) = \exp\{-H(t)\} = \exp\{-\int_0^t h(x)dx\} \tag{2.6}$$

One aim of survival analysis is to fit a model, for any of the above three functions, on a dataset. There are several ways to fit a model to data. Constructing the likelihood function is one of the useful methods which helps us to fit parametric or non-parametric models on a dataset. In the next section we illustrate how to construct the likelihood function for survival datasets.

2.3.1 Likelihood construction

While constructing the likelihood function, it is necessary to take in to account censored and truncated individuals. Whether an observation is censored, truncated, or an exact lifetime, it will have a different effect on the likelihood estimation.
When the event has happened for an individual, the probability of occurrence of the event at the time of happening is taken into account. Therefore, the probability density function at the time of occurrence of the event of interest is integrated into the likelihood function. When a right-censored observation exists, the probability of the individual survived past the censoring time is taken into account in the likelihood function. This probability can be approximated by the survival function evaluated at the censoring time. When a left-censored observation occurs, it means the event has already taken place, and corresponding cumulative density function evaluated at the censoring time contributes in the likelihood function. For an interval-censored observation, it is known that event occurred inside an interval, and

hence the probability of the event occurred during this interval (this can be calculated by using either $S(t)$ or $F(t)$) is added to the likelihood function. Confront with truncated data means we should use a conditional probability function, because observed individuals are those individuals who experience the event of interest within a certain time interval. For example, having right-truncated data means the information, required for likelihood function, is provided by chance of experiencing the event of interest at certain time conditional on not surviving before the end of follow-up.

The likelihood function, $\mathcal{L}(\theta)$, for a data set is constructed by taking the product of each individual component. For example, consider a dataset that consists of the observed lifetimes and right-censored observations. The likelihood for this data set in the case of independent censoring and truncations is:

$$\mathcal{L}(\theta) \propto \prod_{i \in D} f_\theta(x_i) \prod_{i \in R} S_\theta(x_i), \tag{2.7}$$

where D and R represent the set of observed lifetimes and right-censoring times respectively, and x_i is the observed lifetime for ith individual, and θ is the statistical parameters.

The following table shows the information components that we can use for every censoring scheme to construct likelihood function, see Klein and Moeschberger (1997).

Table 2.1 Censoring schemes and the Likelihood Function. Y_l and Y_r are the left and right boundaries of follow-up time for truncated data. C_r and C_l are right and left censoring times for censored individuals.

Likelihood Contribution	Censoring Scheme
$f(t)$	Observed lifetime
$S(C_r)$	Right-censoring
$1 - S(C_l)$	Left-censoring
$S(C_r) - S(C_l)$	Interval-censoring
$f(t)/[1 - S(Y_r)]$	Right-truncation
$f(t)/S(Y_l)$	Left-truncation
$\frac{f(t)}{[S(Y_l) - S(Y_r)]}$	Interval-truncation

2.4 Survival models

Analysing survival data requires to estimate the basic functions related to the data, like survival function or hazard rate function. This estimation is possible through parametric, nonparametric or semiparametric methods. Here, we introduce these methods briefly.
Different types of censoring lead to the different likelihood function. To avoid complexity and redundant discussion, the concept of censored observation is limited to the right censored observations in this study. The other types of censoring have equivalent models and inferences. During the current study, it has been considered that dataset includes typical right censored individuals who are identified by the pair (x_i, δ_i), in which x_i is the observed time, and δ_i is a Bernoulli indicator with value 1 for uncensored individuals and value 0 for censored individuals.

2.4.1 Nonparametric survival models

To draw an inference about the distribution of some time to event variable, based on a sample of right-censored dataset, Kaplan and Meier estimator (KME) is a replacement for empirical distribution function for ordinary data. KME is a nonparametric estimation method; Klein and Moeschberger (1997).

Kaplan-Meier estimator

Suppose Y_i provides the number of individuals who are at risk at time x_i. In other words, Y_i is the number of individuals who are alive and have not yet been censored up to time x_i or have observed time equal to x_i. Again, assume E_i is the number of individuals who experience the event of interest at time x_i. Therefore, the conditional probability that an individual who survives just prior to time x_i experiences the event of interest at time x_i is equal to E_i/Y_i.

The Product-Limit estimator proposed by Kaplan and Meier, in order to non-parametrically estimate the proportion of the population whose lifetimes surpass time t, is defined by

$$\hat{S}(t) = \prod_{x_i \leq t} \{1 - \frac{E_i}{Y_i}\}. \tag{2.8}$$

When t is less than the smallest observed survival time, then $\hat{S}(t) = 1$. Unfortunately KME

is not well defined when t is greater than the largest observed survival time.
The variance of KME is estimated by

$$\hat{V}\{\hat{S}(t)\} = \hat{S}(t)^2 \sum_{x_i \leq t} \frac{E_i}{Y_i(Y_i - E_i)}. \tag{2.9}$$

2.4.2 Parametric survival models

Parametric survival models have been used widely to analyze survival data. Parametric survival models fit a parametric survival function to a dataset. Then it is necessary to learn more about the standard parametric survival functions, and other related functions that have been introduced in Section 2.3.
In this section, we present briefly three parametric functions in survival analysis which have frequently been used in analyzing such data. Exponential, Weibull, and lognormal are three parametric functions we will discuss among all other models such as gamma, loglogistic, normal, exponential power law, and so on.

Exponential

This model suggests the density function for survival data with exponential distribution. Using equations (2.1) to (2.6), we can easily obtain other related functions of exponential survival model.

$$f(t) = \lambda \exp(-\lambda t), \quad \lambda > 0, \;\; t \geq 0 \tag{2.10}$$

$$S(t) = \exp(-\lambda t), \tag{2.11}$$

$$h(t) = \lambda. \tag{2.12}$$

Because of the well-known feature of exponential distribution which is *lack of memory* or *memoryless* property, we cannot fit this model for many real types of survival data. Because this property provides $E(T) = 1/\lambda$, that means, the mean residual life time is constant which does not apply for many types of real data. We can see that exponential hazard function is also constant, which adds another restriction in applying this model for real survival datasets.

Weibull

Weibull distribution is being used commonly in survival analysis. With Weibull distribution, we can bypass the restrictions of using exponential distribution, and have more freedom in fitting different kinds of real datasets. The exponential distribution is a special case of Weibull distribution. The related functions and parameters with this survival model are given below:

$$f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}), \qquad \alpha, \lambda > 0, \quad t \geq 0 \tag{2.13}$$

$$S(t) = \exp(-\lambda t^{\alpha}), \tag{2.14}$$

$$h(t) = \alpha\lambda t^{(\alpha-1)}, \tag{2.15}$$

where α is the shape parameter and λ is the scale parameter. This model has more flexibility to fit on real survival datasets. Because the hazard function of Weibull survival model can take any form of increasing, decreasing or constant for different values of α, and λ.

Lognormal

A random variable T is said to follow the lognormal distribution if its logarithm ($Y = \ln T$) follows the normal distribution. This distribution has been widely used for time to event datasets because of its relation to the normal distribution. We can specify this model by two parameters, the location (μ), and the scale (σ) of Y. The following equations describe the parameters of this model explicitly:

$$f(t) = \frac{\exp[-1/2\{\frac{\ln(t)-\mu}{\sigma}\}^2]}{t(2\pi)^{1/2}\sigma}, \qquad \sigma > 0, \quad t \geq 0 \tag{2.16}$$

$$S(t) = 1 - \Phi\{\frac{\ln(t)-\mu}{\sigma}\}, \tag{2.17}$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal variable. The hazard ratio function can be obtained by the $\frac{f(t)}{S(t)}$ formula.
This survival model is very popular in applied survival analysis, because of its link to the normal distribution function which is very powerful in modeling natural phenomenon.

Accelerated failure time (AFT) model

So far, all the survival models that we have introduced are univariate survival models. But, In many cases, we are interested in figuring out how some factors can affect the surviving time. The factors are usually called covariates in the statistical literature. In other words, the event time, $T > 0$ could be associated with a vector of explanatory covariates, $\mathbf{z}^\top = (z_1, z_2, ..., z_p)$. The $\mathbf{z}^\top$ may include quantitative or qualitative covariates or time-dependent covariates. In this case, we are interested in ascertaining the relationship between time to failure variable T, and the covariates $\mathbf{z}^\top$. This would be the case when we want to compare survival functions for more than one treatment, or controlling of the confounders.
One approach to the modeling of covariate effects is to use classical linear regression. In this approach, the natural logarithm of the time to event variable is denoted $Y = \ln(T)$. This transformation is used in order to convert positive variables to observations on the entire real line. The linear model for Y is:

$$Y = \beta_0 + \mathbf{z}^\top \beta + \varepsilon, \tag{2.18}$$

where β is a vector of regression coefficients, β_0 is a constant, and ε is a random error. Each error distribution yields a specific model for survival time. For example, if the error distribution is the standard normal distribution, the survival time has lognormal regression model. If the error distribution is the logistic distribution, survival time has the log-logistic model, and the extreme value error distribution gives Weibull regression model. To know more about error distributions, the reader can refer to Klein and Moeschberger (1997). This method of modelling is called accelerated failure time model. We can find coefficients and unknown parameters of the specific parametric model that we are using, by constructing and maximizing a likelihood function.

2.4.3 Semiparametric survival models

Semiparametric models combine a parametric model for some components of the model and keep nonparametric estimation for other components. The following sections describe some of the semi-parametric methods in survival analysis.

Modeling with hazard rate function

Although the modelling of the time to failure provides a very useful framework for a considerable number of cases in the real application, its use is restricted by the error distributions that one considers. Conditional hazard rate as a function of the covariates is the major method for modelling the effects of covariates on survival data. Two popular models are used for this reason, one is the multiplicative hazard model, and the other one is the additive hazard rate model. The following paragraphs describe these two approaches briefly.

Multiplicative hazard rate models: Consider for instance that we want to compare the survival function of cancer patients on two different treatments. One form of a regression model for the hazard function that could be used in such a model is:

$$h(t, \mathbf{z}, \beta) = h_0(t) r(\mathbf{z}, \beta), \tag{2.19}$$

in which the $h_0(t)$ could have any arbitrary parametric form or it can be any nonnegative function of t, and $r(\mathbf{z}, \beta)$ is a nonnegative function of covariates which does not depend on t. This model could contain both parametric and nonparametric factors. These factors must be chosen so that $h(t, \mathbf{z}, \beta) > 0$. The function $h_0(t)$ is called the baseline hazard function, the hazard function for the subjects with covariates set to zero. It is equal to the hazard function when $r(\mathbf{z}, \beta) = 1$. The ratio of model (2.19), for two individuals with covariate values denoted by $\mathbf{z}_1$, and $\mathbf{z}_2$, is:

$$HR(t, \mathbf{z}_1, \mathbf{z}_2) = \frac{h(t, \mathbf{z}_1, \beta)}{h(t, \mathbf{z}_2, \beta)} = \frac{h_0(t) r(\mathbf{z}_1, \beta)}{h_0(t) r(\mathbf{z}_2, \beta)} = \frac{r(\mathbf{z}_1, \beta)}{r(\mathbf{z}_2, \beta)}. \tag{2.20}$$

It can be seen that the hazard ratio (HR) depends only on the function $r(\mathbf{z}, \beta)$. The important part of estimation for this regression model is to determine a parametric form for $r(\mathbf{z}, \beta)$.

Cox (1972), one of the leaders in survival analysis, proposes the model (2.20). He suggested the function $e^{\mathbf{z}^\top \beta}$ as a replacement for function $r(\mathbf{z}, \beta)$. In this case the hazard ratio (HR) is equal to:

$$HR(t, \mathbf{z}_1, \mathbf{z}_2) = e^{(\mathbf{z}_1 - \mathbf{z}_2)^\top \beta}. \tag{2.21}$$

If we use the equation (2.6), the survival function for this model has the following form:

$$S(t, \mathbf{z}, \beta) = e^{-H(t,\mathbf{z},\beta)}. \tag{2.22}$$

To obtain $H(t, \mathbf{z}, \beta)$, which is the cumulative hazard function at time t for a subject with covariate $\mathbf{z}$, we can use the following method. Here, we assume the survival time is continuous. For more information refer to Cox (1972).

$$H(t, \mathbf{z}, \beta) = \int_0^t h(u, \mathbf{z}, \beta)du = r(\mathbf{z}, \beta) \int_0^t h_0(u)du = r(\mathbf{z}, \beta)H_0(t), \tag{2.23}$$

in which $H_0(t)$ is called the cumulative baseline risk. Thus it follows that

$$S(t, \mathbf{z}, \beta) = [e^{-H_0(t)}]^{r(\mathbf{z},\beta)} = \{S_0(t)\}^{r(\mathbf{z},\beta)}, \tag{2.24}$$

where, $S_0(t) = e^{-H_0(t)}$ is the base line survival function. So, under the Cox model the survival function is equal to

$$S(t, \mathbf{z}, \beta) = \{S_0(t)\}^{\exp(\mathbf{z}^\top \beta)}. \tag{2.25}$$

Cox proportional hazard model (CPH) is useful when we want to compare two or more groups of survival data, because by applying equation (2.20), there is no need to specify the base line hazard function in the analysis of such datasets.

Additive hazard rate models: Consider we have an event time T whose distribution depends on a vector of possibly time-dependent covariates, $\mathbf{z}^\top(t) = (z_1(t), ..., z_p(t))$. We assume that the hazard rate at time t, for each individual, is a linear combination of the $z_k(t)'s$:

$$h(t|\mathbf{z}(t)) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) z_k(t), \tag{2.26}$$

The p regression functions can be positive or negative, but the values are constrained because $h(t|\mathbf{z}(t))$ must be positive. Estimation for additive models is typically made by nonparametric (weighted) least-squares methods; see Klein and Moeschberger (1997) for more details.

Linear transformation models

A generalization of semi-parametric models proposed above is the linear transformation model which has the form:

$$h(T) = -\mathbf{z}^\top \beta + \varepsilon, \tag{2.27}$$

where $\mathbf{z}$ is a vector of covariates and ε is a random error with distribution function F, and β is a vector of coefficients. Another equivalent form of the linear transformation model is defined by:

$$g\{S(t|\mathbf{z})\} = h(t) + \mathbf{z}^\top \beta, \tag{2.28}$$

in which $h(t)$ is a completely unspecified strictly increasing function, and $g(t)$ is a known decreasing function such that $g^{-1}(t) = \{1 - F(t)\}$. By specifying the distribution of ε, $F(t)$, or specifying $g(t)$ we can precisely determine the transformation model of equation (2.27). If we consider F to be the extreme value distribution, $F(x) = 1 - \exp\{-\exp(x)\}$, then equation (2.27) has the form of proportional hazard model, since $g(x) = \log(-\log(x))$, and equation (2.28) will convert to:

$$\log[-\log\{S(t|\mathbf{z})\}] = h(t) + \mathbf{z}^\top \beta, \tag{2.29}$$

which (this form) is equivalent to equation (2.25), the CPH model. The advantage of linear transformation model is its generality, since F could be any distribution function. The reader can find the estimating equations of β in Cheng et al. (1995). The consistency and asymptotic normality of the estimate of β has been proven in this article.

CHAPTER 3 CURE RATE MODELS: CLASSICAL APPROACH

3.1 Introduction

Cure rate models are a special case of survival models where a portion of subjects in the population never experience the event of interest. Such subjects are called immune or *cured*. There are two platforms for cure models i) the mixture cure rate models, also known as standard cure rate models, which have been widely used for modelling survival data with a cured fraction. ii) the non-mixture cure rate models which have attracted less attention so far.

The survival time of cured individuals might be censored at the end of the follow-up study. Hence, if the follow-up time is long enough, there might be cured individuals in the dataset. In general, however, cured individuals comprise a subset of censored individuals. The challenging aspect of fitting cure models on survival data is that the presence of cured fraction in the sample is not obvious. It is recommended to test the existence of an immune fraction in the sample before fitting cure models, and also test whether the follow-up is long enough or not. In this chapter some preliminary tests are first introduced. We then present parametric and non-parametric mixture cure models.

3.2 Preliminaries

One main reference for the materials in this section is Maller and Zhou (1996).

Farewell (1986) illustrates some restrictions that one confronts when applying cure models, in particular, the mixture models. One restriction is that one needs strong scientific evidence for the existence of two or more substructures in the population when applying mixture models on a dataset. It is reasonable to think that cure models in a clinical setting are sensible only if the data are based on a long-term follow-up study. The use of such models therefore requires careful attention. Visually, the presence of cured fraction in the dataset can be seen when the Kaplan-Meier survival curve reach a plateau, see in Figure 3.1. A test statistics to verify the presence of a cured fraction is proposed below.

Let T with cumulative distribution function F, and C with cumulative distribution function G, respectively, represent the failure and censoring time. Suppose δ is the censoring indicator,

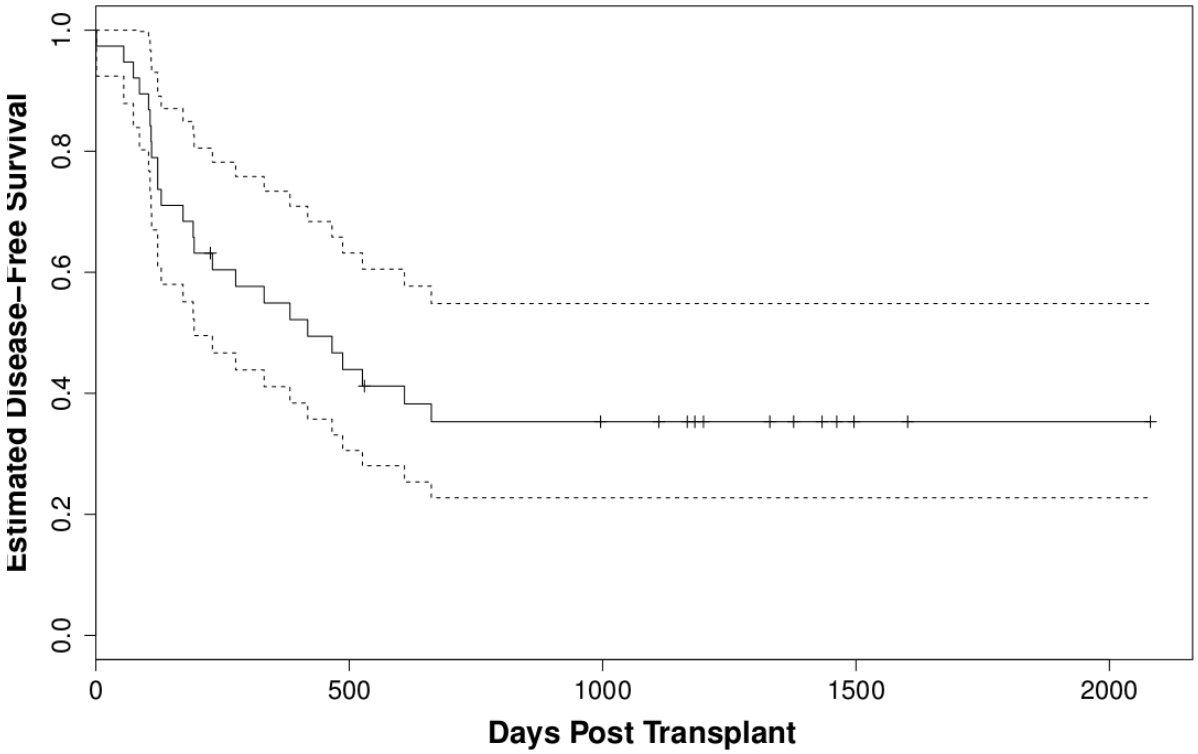

Figure 3.1 Kaplan-Meier estimator for group 1 of patients in *bone marrow transplant* data. Doted lines are a confidence interval.

i.e. $\delta = 1$ if $T \leq C$, and 0 otherwise. Then the observations are $D = \{(x_i, \delta_i), i = 1, 2 \ldots, n\}$ where $x_i = \min(t_i, c_i)$, $i = 1, 2, \ldots, n$ and n is the sample size.
In general a cumulative distribution function A is a proper distribution function if

1. $A(\infty) = \lim_{t\to\infty} A(t) = 1$,
2. $A(-\infty) = 0$.

The distribution function $A(t)$ is improper if one of the above conditions does not apply. In mixture cure rate formulations $G(t)$, the distribution function of censoring times, is required to be proper, but $F(t)$, the distribution function of the event times, is not required to be a proper cumulative distribution function; Maller and Zhou (1996, p. 31).
Suppose

$$p = P(T < \infty) = F(\infty) = \lim_{t\to\infty} F(t). \tag{3.1}$$

The probability p is the proportion of individuals in the population who eventually experience the failure, or the event of interest, if the follow-up time is long enough. This probability p which is always less than or equal to 1, sometimes can be strictly less than 1. Therefore, the immune fraction of population is $1 - p$, i.e. the proportion of individuals in the population

who never experience the event of interest. From equation (3.1) it can be seen that $F(t)$ could be an improper distribution function.
Let

$$\tau_A = \inf\{t \geq 0 : A(t) = 1\}; \tag{3.2}$$

The τ_A is called the right extreme of the distribution $A(t)$. If the function $F(t)$ is an improper distribution function τ_F is ∞, since $F(t) \leq F(\infty) < 1, \text{ for all } t$.

In Section 3.2.1, we highlight some questions that should be addressed prior to analysing time-to-event data.

3.2.1 Determination of cured fraction

Assume that $\tau_G < \tau_F$. This assumption is crucial for identifiability in nonparametric and semi-parametric settings, but can be relaxed in parametric settings.
From equation (3.1) it is concluded that immune subjects are present in the population if and only if

$$p = F(\infty) < 1.$$

Since, if $F(\infty) = 1$ there is no immune fraction. Given that it is impossible to calculate $F(\infty)$ when $F(t)$ is unknown, we consider $F(\tau_G)$ as the base for developing a test of hypothesis, as it is suggested in Maller and Zhou (1996, p. 36).
We therefore consider $\mathbf{H}_{01} : F(\tau_G) = 1$ to test for the presence of immune fraction. If $\mathbf{H}_{01}$ is accepted, there is no evidence of having an immune fraction. If $\mathbf{H}_{01}$ is rejected, we test for having a long enough follow-up.
To obtain a value for τ_G, it is needed to know the distribution of the censoring time, G. In most practical applications, however, G is unknown as well. Therefore, the largest observed time in the sample is used instead of τ_G. Suppose $x_{(n)}$ is the largest observation, censored or uncensored, in the sample, and $\hat{F}_n\{x_{(n)}\}$ is the value of Kaplan-Meier estimator at $x_{(n)}$. A nonparametric estimator of p is

$$\hat{p}_n = \hat{F}_n\{x_{(n)}\}. \tag{3.3}$$

It has been proved that $\hat{F}_n\{x_{(n)}\}$ is a consistent estimator for $F(\tau_G)$ under a mild continuity condition and $\tau_G \leq \tau_F$; see Maller and Zhou (1996, chap. 3). Since $\hat{F}_n\{x_{(n)}\}$ is an estimate for $F(\tau_G)$, if $\hat{p}_n = 1$, $\mathbf{H}_{01}$ is accepted and there is no immune fraction in the population. To summarize,

$$\text{Reject} \quad \mathbf{H}_{01} \ \text{ if } \ \hat{p}_n < c_\alpha, \tag{3.4}$$

where c_α is the α^{th} percentile of the distribution of $\hat{p}_n$ under $\mathbf{H}_{01}$. To obtain the value of c_α, the distribution of $\hat{p}_n$ is needed, and in general this distribution is unknown; see Maller and Zhou (1996) [1] who performed simulations to estimate the value of c_α, when α is equal to 1%, 5%, 10%, or 20%.

3.2.2 Is the follow-up time long enough?

Observations in a sample may consist of two different groups, the *cured* individuals (we sometimes refer to them as *insusceptible* or *immune* individuals), and the *uncured* (we sometimes refer to them as *susceptible* or *nonimmune*). Note that susceptible individuals could be either censored or not censored, i.e. those individuals whose event time is observed during the study period are not censored.
Suppose

$$F^*(t) = \frac{F(t)}{F(\infty)}.$$

The function $F^*(t)$ is the proper distribution function of susceptible individuals, i.e.

$$F^*(\infty) = \frac{F(\infty)}{p} = 1.$$

Analogously, τ_{F^*} is the extreme value of the survival times of susceptible individuals.

When $\tau_F \leq \tau_G$, we may observe the largest possible event time up to the maximum possible. In contrast, when $\tau_F > \tau_G$, it means censoring is so heavy and we may not be able to observe all event times.Therefore, $\tau_F \leq \tau_G$ can be used to examine whether the follow-up time is enough. Note that for susceptible individuals if $t_i \leq c_i$ then the failure time for individual i^{th} is observed. It is straightforward that $\tau_{F^*} < \tau_F$. In Maller and Zhou (1996) $\tau_{F^*} \leq \tau_G$ is considered as the reference for enough follow-up time. For more details see Maller and Zhou (1996, p. 33).

[1] Maller and Zhou (1996), Appendix A.1 and A.2.

Therefore the desired hypothesis test for sufficient follow-up time is

$$\begin{cases} \mathbf{H}_{02}: & \tau_{F^*} \leq \tau_G \\ \mathbf{H}_{02}^c: & \tau_{F^*} > \tau_G. \end{cases} \tag{3.5}$$

This means only the magnitude of $\tau_G - \tau_{F^*}$ needs to be estimated, the distance between largest possible censored time and largest possible failure time of susceptible individuals. Assume $x_{(n)}$ is the largest observed survival time; and $t_{(n)}$ is the largest uncensored failure time, it has been proved that

$$x_{(n)} - t_{(n)} \to \begin{cases} \tau_G - \tau_{F^*} & \text{if } \tau_G \leq \tau_{F^*} \\ 0 & \text{if } \tau_G > \tau_{F^*}, \end{cases}$$

almost surely as $n \to \infty$ (Maller and Zhou (1996)).

Consequently, long enough follow-up time is the result of two conditions:

i) whether the presence of immune fraction is accepted in the sample, i.e. $\mathbf{H}_{01}$ is rejected.

ii) if $x_{(n)} - t_{(n)}$ is large *enough*.

Then one can accept that the follow-up is sufficient. However, since the distribution of $x_{(n)} - t_{(n)}$ is unknown, a simpler quantity should be found to perform this hypothesis test. To this end Maller and Zhou (1996) defined

$$q_n = \frac{N_n}{n} = \frac{\text{Number of uncensored } x_i \text{ in the interval } (2t_{(n)} - x_{(n)}, t_{(n)}]}{\text{Number of sample individuals}} \tag{3.6}$$

to test this hypothesis.

Large values of q_n suggest rejecting $\mathbf{H}_{02}^c$ or equivalently accepting $\mathbf{H}_{02}$. This means the follow-up time is long enough. However, how large q_n must be to accept $\mathbf{H}_{02}$ is still unclear, since the distribution of q_n (and N_n) is unknown. Again the simulation results of Maller and Zhou (1996) can be used to evaluate the percentiles of distribution of q_n. If the value of q_n exceeds the tabulated $(1-\alpha)$th percentile (for α equal to 1%, 5%, 10%, or 20%), $\mathbf{H}_{02}$ is accepted.

3.3 Mixture cure rate models

Most of the early works in cure rate models are based on parametric mixture models. In this section we aim to explain one of the earliest work in mixture cure rate models. Some other parametric, nonparametric, and semi-parametric methods are discussed in sections 3.4, 3.5, and Chapter 4. We start with a simple parametric version of mixture cure rate model, using the exponential distribution function for the survival time.

3.3.1 History

The paper by Berkson and Gage (1952) is our main reference in this section. The idea of mixture cure rate models originates from comparison between survival curves of two groups of subjects. One group is the patients under a specific treatment, and the other is a sample from a control group. Consider logarithm of survival times in drawing two survival curves and draw them on one figure. The ratio of survival times of two groups can be found by vertical difference between the two curves. It has been shown that the two curves approximately become parallel after passing some time. It means the instantaneous failure rate for the two curves is equal, and failure rate for group of patients who are under the treatment becomes equal to failure rate for the controlled group, at some point. Cured fraction is the proportion of patients who are subject to normal failure rate in the controlled group. This definition of cured fraction first appeared in Berkson and Gage (1952).

Berkson and Gage (1952) divide the population in two hypothetical groups, one group is just subject to normal failure rate which is called cured fraction. The other group is subject to normal failure rate *and* a specific failure rate which could be the failure rate for a disease under the study. These two failure rates are represented by q_0, for normal failure rate, and q_{ca}, for the specific failure rate of the disease. These two failure rates act independently and simultaneously. Assume the two groups act separately, then the probability of survival for the cured fraction is $l_0 = \prod_{i=1}^{n}(1 - q_0)$, and the probability of survival for uncured fraction is $l_0 l_{ca}$, in which $l_{ca} = \prod_{i=1}^{n}(1 - q_{ca})$.

Berkson and Gage (1952) made another assumption to simplify their modelling. They considered the rate of death caused by the specific disease to be a constant (say β), i.e. the hazard function for uncured population is β. So the survival probability for the uncured population should decrease exponentially and be time-varying, implying that $l_{ca}(t) = e^{-\beta t}$. By considering p as the cured fraction in our population, the probability of survival up to

time t for total population is

$$\text{Probability of survival up to time } t = pl_0 + (1-p)l_0 e^{-\beta t}. \tag{3.7}$$

The unknown parameters in equation (3.7) are p and β. Dividing both sides of equation (3.7) by l_0 we have

$$\frac{\text{Probability of survival up to time } t}{l_0} = p + (1-p)e^{-\beta t}. \tag{3.8}$$

The equation (3.8) can be interpreted as the probability of survival in the whole population. This interpretation is valid only if the population is free of death by any other causes, except the disease of interest. Parameters p and β are estimated by least squares method using one of the numerical minimization routines. For complete details of minimization methods see Berkson and Gage (1952).

Berkson and Gage (1952) did not consider any censorship for the data during the experiment. Cured individuals are those who survive by the end of the experiment, and uncured individuals are those who are faced with the event of interest during the follow-up.

3.3.2 Formulation of mixture models

In survival analysis, observations usually consist of the following random variables.
Random variable $X = \min\{T, C\}$, where T is the failure time, and C is censoring time, and Bernoulli variable $\delta = 0$ if the individual is censored, and is equal to 1 otherwise. Also, some covariates may be added to these variables. We often assume that T and C are independent.

Subjects can be divided in another category which divides individuals between susceptible (uncured) and insusceptible (cured) individuals. A Bernoulli random variable η is an indicator for susceptibility of each individual. This variable η takes value 1 if the individual is susceptible and 0 otherwise. Of course, η is not observed; essentially, there is no information from a study, but it has been used as a latent variable in the model formulation. To distinguish properly the difference between variables η and δ, consider η as the true event status and δ as the observed failure status; essentially, $\{\delta_i = 1, i = 1, \ldots, n\} \subset \{\eta_i = 1, i = 1, \ldots, n\}$, where i denotes an individual and n is sample size. The latent variable η enable us to divide event time in two categories: those individuals who meet the event, and those who never

meet the event of interest. Following the decomposition for event (failure) time:

$$T = \eta T^* + (1 - \eta)\infty. \tag{3.9}$$

Equation (3.9) illustrates that failure time T is decomposed by T^*, the survival time of susceptible individuals, and survival time of insusceptible individuals which conventionally is considered to be ∞ and never happens. By introducing η the true survival time, $T = \infty$ becomes reasonable. We introduced before $F^*(t)$ as the proper cumulative distribution for survival time; thus,

$$P(T \leq t|\eta = 1) = F^*(t), \tag{3.10}$$
$$P(T \leq t|\eta = 0) = 0. \tag{3.11}$$

The density and survival functions of cured individuals are set to zero and one, respectively, for all finite values of t because cured (insusceptible) individuals will never experience the failure. Thus,

$$F(t) = P(T \leq t) \quad = P(T \leq t|\eta = 1)P(\eta = 1) + P(T \leq t|\eta = 0)P(\eta = 0) \tag{3.12}$$
$$= pF^*(t) + 0 = pF^*(t). \tag{3.13}$$

The last equality holds since $P\{\eta = 1\} = p$. The overall formulation of a mixture cure rate model is derived from the above equation. Briefly, the mixture cure rate models have the form

$$F(t) = pF^*(t), \tag{3.14}$$

or equivalently,

$$S(t) = (1 - p) + pS^*(t). \tag{3.15}$$

The functions $S(t)$ and $S^*(t)$ are improper and proper cumulative survival functions of T. Consequently, $f^*(t)$ is a proper density function of T.
In the previous part, equation (3.7) is equivalent to the mixture formulation of equation (3.15), if T has exponential distribution with parameter β.

3.3.3 Likelihood function for mixture models

The general likelihood function for mixture rate cure model is

$$\mathcal{L}(\theta, p) = \prod_{i=1}^{n} \{p f_\theta^*(t)\}^{\delta_i} \{p S_\theta^*(t) + (1-p)\}^{1-\delta_i}, \tag{3.16}$$

where θ is a vector of statistical parameters, and the functions $S_\theta^*(t)$ and $f_\theta^*(t)$ are identified with these parameters. To simplify notation we stop writing index θ repeatedly unless for emphasizing and remembering this.
This likelihood is derived from the fact that the probability of experiencing the event for those individuals who are not censored at time t is $p f_\theta^*(t)$, and the probability of staying alive up to time t for those individuals who have been censored at time t is $p S_\theta^*(t) + (1-p)$. The specification of $S_\theta^*(t)$, or equivalently $f_\theta^*(t)$, can be parametric or nonparametric, which leads to parametric and nonparametric mixture models.

3.4 Parametric mixture models

Parametric mixture cure rate models are obtained by simply considering a parametric model for $S_\theta^*(t)$ in equation (3.15). The most frequently used parametric models for $S_\theta^*(t)$ are Weibull, logistic, and exponential.

3.4.1 Weibull and conditional logistic mixture model

The susceptibility variable, η, has been defined by **?**. He divides individuals into two cohorts, one cohort is those individuals who face by the event of interest during the follow-up, $\eta = 1$, and another is those individuals who survive by the end of experiment, and these could have either $\eta = 0$ or $\eta = 1$.

The term $\pi(\gamma^\top \mathbf{w})$ has been used here as the cured fraction, p in equation (3.15) to emphasize the dependency of cured fraction on some covariates, $\mathbf{w}$. Obviously γ is the vector of related covariate coefficients to be estimated. To connect the cure fraction with the vector of covariates, a link function is needed. Logistic regression is an option for link function as follows:

$$P(\eta = 1|\mathbf{w}) = \pi(\gamma^\top \mathbf{w}) = \frac{\exp(\gamma^\top \mathbf{w})}{1 + \exp(\gamma^\top \mathbf{w})}. \tag{3.17}$$

Also, to emphasize the possibility of a connection between survival time of individuals and some covariates like vector $\mathbf{z}$, the mixture model formulation which is introduced by equation (3.15) can be reformulated with the following notation:

$$S(t|\mathbf{w}, \mathbf{z}) = 1 - \pi(\gamma^\top \mathbf{w}) + \pi(\gamma^\top \mathbf{w}) S^*(t|\mathbf{z}). \tag{3.18}$$

In the above equation for time to event variable, T, two different parametric models have been assumed, one for individuals who are susceptible and the other one for those who are not susceptible. Consider the probability of survival for susceptible individuals ($\eta = 1$), with covariate vector $\mathbf{z}$, is obtained by Weibull distribution, Farewell (1982). The Weibull distribution function is defined in (2.13), where $\lambda = \exp(-\beta^\top \mathbf{z})$ is replaced with scale parameters, and vector β represents unknown regression coefficients. Note that the parameter λ differs for each individual because of covariate the vector $\mathbf{z}$.
We refer to the combination of (3.17) and Weibull distribution function as Weibull mixture cure rate model.
Assume that no individual with $\eta = 0$ experience failure during the follow-up. The unknown parameters are estimated using maximum likelihood. The probability of failure at time t for an individual is $P(\eta = 1|\mathbf{z}) f(t|\eta = 1, \mathbf{z})$. If the individual has been followed completely during the study, the probability of survival by the time t becomes

$$\{1 - P(\eta = 1|\mathbf{z})\} + P(\eta = 1|\mathbf{z}) \int_t^\infty f^*(s|\eta = 1, \mathbf{z}) ds. \tag{3.19}$$

The likelihood function has the form of equation (3.16), where $S^*(t) = \int_t^\infty f^*(s|\eta = 1, \mathbf{z}) ds$, and $f^*(t)$ is replaced by Weibull distribution function. The cured fraction, p, also is replaced by parametric model of equation (3.17). An iterative method, like Newton-Raphson, is adopted to maximize the log likelihood function, numerically.

3.5 Nonparametric mixture models

In this section a nonparametric estimation is assumed for F^*, the cumulative distribution of the survival times for susceptible individuals. The first nonparametric model that crosses the mind to fit on survival data is Kaplan-Meier estimator. In statistical modelling context usually the aim is to examine the effects of multiple covariates on the response variable. In

cure rate models, one of the goals is to investigate how other factors can affect the cured fraction and the survival time of susceptible patients. Cox proportional hazard (CPH) model is a well-known survival model to investigate the effect of covariates on survival time. The following section aims to explain a nonparametric approach for mixture cure rate models, using CPH model.

3.5.1 Proportional hazard mixture model

The material of this section is acquired from Peng and Dear (2000). To study a general nonparametric mixture model, the CPH model can be a good choice for connecting failure time to some covariates. Because CPH model is specifically appropriate for survival data and relaxes the normality assumption. The EM algorithm is another tool that has been used for estimating parameters in this method Dempster et al. (1977).
Again, suppose that $\mathbf{z}$ and $\mathbf{w}$ are two covariate vectors related to each individual, in which the distribution of T^*, and cured fraction, respectively, may depend on them.

For susceptibility indicator a model similar to equation (3.17) can be considered, and $S^*(t|\mathbf{z})$ is replaced with survival CPH function to model the effects of covariate $\mathbf{z}$ on the failure distribution of susceptible individuals. The CPH model takes the following form:

$$h^*(t|\mathbf{z}) = h_0^*(t)\exp(\mathbf{z}^\top\beta), \tag{3.20}$$

in which $h^*(t)$ is hazard function, and $h_0^*(t)$ is the baseline hazard function and can be any arbitrary specified hazard function but not a function of $\mathbf{z}$. This leads to $S^*(t|\mathbf{Z}) = S_0^*(t)^{\exp(\beta^\top\mathbf{z})}$ as a model for survival function of susceptible individuals, where $S_0^*(t) = \exp\{-\int_0^t h_0^*(s)ds\}$. EM algorithm can be used to estimate γ in (3.17), and β in (3.20).

EM algorithm is an iterative method for finding locally maximum likelihood and for estimating parameters in a model. This algorithm has been applied usually when it is assumed there are some unmeasured parameters in the dataset. These unmeasured parameters is sometimes called missing values. This method consists of two steps, the E-step and the M-step. In the E-step the aim is to find expectation of complete log-likelihood (as a function of missing value) conditional on the observed values and estimated parameters in the previous iteration, and in the M-step the goal is to find a new estimation for the parameters of the model by maximization of the expectation of complete log-likelihood. For complete details of EM

algorithm on mixture models, see Dempster et al. (1977).

To use the EM algorithm, η is treated as the missing parameter, which is an indicator for susceptibility. It is obvious that when $\delta_i = 1$, i.e. censoring indicator is equal to 1, therefore $\eta_i = 1$, which means individual ith is susceptible. If ith individual is censored, $\delta_i = 0$, then η_i is one or zero. Therefore η is partially missing information (or parameter); it could be used to illustrate a new complete likelihood function for mixture cure model given η, and observed variables.
Note that $\pi(\gamma^\top \mathbf{w}_i) = P(\eta_i = 1|\mathbf{w}_i)$.

Let's suppose $\mathbf{O} = (x_i, \delta_i, \mathbf{w}_i, \mathbf{z}_i)$ denotes the observed data for ith individual, $i = 1, \ldots, n$, and $\theta = (\gamma, \beta, S_0^*)$ as unknown parameters to be estimated.
Given η and $\mathbf{O}$ the complete log-likelihood function is

$$\ell_c(\theta) = \log\left\{ \prod_{i=1}^{n} \Big[\{\pi(\gamma^\top \mathbf{w}_i) f^*(x_i|\mathbf{z}_i)\}^{\eta_i} \Big]^{\delta_i} \times \Big[\{1 - \pi(\gamma^\top \mathbf{w}_i)\}^{1-\eta_i} \{\pi(\gamma^\top \mathbf{w}_i) S^*(x_i|\mathbf{z}_i)\}^{\eta_i} \Big]^{1-\delta_i} \right\}, \tag{3.21}$$

and the probability of being uncured for ith individual is

$$g_i = P(\eta_i = 1|\mathbf{O}, \theta) = \delta_i + \frac{(1-\delta_i)\pi(\gamma^\top \mathbf{w}_i) S^*(x_i|\mathbf{z}_i)}{1 - \pi(\gamma^\top \mathbf{w}_i) + \pi(\gamma^\top \mathbf{w}_i) S^*(x_i|\mathbf{z}_i)}. \tag{3.22}$$

The first term on the right-hand side of equation (3.22) represents the probability of observing failure when the individual is not censored, which is one ($\delta_i = 1$). The second term is the probability of susceptibility when the individual is censored at time x_i.
Note that the probability of susceptibility for an individual with covariate vector $\mathbf{w}_i$ and $\mathbf{z}_i$, given that subject be censored, is equal to

$$P(\eta_i = 1|\delta_i = 0) = \frac{\pi(\gamma^\top \mathbf{w}_i) S^*(x_i|\mathbf{z}_i)}{1 - \pi(\gamma^\top \mathbf{w}_i) + \pi(\gamma^\top \mathbf{w}_i) S^*(x_i|\mathbf{z}_i)}.$$

For the E-step conditional expectation of the complete log-likelihood with respect to η at the current estimates of $F^*(x_i|\mathbf{z}_i)$ and $\pi(\gamma^\top \mathbf{w}_i)$ is computed. Suppose the complete log-likelihood function is split into two functions like below:

$$\ell_c(\theta) = \ell_{c1}(\gamma) + \ell_{c2}(\beta, S_0^*).$$

Indicator $E_\eta(\ell_c|\mathbf{O},\theta)$ denotes the expectation of complete log-likelihood and can be written as the sum of following equations:

$$E(\ell_{c1}|\mathbf{O},\gamma)=\sum_{i=1}^{n}\Big[g_i\log\{\pi(\gamma^\top\mathbf{w}_i)\}+(1-g_i)\log\{1-\pi(\gamma^\top\mathbf{w}_i)\}\Big], \tag{3.23}$$

$$E(\ell_{c2}|\mathbf{O},\beta,S_0^*)=\sum_{i=1}^{n}\Big[g_i\log\{S^*(x_i|\mathbf{z}_i)\}+\delta_i\log\{h^*(x_i|\mathbf{z}_i)\}\Big], \tag{3.24}$$

where $h^*(t)=\frac{f^*(t)}{S^*(t)}$ is the hazard function of susceptible individuals.
The next phase of the EM algorithm is the M-step. In this step equations (3.23) and (3.24) are maximized with respect to θ for a fixed g_i. In the E-step, the expectation of log-likelihood is broken into two separate functions, (3.23) and (3.24). The equation (3.23) depends on the unknown parameter γ only, and equation (3.24) contains only the unknown parameter β. Therefore, the maximization step can be carried out by maximizing these two functions separately. Equation (3.23) can be maximized analytically, but maximizing equation (3.24) is not as easy, because it depends on the relation between covariates and failure times. Equation (3.20) represents how covariates vector $\mathbf{z}$ are connected to failure time. The only missing part is the baseline hazard function, $h_0^*(t)$, in equation (3.20). Equivalently $S_0^*(t)$ can be specified instead of $h_0^*(t)$, because when $S_0^*(t)$ is determined, $h_0^*(t)$ can be specified. In the following we discuss the estimation of $S_0^*(t)$.

Let $t^*_{(1)},t^*_{(2)},\ldots,t^*_{(q)}$ be the ordered distinct uncensored failure times. The set of δ_j tied uncensored failures at $t^*_{(j)}$ is denoted by D_j. Let E_j be the set of individuals with censoring time in $[t^*_{(j)},t^*_{(j+1)})$, $j=0,\ldots,q$, in which $t^*_{(0)}=0$, and $t^*_{(q+1)}=\infty$. R_j is the risk set at time $t^*_{(j)}$ and contains those individuals who are alive and uncensored just prior to time $t^*_{(j)}$. Again, let $\mathbf{z}_{(1)},\mathbf{z}_{(2)},\ldots,\mathbf{z}_{(q)}$ be the covariates related to $t^*_{(1)},t^*_{(2)},\ldots,t^*_{(q)}$. Here, the final equation that has been obtained in Peng and Dear (2000) is introduced briefly; for more details the reader can refer to this paper. Equation (3.24) can be approximated by:

$$\log\prod_{j=1}^{q}\frac{\exp(\beta^\top\mathbf{s}_j)}{\{\sum_{i\in R_j}g_i\exp(\beta^\top\mathbf{z}_i)\}^{\delta_j}}, \tag{3.25}$$

where $\mathbf{s}_j=\sum_{i\in D_j}\mathbf{z}_i$ is the sum of covariates vector for individuals in D_j. Maximizing (3.24) is equivalent to maximizing (3.25). This new equation involves the regression parameter β but not the baseline hazard function, and it can be maximized numerically using Newton-Raphson. The estimation for baseline survival function $S_0^*(t)$ is required as well; see details in Peng and Dear (2000). The estimation of $S_0^*(t)$ based on information obtained from estimating β, γ and g_i is as follows:

$$\hat{S}_0^*(t) = \exp\left(- \sum_{j:t^*_{(j)}<t} \frac{d_j}{\sum_{i\in R_j} g_i \exp(\beta^\top \mathbf{z}_i)} \right). \tag{3.26}$$

Here a pseudo code is given for the whole steps of this method.

1. Consider initial values for γ, β and $S_0^*(t)$, like γ_0, β_0 and $S_{00}^*(t)$ ($S_{00}^*(t)$ obtained by injecting β_0 in equation (3.26)).

2. In j^{th} iteration, find g_i from equation (3.22), for each $i \in 1, \ldots, n$, regarding to estimated parameters in $(j-1)^{th}$ iteration. Name it g_i^j.

3. Insert g_i^j and $\gamma^{(j-1)}$, $\beta^{(j-1)}$ in equations (3.23), (3.25) and obtain $\gamma^{(j)}$, $\beta^{(j)}$.

4. In j^{th} iteration $S_0^{*(j)}(t)$ can be found by using $\beta^{(j)}$ and g_i^j in equation (3.26).

5. Steps 2 to 4 should be repeated until a predefined convergence criterion is verified.

CHAPTER 4 CURE RATE MODELS: MODERN APPROACH

4.1 Introduction

Recent studies in cure rate models are more focused on semiparametric mixture and non-mixture cure rate models. In this chapter some of the more recent studies in mixture and non-mixture models are discussed.

4.2 Semiparametric models

While parametric models force to impose models on data generating process, up to finite many unknown parameters, non-parametric methods are procedures that work rather with minimal distributional assumptions. The semi-parametric approach is an alternative method which tries to retain the strength of both parametric and nonparametric models: the efficiency and interpretability of parametric models, and robustness of the nonparametric models to departure from the underlying assumptions about the data generality mechanism. In the following, some of the semiparametric mixture and non-mixture cure models are introduced.

4.2.1 Cure rate quantile regression mixture model

Most of the previous two-component mixture cure rate models are mean-based regression models, which give an overall quantification for the central covariate effects. The method of the current section is based on quantile regression. Quantile regression finds a better assessment of covariate effects, especially when the cure fraction is high. To estimate both quantile regression coefficients and cure fraction coefficients, martingale-based estimating equations have been proposed in the literature (Wu and Yin (2013)). Two estimation methods have been suggested for the cure rate parameters: one method is based on the iteration between the cure rate parameters and the quantile regression coefficients, and the other method separates them by applying the nonparametric kernel smoothing technique.

To connect covariates with cured fraction again, the logistic regression of equation (3.17) has been suggested. Equation (3.9) is used to represent the true failure time as a convex combination of failure time of susceptible individuals, T^*, and ∞ if the individual is non-

susceptible. To connect T^* with covariates the following linear model can be postulated:

$$\log T^* = \beta^\top \mathbf{z} + \varepsilon, \tag{4.1}$$

where $\mathbf{z}$, is a vector of covariates related to T^*, and the error ε may depend on $\mathbf{z}$. The aim is to model a set of quantiles of susceptible survival times. Assume $\tau \in (0,1)$, according to Wu and Yin (2013) the definition of τ^{th} conditional quantile function is

$$Q_{T^*}(\tau|\mathbf{z}) = \inf\{t : P(T^* \leq t|\mathbf{z}) \geq \tau\}, \tag{4.2}$$

and quantile regression model is given by

$$Q_{T^*}(\tau|\mathbf{z}) = \exp\{\mathbf{z}^\top \beta(\tau)\}, \tag{4.3}$$

where $\beta(\tau)$ is an unknown vector of regression coefficients for each $\tau \in (0,1)$. Now, the estimation of the coefficients β and γ is required. At first the method for estimating β is discussed, and then two different methods for estimating γ are explained.

By applying the martingale formulation of censored quantile regression in Peng and Huang (2008), the following estimating function can be used to estimate $\beta(\tau)$. To find out more about how this equation is obtained you can refer to Wu and Yin (2013).

$$\mathbf{U}_n(\beta, \tau; \gamma) = n^{-1} \sum_{i=1}^{n} \mathbf{z}_i \Big\{ (N_i[\exp\{\mathbf{z}_i^\top \beta(\tau)\}] - \int_0^\tau I[x_i \geq \exp\{\mathbf{z}_i^\top \beta(\tau)\}] dH(u|\mathbf{w}_i) \Big\}, \tag{4.4}$$

where,

$$H_\gamma(u|\mathbf{w}) = -\log(1 - \pi(\gamma^T \mathbf{w})u) \text{ for, } 0 \leq u \leq 1,$$

and $N_i(t) = I(x_i \leq c_i) I(x_i \leq t)$.

Let L indicates the duration of the study and τ_{max}, the upper bound of the quantile levels that can be estimated, be a constant in $(0,1)$. To ensure the identifiability for all regression quantiles below τ_{max}, it is required that τ_{max} be smaller than $\inf_{\mathbf{z}} F^*(L|\mathbf{z})$. Here, L is the duration of the follow-up time. The proof of this statement is given in Wu and Yin (2013). Estimation of $\beta(\tau, \gamma)$, for a fixed τ and a fixed γ, can be obtained by solving the equation $\mathbf{U}_n(\beta, \tau; \gamma) = 0$. Equivalently a grid-based estimation procedure to estimate $\beta(\tau, \gamma)$ is ob-

tained by sequentially locating the minimizer of the following convex objective function (4.5). We denote a partition over the interval $[0, \tau_{max}]$, as $S = \{0 \equiv \tau_0, \ldots, \tau_{q_n} \equiv \tau_{max}\}$, where the number of grid points q_n is allowed to depend on n, the number of subjects. The estimation of $\beta(\tau_j, \gamma)$ is needed, for $j = 1, \ldots, q_n$, by sequentially minimizing the following convex function:

$$\mathbb{L}_j(\mathbf{b}) = \sum_{i=1}^{n} \left|\delta_i \log x_i - \delta_i \mathbf{b}^\top \mathbf{z}_i\right| + \left|\mathbf{R}^* - \mathbf{b}^\top \sum_{i=1}^{n}(-\delta_i \mathbf{z}_i)\right| + \left|\mathbf{R}^* - \mathbf{b}^\top \sum_{i=1}^{n}\left[2\mathbf{z}_i \sum_{k=0}^{j-1} I\left[x_i \geq \exp\{\mathbf{z}_i^\top \hat{\beta}(\tau_k, \gamma)\}\right]\{H_\gamma(\tau_{k+1}|\mathbf{w}_i) - H_\gamma(\tau_k|\mathbf{w}_i)\}\right]\right|, \quad (4.5)$$

where $\mathbf{R}^*$ is a large number greater than $10^3(p+1) \times \max\{\sum_{j=1}^{p} z_{ij}^2 : 1 \leq i \leq n\}$.

Two methods have been proposed for estimating γ; the iterative method and the nonparametric approach.

In Section 3.5.1 the probability of being susceptible for i^{th} individual is given by equation (3.22). Thus, equation (3.22) is adopted to construct an estimating equation for γ in iterative approaches. Since $P(\eta_i = 1|\mathbf{w}_i) = \pi(\gamma^\top \mathbf{w}_i)$, it follows that

$$E\left[\mathbf{w}_i\{P(\eta_i = 1|\mathbf{O}) - P(\eta_i = 1|\mathbf{w}_i)\}\right] = 0.$$

The above equation leads to the following estimating function for γ:

$$\mathbf{S}_n(\gamma, F^*) = n^{-1} \sum_{i=1}^{n} \frac{\mathbf{w}_i\{1 - \pi(\gamma^\top \mathbf{w}_i)\}}{1 - \pi(\gamma^\top \mathbf{w}_i)F^*(x_i|\mathbf{z}_i)}\{\delta_i - \pi(\gamma^\top \mathbf{w}_i)F^*(x_i|\mathbf{z}_i)\}. \quad (4.6)$$

Consequently, equation (4.7), which is equivalent to equation (4.6), can be considered as an estimating function for estimating γ. See Wu and Yin (2013).

$$\mathbf{R}_n(\gamma;\beta(.)) = n^{-1}\sum_{i=1}^{n}\int_0^{\tau_{max}} \frac{\mathbf{w}_i\{1-\pi(\gamma^\top \mathbf{w}_i)\}}{1-\pi(\gamma^\top \mathbf{w}_i)u}\Big[dN_i(\exp\{\mathbf{z}_i^\top\beta(u)\})- I[x_i \geq \exp\{\mathbf{z}_i^\top\beta(u)\}]H_\gamma(du|\mathbf{w}_i)\Big]. \quad (4.7)$$

To solve equation $\mathbf{R}_n(\gamma;\beta(.)) = \mathbf{0}$, the parameter vector $\beta(.)$ should be known. An equivalent equation for equation (4.7) is given below.

$$n^{-1}\sum_{i=1}^{n}\sum_{k=0}^{q_n-1}\mathbf{w}_i\Big[\frac{1-\pi(\gamma^\top \mathbf{w}_i)}{1-\pi(\gamma^\top \mathbf{w}_i)(\tau_k+\tau_{k+1})/2}\delta_i\times I[\exp\{\mathbf{z}_i^\top\hat{\beta}(\tau_k,\hat{\gamma}^{(m)})\} \leq x_i < \exp\{\mathbf{z}_i^\top\hat{\beta}(\tau_{k+1},\hat{\gamma}^{(m)})\}]- I[x_i \geq \exp\{\mathbf{z}_i^\top\hat{\beta}(\tau_k,\hat{\gamma}^{(m)})\}]\int_{\tau_k}^{\tau_{k+1}}\frac{1-\pi(\gamma^\top \mathbf{w}_i)}{1-\pi(\gamma^\top \mathbf{w}_i)u}H_\gamma(du|\mathbf{w}_i)\Big] = 0 \quad (4.8)$$

Following is the algorithm for estimating γ, and $\beta(.)$:

1. Choose an initial value $\hat{\gamma}^{(0)}$ for γ.
2. At the m^{th} iteration, set $\exp\{\mathbf{z}_i^\top\hat{\beta}(\tau_0,\hat{\gamma}^{(m)})\} = 0$ and obtain $\hat{\beta}(\tau_j,\hat{\gamma}^{(m)})$, for $j = 1,\ldots,q_n$, by sequentially minimizing (4.5).
3. Obtain $\hat{\gamma}^{(m+1)}$ by solving equation (4.8), using Newton-Raphson algorithm.
4. Repeat steps (2) and (3) to reach to a predetermined convergence criterion.

This algorithm may be sensitive to initial values. This issue has been discussed by Mao and Wang (2010) for a class of proportional odds cure rate models. Also, the entanglement of $\hat{\beta}(.)$, and $\hat{\gamma}$ makes the derivation of asymptotic properties difficult. To solve these challenges another nonparametric approach for estimating parameter γ has been introduced by Wu and Yin (2013).

The approach suggested by these authors estimates γ and β separately. The function F^* in (4.6) is replaced by a nonparametric estimator $\hat{F}^*$, and then γ is estimated accordingly. Also,

a Nelson–Aalen type estimator is considered for the cumulative hazard function, $\Lambda^*(t|\mathbf{z})$, of susceptible individuals

$$\hat{\Lambda}^*(t|\mathbf{z}) = \int_0^t \frac{\sum_{i=1}^n B_{ni}(\mathbf{z}) dN_i(u)}{\sum_{k=1}^n I(x_k \geq u) w_k(\hat{\gamma}, \hat{\Lambda}^*) B_{nk}(\mathbf{z})}, \tag{4.9}$$

where

$$w_k(\gamma, \Lambda^*) = \delta_k + (1-\delta_k)\frac{\pi(\gamma^\top \mathbf{w}_k)\exp\{-\Lambda^*(x_k|\mathbf{z})\}}{1-\pi(\gamma^\top \mathbf{w}_k)+\pi(\gamma^\top \mathbf{w}_k)\exp\{-\Lambda^*(x_k|\mathbf{z})\}}, \tag{4.10}$$

and $B_{ni}(z) = \frac{K_p\{(z-\mathbf{z}_i)/h_n\}}{\sum_{k=1}^n K_p\{(z-\mathbf{z}_k)/h_n\}}$. Here $K_p(.)$ is a p-variate kernel function with p equal to number of different covariates in $\mathbf{z}$, and $h_n > 0$ is the bandwidth converging to zero as $n \to 0$. Note that $\hat{F}^*(t|\mathbf{z})$ can be estimated by $\hat{F}^*(t|\mathbf{z}) = 1 - \exp(-\hat{\Lambda}^*(t|\mathbf{z}))$. Following the nonparametric approach is illustrated step by step.

At first an initial value for γ like γ^0 is assumed, and then plug γ^0 and $w_k = 1$, for $k = 1, \ldots, n$ in equation (4.9) to obtain $\hat{\Lambda}^{*0}(t|\mathbf{z})$ as the initial value for $\hat{\Lambda}^*(t|\mathbf{z})$. Then by plugging $\hat{\Lambda}^{*0}(t|\mathbf{z})$, and γ^0 into equation (4.10), w_k^0 is obtained. At the m^{th} iteration we perform:

1. Plug w_k^m into (4.9) to obtain $\hat{\Lambda}^{*(m+1)}(t|\mathbf{z})$.
2. Plug $\hat{\Lambda}^{*(m+1)}(t|\mathbf{z})$ into (4.6) and solve the equation by Newton-Raphson algorithm to obtain $\hat{\gamma}^{(m+1)}$.
3. Plug $\hat{\gamma}^{(m+1)}$ and $\hat{\Lambda}^{*(m+1)}(t|\mathbf{z})$ into (4.10) to get $w_k^{(m+1)}$.
4. Repeat steps 1, 2, and 3 until a pre-set convergence criterion is met.

For some identifiability and computational problems, $\hat{\Lambda}^*(t|\mathbf{z}) = \infty$ has been considered when t is greater than the largest observed failure time. The resultant estimator which is obtained from the above algorithm is denoted by $\hat{\gamma}_N$, and is used to plug in equation (4.5), to obtain $\hat{\beta}(\tau, \hat{\gamma}_N)$.

Finally, consider that $\hat{\beta}_I$ is obtained, as an estimate of β, and $\hat{\gamma}_I$ is an estimate of γ which has been obtained by one of the above algorithms. Under some conditions, which are given in Wu and Yin (2013), the resultant estimators are consistent in probability and have asymptotically normal distributions.

4.2.2 Two groups trial with semiparametric method

The material of this section is adopted from Shen et al. (2007).
In this section a semi-parametric cure rate model is introduced to analyse data which specifically consist of two different treatment groups. Suppose the effect of two different treatments is going to be studied by dividing the patients in two groups randomly, and applying each treatment to each group separately. We call these two groups, cohort 1 and cohort 2. Here, cohort 2 is the control group. The interest is to estimate the distribution of time to diagnosis of the disease, and the incidence of the disease within each group. Suppose there are two different cure rates and two different survival time distributions for each cohort. Consider that B is a variable which takes value 1 if the individual belongs to cohort 1 and takes value 2 otherwise. Furthermore, $p_1 = P(\eta = 1|B = 1)$ and $p_2 = P(\eta = 1|B = 2)$ are the probability of susceptibility for cohort 1 and cohort 2, respectively. Therefore, the probability of being cured in each cohort is $1 - p_1$ and $1 - p_2$.

Suppose $S_1^*(t|\eta = 1)$ is the survival function for the time to disease diagnosis, conditional on the ultimate occurrence of the disease in cohort 1. Consequently, $S_1(t)$ is an improper survival function for cohort 1, which follows the mixture cure model of equation (3.15), with p_1 replacing p in this equation. Subsequently, the same formula applies for cohort 2 of patients with S_1 replacing S_2, and S_1^* replacing S_2^*.
Mixture model in each group allows to study the distribution of time to disease diagnosis or the probability of incidence in each group separately, but finding a pattern to fit the data or connect the two groups of data together is needed. When the nonparametric discrete density curves for the time to disease diagnosis in each group of data show similar shapes, but in different scales for the two treatment groups, the change in the time to disease diagnosis between the two cohorts could be shown by the natural logarithm of the two densities ratio fitted in a specific form.
Suppose f_1^* and f_2^* are the proper probability density functions of time-to-disease diagnostic in cohort 1 and cohort 2 respectively. The following equation shows the connection between these two probability density functions:

$$f_2^*(t) = \exp\{\alpha^\star + \beta h(t)\} f_1^*(t), \tag{4.11}$$

where h is a specific function for transforming the observed time t, and $\alpha^\star$ and β are a constant and a vector of coefficients respectively. Note that $\alpha^\star$ is a new notation to emphasize difference between $\alpha^\star$ and α which is introduced later.
The model (4.11) presents a natural logistic regression connection between cohort 1 and cohort 2, or equivalently giving the following equation.

Note that in the following equations T represents the group of time-to-disease diagnostic in both cohorts, and t represents a specific time-to-disease diagnostic.

$$\begin{aligned}
P(B=2|t) &= \frac{P(B=2,T=t)}{P(T=t)} \\
&= \frac{P(B=2,T=t)}{P(T=t,B=2)+P(T=t,B=1)} \\
&= \frac{P(B=2)P(T=t|B=2)}{P(B=1)P(T=t|B=1)+P(B=2)P(T=t|B=2)} \\
&= \frac{P(B=2)f_2^*(t)}{P(B=1)f_1^*(t)+P(B=2)f_2^*(t)} \\
&= \frac{P(B=2)\exp\{\alpha^\star+\beta h(t)\}f_1^*(t)}{P(B=1)f_1^*(t)+P(B=2)\exp\{\alpha^\star+\beta h(t)\}f_1^*(t)} \\
&= \frac{\frac{P(B=2)}{P(B=1)}\exp\{\alpha^\star+\beta h(t)\}}{1+\frac{P(B=2)}{P(B=1)}\exp\{\alpha^\star+\beta h(t)\}}.
\end{aligned}$$

Thus, if we consider $\alpha_1 = \alpha^\star + \log\{\frac{P(B=2)}{P(B=1)}\}$, then $P(B=2|t) = \frac{\exp\{\alpha_1+\beta h(t)\}}{1+\exp\{\alpha_1+\beta h(t)\}}$, and it has the form of logistic regression.

Model (4.11) with a proper choice of h provides a good fit to the data. This model needs to be fitted to right-censored data with a cure rate fraction. In order to illustrate the full likelihood function according to the model (4.11) for censored data, assume $x_1,\ldots,x_{n_0}$ are the observed times in cohort 1 of patients and $x_{n_0+1},\ldots,x_n$, $n=n_0+n_1$, are observed time in control group or cohort 2. Therefore,

$$\begin{aligned}
\mathcal{L}(\theta) = &\prod_{i=1}^{n_0}\{p_1 f_1^*(x_i)\}^{\delta_i}\{(1-p_1)+p_1 S_1^*(x_i)\}^{1-\delta_i}\times \\
&\prod_{j=n_0+1}^{n}\{p_2 f_2^*(x_j)\}^{\delta_j}\{(1-p_2)+p_1 S_2^*(t_j)\}^{1-\delta_j},
\end{aligned} \tag{4.12}$$

in which θ is statistical parameter vector related to $f_1^*(t)$, $f_2^*(t)$, $S_1^*(t)$ and $S_2^*(t)$.
The equation is the full likelihood function for mixture cure rate models with two groups sample datasets. In general, estimating the unknown parameters requires EM algorithm or one of the MCMC algorithms which could be computationally expensive. Therefore, a conditional estimating approach by generalizing the profile likelihood method is applied; see Qin and Zhang (1997).

An alternative method to the full likelihood maximization is based on estimating equations conditional on the observed failure only ($\delta = 1$). Suppose f_1 represents the probability density function of T conditional on $\delta = 1$ in the cohort 1 of patients, and f_2 is the probability density function of T conditional on $\delta = 1$ in the cohort 2 of patients. So,

$$f_1(t) = \lim_{\Delta t \to 0} \frac{P(\Delta t < T < t + \Delta t | \delta = 1)}{\Delta t} \tag{4.13}$$

$$= \frac{\lim_{\Delta t \to 0} \frac{P(\Delta t < T < t + \Delta t, C > t | \delta = 1)}{\Delta t}}{P(T < C)} \tag{4.14}$$

$$= \frac{p_1 f_1^*(t) \bar{H}_1(t)}{\int_0^\infty p_1 f_1^*(v) \bar{H}_1(v) dv} \tag{4.15}$$

$$= \frac{f_1^*(t) \bar{H}_1(t)}{\mu(f_1^*, H_1)}, \tag{4.16}$$

where $\bar{H}_1(t)$ is the survival function of censoring time variable C, and

$$\mu(f_1^*, H_1) = \int_0^\infty p_1 f_1^*(v) \bar{H}_1(v) dv$$

is a constant with unknown quantities for f_1^* and H_1. The density $f_2(t)$ is the equivalent definition of f_1 for cohort 2 of patients. Thus,

$$f_2(t) = \frac{f_2^*(t) \bar{H}_2(t)}{\mu(f_2^*, H_2)}.$$

Under the model (4.11),

$$\frac{f_2(t)}{f_1(t)} = \exp\{\alpha + \beta h(t) + \psi(t)\}, \tag{4.17}$$

where $\alpha = \alpha^\star + \log\{\mu(f_1^*, H_1) - \mu(f_2^*, H_2)\}$, and $\psi(t) = \log \bar{H}_2(t) - \log \bar{H}_1(t)$.

Let $t_1^*, \ldots, t_{m_0}^*$, $t_{m_0+1}^*, \ldots, t_m^*$ denote the observed time-to-disease diagnosis in cohort 1 and cohort 2, among $x_1, \ldots, x_{n_0}$, and $x_{n_0+1}, \ldots, x_n$, where $m = m_0 + m_1$, and m_0 and m_1 are the number of failures in cohort 1 and cohort 2. Suppose l_{con} is the corresponding conditional log-likelihood:

$$\ell_{\text{con}}(\alpha, \beta, F^*) = \sum_{i=1}^{m} \log f_1(t_i^*) + \sum_{j=m_0+1}^{m} \{\alpha + \beta h(t_j^*) + \psi(t_j^*)\}. \tag{4.18}$$

The conditioning is over the following constraints:

$$\sum_{i=1}^{m} f_1(t_i^*) = 1, \tag{4.19}$$

$$\sum_{i=1}^{m} f_1(t_i^*)\{\exp(\alpha + \beta h(t_i^*) + \psi(t_i^*)) - 1\} = 0. \tag{4.20}$$

These two constraints are to guarantee that $\int f_1(x)dx = 1$, and $\int g_1(x)dx = 1$. When the censoring distribution of two groups are the same, $\psi(t^*) = 0$. Therefore maximizing the equation (4.18) is equivalent to maximizing the ordinary logistic regression likelihood, and the estimators of α and β are consistent. But when $\psi(t^*) \neq 0$, the following score equations hold for (α, β):

$$\frac{\partial \ell_{\text{con}}}{\partial \alpha} = m_1 - \sum_{i=1}^{m} \frac{\rho \exp\{\alpha + \beta h(t_i^*) + \psi(t_i^*)\}}{1 + \exp\{\alpha + \beta h(t_i^*) + \psi(t_i^*)\}},$$
$$\frac{\partial \ell_{\text{con}}}{\partial \beta} = \sum_{i=m_0+1}^{m} h(t_i^*) - \sum_{i=1}^{m} \frac{\rho h(t_i^*) \exp\{\alpha + \beta h(t_i^*) + \psi(t_i^*)\}}{1 + \exp\{\alpha + \beta h(t_i^*) + \psi(t_i^*)\}},$$

where $\rho = m_1/m_0$.

The estimate of (α, β) is obtained by solving the equations $\frac{\partial l_c}{\partial \alpha} = 0$ and $\frac{\partial l_c}{\partial \beta} = 0$. These estimators are unbiased and consistent (Shen et al., 2007). In general, ψ is unknown; therefore it is replaced by its consistent estimator, $\hat{\psi} = \log(\hat{H}_2) - \log(\hat{H}_1)$, where $\hat{H}_1$ and $\hat{H}_2$ are the Kaplan-Meier estimators for the survival functions of the censoring variables in cohort 1 and cohort 2 of patients, respectively.

After obtaining $(\hat{\alpha}, \hat{\beta})$, the other unknowns are estimated:

$$\hat{f}_1(t_i^*) = m_0^{-1}\{1 + \rho \exp\{\hat{\alpha} + \hat{\beta} h(t_i^*) + \hat{\psi}(t_i^*)\}^{-1}, \tag{4.21}$$

$$\hat{f}_2(t_i^*) = \exp(\hat{\alpha} + \hat{\beta} h(t_i^*) + \hat{\psi}(t_i^*))\hat{f}_1(t_i^*). \tag{4.22}$$

Since $dF_1(t) = \frac{dF_1^*(t)\bar{H}_1(t)}{\mu(f_1^*, H_1)}$ and $\int_0^\infty dF_1^*(t) = 1$, it follows that the constant μ can be estimated by

$$\hat{\mu}(f_1^*, H_1) = \frac{1}{\int \hat{H}_1^{-1}(t) d\hat{F}_1(t)} = \left(\sum_{i=1}^{m} \frac{\hat{f}_1(t_i^*)}{\hat{\bar{H}}_1(t_i^*)} \right)^{-1}, \tag{4.23}$$

where $\hat{\bar{H}}_1(t_i^*) = 1 - \hat{H}_1(t_i^*)$.
Also, $F_1^*(t)$ can be estimated by $\hat{F}_1^*(t) = \hat{\mu}(f_1^*, H_1) \sum_{i=1}^m \hat{\bar{H}}_1^{-1}(t_i^*) \hat{f}_1(t_i^*) I(t_i^* \leq t)$, and in the same way

$$\hat{F}_2^*(t) = \hat{\mu}(f_2^*, H_2) \sum_{i=1}^m \hat{\bar{H}}_2^{-1}(t_i^*) \hat{f}_2(t_i^*) I(t_i^* \leq t),$$

where

$$\hat{\mu}(f_2^*, H_2) = \left(\sum_{i=1}^m \frac{\hat{f}_2(t_i^*)}{\hat{\bar{H}}_2(t_i^*)} \right)^{-1}.$$

To estimate the cumulative incidence in each group, the non-parametric conventional Kaplan-Meier estimator is suggested. Suppose $1 - \tilde{F}_{n_0}$ and $1 - \tilde{F}_{n_1}$ are survival Kaplan-Meier estimators of cohort 1 and cohort 2 ($\tilde{F}_{n_0}$ is an estimator for F). According to equation (3.3), $\hat{p}_1 = \tilde{F}_{n_0}(t_{(n_0)})$ and $\hat{p}_2 = \tilde{F}_{n_1}(t_{(n_1)})$, where $t_{(n_0)}$ and $t_{(n_1)}$ are the last observed failure time and censored time for cohort 1 and 2.

4.3 Non-mixture models

Another method for modelling time-to-event data is non-mixture cure models. Despite the fact that standard cure rate model is more attractive and has been widely used, it has some disadvantages such as it cannot have a proportional hazard structure; see Ming-Hui Chen (1999). This disadvantage can be overcome with an alternative definition of cure rate model. Although non-mixture cure models have attracted less attention compared to mixture cure models, they remain useful in a number of particular cases, especially those related to cancer studies.

Among parametric, nonparametric, and semi-parametric methods for developing the non-mixture cure models, it seems semi-parametric methods are applied more often. Therefore, it is adequate to explain one of the semi-parametric models to illustrate the rationale behind this modelling technique. In this section an alternative approach to cure models is introduced and resimulate the method of Ming-Hui Chen (1999).

4.3.1 Semi-Parametric non-mixture models

In non-mixture modelling formulation, the latent variable, η, of mixture model is replaced by a new unknown variable N, with a new definition. Suppose N denotes the number of defective tumor cells for an individual in the population which have been left active after the initial treatment. A defective tumor cell is a tumor cell which has the potential of metastasizing. Consider N has a Poisson distribution with parameter μ. Suppose T_i is the random time variable for the ith defective tumor cell to produce detectable metastatic disease. Given N, the random variables T_i, $i = 1, 2, \ldots$ are assumed to be independent and identically distributed with a common distribution function $F_T(t)$ that does not depend on N. The time to relapse of cancer can be defined by random variable $X = \min\{T_i, 0 \leq i \leq N\}$, where $P(T_0 = \infty) = 1$. Note that in this section T and X have slightly different definitions from the previous one, and T_0 represents time to detectable metastatic when there is no defective tumor cell. The survival function for X is given by

$$S(t) = P(\text{no metastatic cancer by time } t) = P(N = 0) + P(T_1 > t, \ldots, T_N > t, N \geq 1).$$

After some algebra we have:

$$\begin{aligned} S(t) &= \exp(-\mu) + \sum_{k=1}^{\infty} S_T(y)^k \frac{\mu^k}{k!} \exp(-\mu) \\ &= \exp(-\mu + \mu S_T(t)) \\ &= \exp(-\mu F_T(t)), \end{aligned} \tag{4.24}$$

where $S(t) = \exp(-\mu F_T(t))$ is an improper survival function, because $S(\infty) = \exp(-\mu) > 0$. Therefore the cure fraction is $S(\infty) \equiv P(N = 0) = \exp(-\mu)$. This model is suitable for survival data that include cured fraction. According to formulation (4.24),

$$f(t) = \mu f_T(t) \exp\{-\mu F_T(t)\} \tag{4.25}$$

and

$$h(t) = \mu f_T(t). \tag{4.26}$$

Here, $f(t)$ is not a proper probability density, but $f_T(t)$ is. The hazard function obtained by this formulation has the proportional hazard structure. This form of the hazard function is more common in analysis of survival data than the one from standard cure model in case of application and computation.

The survival function for the susceptible population is given by

$$\begin{aligned} S^*(t) &= P(X > t | N \geq 1) \\ &= \frac{\exp(-\mu F_T(t)) - \exp(-\mu)}{1 - \exp(-\mu)}. \end{aligned} \tag{4.27}$$

Because $S^*(0) = 1$ and $S^*(\infty) = 0$, it follows that $S^*(t)$ is a proper survival function. The survival density and hazard function for the susceptible fraction of individuals are

$$\begin{aligned} f^*(t) &= -\frac{d}{dt} S^*(t) \\ &= \Big[\frac{\exp\{-\mu F_T(t)\}}{1 - \exp(-\mu)}\Big] \mu f_T(t) \end{aligned} \tag{4.28}$$

and

$$h^*(t) = \frac{h(t) \exp\{-\mu F_T(t)\}}{\exp\{-\mu F_T(t)\} - \exp(-\mu)} = \frac{h(t)}{P(X < \infty | X > t)}. \tag{4.29}$$

According to the equation (4.27) and equation (4.24), it can be shown that

$$S(t) = \exp(-\mu) + \{1 - \exp(-\mu)\} S^*(t). \tag{4.30}$$

By comparing equation (3.15) and equation (4.30), it can be seen that there is a natural connection between standard cure model and non-mixture cure model. In other words one may reconstruct equation (3.15) from equation (4.30) replacing p by $\exp(-\mu)$.

We incorporate the covariates in the model (4.24) through μ according to the relationship $\mu = \exp(\mathbf{z}^\top \beta)$, where $\mathbf{z}$ is a $p \times 1$ vector of covariates and β is a $p \times 1$ vector of regression coefficients. By this relationship between covariates and cured fraction, the regression coefficients become interpretable for cured and non-cured fractions. For cured fraction, the negative sign of regression coefficient leads to a larger cure fraction for positive covariates value. For the non-cured fraction the regression coefficients affect the hazard function in (4.29). For example, a negative regression coefficient leads to a larger hazard, for a positive covariate.
The likelihood function can be constructed as follows. Suppose n individuals are under the study, and let N_i denote the number of metastasis-competent tumor cells for ith individual. Further, suppose that N_i's are Poisson random variables with mean μ, for $i = 1, 2, \ldots, n$, and N_i's are not observed. Let $T_{i1}, T_{i2}, \ldots, T_{i,N_i}$, be i.i.d. time-to-metastasis for the i^{th} individual, which are unobserved and have the same proper cumulative distribution function, $F_T(.)$. We

will specify a parametric form for $F_T(.)$, such as Weibull distribution. This distribution has its indexing parameter vector, ψ, thus there are two new notations: $F_T(.|\psi)$ and $S_T(.|\psi)$. By assuming Weibull distribution for $F_T(.|\psi)$, the indexing parameter vector is $\psi = (\alpha, \lambda)^\top$, where α is the shape parameter and λ is the scale parameter of the Weibull distribution. Our observed data is $\mathbf{O} = (n, x, \delta)$, where x_i denotes the observed survival time for ith subject, and δ_i denotes the censoring indicator for ith individual, $i = 1, 2, \ldots, n$. The complete data is given by $\mathbf{D} = (n, x, \delta, N)$, where N is an unobserved variable. The complete data likelihood function of the parameter $\theta = (\psi, \beta)$ can be written as

$$\mathcal{L}(\theta|\mathbf{D}) = \Big[\prod_{i=1}^{n} S_T(x_i|\psi)^{N_i-\delta_i}\{N_i f_T(x_i|\psi)\}^{\delta_i}\Big] \times \exp\left[\sum_{i=1}^{n}\{N_i \mathbf{z}_i\beta_i - \log(N_i!) - \exp(\mathbf{z}_i^\top\beta)\}\right]. \tag{4.31}$$

Since the covariates are weighted in this model through μ, this implies that each individual has different cure rates, $\mu_i \equiv \exp(\mathbf{z}_i^\top\beta)$. To estimate the corresponding parameters in this likelihood function there are several methods like the EM algorithm and the MCMC methods. In Ming-Hui Chen (1999), one of the MCMC methods is implemented. To run the MCMC method of Ming-Hui Chen (1999), one needs to sample from the following complete conditional posterior distributions:

i) Sample $\beta, N|\psi, \mathbf{O}$.

ii) Sample $\psi|\beta, N, \mathbf{O}$.

By applying the collapsed Gibbs procedure, we have:

$$[\beta, N|\psi, \mathbf{O}] = [\beta|\psi, \mathbf{O}][N|\beta, \psi, \mathbf{O}]. \tag{4.32}$$

Briefly the MCMC algorithm samples from the posterior is:

1. Sample $N_i \sim \text{Pois}\left\{S_T(x_i|\psi)\exp(\mathbf{z}_i^\top\beta)\right\} + \delta_i$.
   (In this step we add δ_i to the Poison distribution, because when $\delta_i = 1$, i.e. failure happened for subject ith, then the number of defective cells should be at least 1.)

2. Sample β from the following conditional posterior density:

$$\pi[\beta|\psi, \mathbf{O}] \propto \exp\left\{\sum_{i=1}^{n}[\delta_i \mathbf{z}_i^\top\beta - \exp(\mathbf{z}_i^\top\beta)\{1 - S_T(x_i|\psi)\}]\right\} \times \pi_0(\theta).$$

3. Sample $\psi = (\alpha, \lambda)^\top$ from the following joint conditional posterior density:

$$\pi[\psi|\beta, \mathbf{N}, \mathbf{O}] \sim \prod_{i=1}^{n} S_T(x_i|\psi)^{N_i - \delta_i} f_T(x_i|\psi)^{\delta_i} \times \pi_0(\theta),$$

where $\pi_0(\theta)$ is the initial prior distribution for $\theta = (\beta, \psi)$. Furthermore, assume that $\pi_0(\theta)$ is log-concave. A gamma prior for α with a small shape and scale parameters, and an independent normal prior for λ with mean 0 and variance c are often suggested. In most cases $\pi[\psi|\beta, \mathbf{N}, \mathbf{O}]$ and $\pi[\beta|\psi, \mathbf{O}]$ are log-concave in each component of β and ψ respectively, therefore the adaptive rejection algorithm can be used to sample from these posterior distributions efficiently. This model is fast in convergence and is compatible with cancer datasets.

CHAPTER 5 SIMULATION

5.1 Introduction

In this section we conduct simulations to test three of the cure rate models proposed in the thesis. Even though cure rate models have been used in different areas ranging from medicine to reliability, there is only a few computational packages available for fitting these models. There are two `R` packages called `smcure`, and `NPHMC`, currently available on CRAN. The `smcure` fits semi-parametric proportional hazard mixture cure rate models and semi-parametric accelerated failure time mixture cure rate models. The `NPHMC` can helps simulate cure datasets. For more recent models such as mixture quantile regression cure models, two groups mixture cure models, and non-mixture cure models there is no package to ease data simulation.
Here, three classes of semi-parametric models, namely mixture quantile regression cure rate model, two groups mixture model, and non-mixture model are fitted and then a Monte Carlo study is conducted for each model to examine their accuracy.

5.1.1 Cure rate quantile regression model

To study finite-sample performance of the quantile regression method, we run the set up of Wu and Yin (2013). The simulation setting is briefly explained bellow, see Wu and Yin (2013) for more details.

As discussed in Section 4.2.1, two groups of covariates are simulated. The matrix $\mathbf{w}$ is the covariates associated with susceptibility of each individual and the matrix $\mathbf{z}$ is covariates related to the survival time of susceptible individuals. The matrix $\mathbf{z}$ includes a two columns. One column is a vector of ones, and other column is $\mathbf{z}$, simulated from Bernoulli(0.5). We set $\mathbf{z} = \mathbf{w}$, $\gamma_0 = 1$ and $\gamma_1 = -0.5$ to generate susceptibility indicator η from equation (3.17). Failure time of susceptible individuals is simulated from the following log-transformed linear model

$$\log T^* = b_0\mathbf{z} + (1 + \mathbf{z})\varepsilon,$$

in which $b_0 = -1$, and $(1+\mathbf{z})\varepsilon$ is the error term. Given $\mathbf{z}$ the corresponding quantile regression

model of equation 4.3 is:

$$Q_{T^*}(\tau|\mathbf{z}) = \exp(\beta_0(\tau) + \beta_1(\tau)z),$$

where $\beta_0(\tau) = Q_\varepsilon(\tau)$, $\beta_1(\tau) = b_0 + Q_\varepsilon(\tau)$ and $Q_\varepsilon(\tau)$ is the τ^{th} quantile regression of ε. Censoring time for each individual is simulated from Unif$(0, L+2)$ if $z < 0.5$, and from Unif$(1, L+2)$ if $z \geq 0.5$, where L is the duration of the study. We simulated two sets of individuals, one with $n = 100$ and the other with $n = 400$, with the Monte Carlo replication of 1000 times. The bootstrap method for estimation of standard errors was used, with 200 samples. The results of simulations is given in Tables 5.1 to 5.2. The `R` code related to this simulation is given in Appendix A

Table 5.1 Estimation of γ, under the iterative quantile regression method. The error term for the following simulation is the standard normal distribution. True indicates true value of the parameter. EST indicates estimated value. SE shows standard error, MSE indicates mean square error, and Bias shows the bias for the estimation.

	$n = 100$			$n = 400$		
True value	$\text{EST}_{(\text{SE})}$	MSE	*Bias*	$\text{EST}_{(\text{SE})}$	MSE	Bias
$\gamma_0 = 1$	$0.8_{(0.22)}$	0.07	0.15	$0.8_{(0.32)}$	0.11	0.11
$\gamma_1 = -0.5$	$-0.4_{(0.13)}$	0.02	-0.02	$-0.5_{(0.25)}$	0.06	0.03

Table 5.2 Simulation results for cure rate quantile regression. This table shows the estimation of quantile coefficients for different τ's, in different simulation sets.

		$n = 100$			$n = 400$		
τ	True value	$\text{EST}_{(\text{SE})}$	MSE	Bias	$\text{EST}_{(\text{SE})}$	MSE	Bias
0.2	$\beta_0 = -1.17$	$-1.9_{(0.23)}$	0.60	0.74	$-1.3_{(0.34)}$	0.16	0.21
0.4	$\beta_0 = -0.17$	$0_{(0.18)}$	0.04	-0.1	$-0.2_{(0.25)}$	0.06	0.03
0.6	$\beta_0 = 0.15$	$0.2_{(0.16)}$	0.03	-0.05	$0.2_{(0.40)}$	0.16	-0.03
0.2	$\beta_1 = -2.17$	$-2.5_{(0.31)}$	0.21	0.34	$-2.3_{(0.43)}$	0.22	0.21
0.4	$\beta_1 = -1.17$	$-0.9_{(0.26)}$	0.10	-0.19	$-1.3_{(0.38)}$	0.17	0.16
0.6	$\beta_1 = -0.85$	$-1.2_{(0.17)}$	0.15	0.35	$-0.5_{(0.27)}$	0.19	-0.34

Proposed cure rate quantile has 28% of non-convergence cases on average; see Wu and Yin

(2013). Implementing quantile method requires extensive coding skill and there is no ready-to-use package to fit this method. This iterative approach is relatively unstable and sensitive to initial values.

5.1.2 Two groups mixture model

This simulation set up is motivated from Shen et al. (2007). In this simulation study the point estimators and their standard error of α and β are estimated under model 4.11 when $h(t) = t$. The p_1 and p_2 are estimated non-parametrically. Simulation is repeated with 1000 Monte Carlo replications with sample size 100 or 400.
We set $p_1 = 0.2$ or 0.3 and $p_2 = 0.2$ as true values for p_1 and p_2. We assumed that $f^*(t)$, the distribution of failure times for group 1 of patients, is an exponential distribution with mean ω. Censoring times for both groups have been generated from the same uniform distribution. The functions g^* and f^* are computed according to the model 4.11, so g^* is also an exponential distribution with mean ν; then,

$$\alpha = \log(\omega) - \log(\nu),$$
$$\beta = 1/\omega - 1/\nu.$$

We set $\omega = 1.5$ and $\nu = 1.5, 2.5, 3$; then we obtain different combinations of α, β to use as the true values. The results of this simulation are presented in Table 5.3. The `R` codes related to this simulation is given in Appendix B.

5.2 Non-mixture models

As in Chapter 4 the cured probability for non-mixture models is obtained from the following equation:

$$p = \exp\{-\exp(\mathbf{z}^\top \beta)\},$$

and depends on each individual. Another simulation is performed, once with $n = 100$, and another time with $n = 400$ for non-mixture models. $\mathbf{z}_1$ is simulated from standard normal distribution. Both parameters β_0 and β_1 are assumed equal to 1. The cure probability for each individual is simulated from equation $p_i = \exp(-\mu_i)$, where p_i is the cured probability

for the ith subject, $i = 1, \ldots, n$. If $F_T(t)$ is known, then survival time for each individual is

$$S(t) = \exp(-\mu F_T(t)). \tag{5.1}$$

The $U = F_T(t)$ is simulated with uniform(0,1), then from the equation (5.1) the value of $S(t)$ is obtained. Suppose the survival probability of each individual is Weibull distribution with the shape parameter equal to 1 and the scale parameter equal to $\exp(-1.5)$. Then the survival time for each individual is calculated by using the inverse of Weibull distribution function. Censoring indicator for each individual is defined by $I(U_i \leq P_i)$.
The MCMC algorithm, of Chapter (4), is iterated 2000 times with 1000 burning samples. Table 5.4 shows that the estimates of parameters in non-mixture model have quite smaller variances compared to standard cure rate models. The `R` codes related to this simulation is given in Appendix C.

Table 5.4 The result of fitting a non-mixture cure model on simulated data. Censoring rate in this case is approximatly between 0.40% to 0.70%, and the link function is equal to $p = \exp(-\exp(\mathbf{z}^\top \beta))$.

	$n = 100$			$n = 400$		
True value	$\text{EST}_{(\text{SE})}$	MSE	Bias	$\text{EST}_{(\text{SE})}$	MSE	Bias
$\beta_0 = 1$	$0.8_{(0.18)}$	0.05	0.14	$0.9_{(0.09)}$	0.01	0.09
$\beta_1 = 1$	$1_{(0.16)}$	0.04	0.16	$1_{(0.09)}$	0.01	-0.08

We have repeated the simulation sets of the original articles for each proposed model in Chapter 4. The results of Table 5.1, Table 5.2, Table 5.3, and Table 5.4 show that implementation of these models have results close to original papers. The calculated variance for parameters in each simulation set becomes smaller as the sample size increases, which implies consistency of parameters in each model.

Table 5.3 Parameter estimation for two groups mixture cure rate model. Here $\omega = 1.5$ for all sets.

			$n_1 = n_2 = 100$			$n_1 = n_2 = 400$		
		True value	$\text{EST}_{(\text{SE})}$	MSE	Bias	$\text{EST}_{(\text{SE})}$	MSE	Bias
$p_1 = 0.2$	$\nu = 1.5$	$\alpha = 0$	$0_{(0.02)}$	0.05	-0.23	$0_{(0.01)}$	0	0.001
		$\beta = 0$	$0_{(0.11)}$	0.012	0	$0_{(0.07)}$	0.005	0
		$p_1 = 0.2$	$0.2_{(0.09)}$	0.01	-0.02	$0.2_{(0.05)}$	0.003	-0.02
		$p_2 = 0.2$	$0.2_{(0.1)}$	0.012	-0.05	$0.2_{(0.06)}$	0.005	-0.04
	$\nu = 2.5$	$\alpha = -0.512$	$-0.5_{(0.34)}$	0.12	0.04	$-0.5_{(0.17)}$	0.03	0.03
		$\beta = 0.267$	$0.2_{(0.38)}$	0.14	-0.01	$0.2_{(0.32)}$	0.10	-0.01
		$p_1 = 0.2$	$0.2_{(0.09)}$	0.01	-0.007	$0.2_{(0.08)}$	0.01	-0.01
		$p_2 = 0.2$	$0.2_{(0.13)}$	0.02	-0.06	$0.2_{(0.1)}$	0.01	-0.05
	$\nu = 3$	$\alpha = -0.693$	$-0.7_{(0.49)}$	0.24	0.07	$-0.7_{(0.36)}$	0.13	0.05
		$\beta = 0.333$	$0.3_{(0.08)}$	0.006	0.01	$0.3_{(0.04)}$	0.002	-0.016
		$p_1 = 0.2$	$0.2_{(0.06)}$	0.003	-0.003	$0.1_{(0.04)}$	0.002	0.015
		$p_2 = 0.2$	$0.2_{(0.12)}$	0.02	-0.065	$0.1_{(0.08)}$	0.007	0.022
$p_1 = 0.3$	$\nu = 1.5$	$\alpha = 0$	$0_{(0.06)}$	0.005	-0.04	$0_{(0.05)}$	0.003	-0.03
		$\beta = 0$	$0_{(0.28)}$	0.08	0.027	$0_{(0.24)}$	0.06	0.02
		$p_1 = 0.3$	$0.3_{(0.15)}$	0.024	-0.04	$0.3_{(0.09)}$	0.01	-0.04
		$p_2 = 0.2$	$0.2_{(0.08)}$	0.008	-0.04	$0.2_{(0.07)}$	0.005	0.01
	$\nu = 2.5$	$\alpha = -0.512$	$-0.5_{(0.29)}$	0.08	0.02	$-0.5_{(0.25)}$	0.06	0.04
		$\beta = 0.267$	$0.2_{(0.41)}$	0.17	0.004	$0.2_{(0.39)}$	0.15	0
		$p_1 = 0.3$	$0.3_{(0.15)}$	0.025	-0.05	$0.3_{(0.12)}$	0.016	-0.05
		$p_2 = 0.2$	$0.2_{(0.08)}$	0.01	-0.031	$0.2_{(0.05)}$	0	-0.03
	$\nu = 3$	$\alpha = -0.693$	$-0.7_{(0.24)}$	0.06	0.07	$-0.7_{(0.21)}$	0.048	0.06
		$\beta = 0.333$	$0.3_{(0.37)}$	0.14	-0.013	$0.3_{(0.3)}$	0.1	-0.011
		$p_1 = 0.3$	$0.3_{(0.1)}$	0.012	-0.05	$0.3_{(0.07)}$	0.007	-0.05
		$p_2 = 0.2$	$0.1_{(0.07)}$	0.005	0.022	$0.2_{(0.04)}$	0.002	-0.02

CHAPTER 6 APPLICATION

6.1 Data summary

In this chapter cure rate models of Chapter 3 and Chapter 4 are fitted to Bone Marrow Transplantation (BMT) dataset obtained from a study on 137 leukemia patients. Bone marrow transplantation is one of the most common remedies for critical leukemia patients. A successful transplantation may depend on several factors, such as patient or donor age, sex, status of disease of the patient at the time of transplantation, patient's time of diagnostic, patient's post-transplantation recovery history, etc. Transplantation is considered to be unsuccessful when the patient's leukemia relapses or when the patient dies. The goal of this study is to illustrate the probability of relapse or death for the patients who were under bone marrow transplantation therapy.

The data were collected between March 1, 1984 to June 30, 1989, and is used by Copelan et al. (1991) to study the effects of pre-transplantation treatment on failure rate of transplantation. John P. Klein (1990) also used this dataset to develop a partially parametric estimator of survival data. The patients from age 7 to 55 were treated with allergenic bone marrow transplantation at five separate hospitals in Australia and the United States. Allergenic bone marrow transplantation (BMT) following high-dose chemotherapy and total body irradiation (TBI) cures many patients with acute myelocytic leukemia (AML). All donors were siblings of recipients.

Different preparation regimes for transplant were applied to each patient. This results in various factors affecting the transplantation process. The dataset contains 22 variables, of which 10 variables report one potential risk factor. These risk factors increase a person's chance of developing a disease. The rest of the variables indicate different preparation process for every patient, time-to-event variables, and failures indicators. The risk factor *group* reports status of disease and different preparation regimes applied to different patients before the surgery. The *group* equals 1 for acute Lymphocytic Leukemia (ALL), containing 38 patients. The *group* equals 2 for low risk acute Lymphocytic Leukemia which consists of 54 patients. The *group* equals 3 for high risk acute Lymphocytic Leukemia with 45 patients. The survival time t, in days, indicates the shortest of the times to death, relapse or date of the most recent follow-up time for each patient. The variable d is the patient failure indicator. This is another categorical variable which takes value 1 if the patient is dead or had disease relapse,

and equals 0 if patient survived by the end of the study. Some other factors, like age and sex of the patients and donors, waiting time to transplant in days, are also recorded.

Among 137 patients in this study, 82 patients died or relapsed (the rate of failure in overall groups are about 60%). For group *ALL*, there were 24 failures (63%), for group *AML low risk* there were 25 failures (46%), and for group *AML high risk* there were 34 failures (75%).

6.2 Survival models

The response variable in this dataset is time to death or relapse, and the dataset contains censored individuals. As the first step in analysis, Kaplan-Meier estimator is calculated to provide an overall idea about the survival function of the population. Then, the effects of various potential risk factors are estimated using Cox regression. Covariates highly correlated with the survival time have mostly been selected.

As developed in Chapter 2, Kaplan-Meier estimator is a non-parametric method for estimating survival probability of each individual up to certain time. Figure 6.1 shows that the survival probability for low risk group of patients is higher than the high risk group of patients as expected. At the beginning of the study, patients who are in *AML high risk* group have higher risk of mortality and relapse. Patients in *AML low risk* group have greater probability of survival. Figure 6.1 shows that *All* group of patients and *AML high risk* group of patients have flat survival curve after 700 days. This means no more deaths are observed for these patients after some point.

In Figure 6.2, the cumulative hazard function for each group of patients is illustrated. This plot demonstrates *AML high risk* group of patients have higher risk of death during this trial.

6.2.1 Cox proportional hazard model

To examine the effects of covariates on the outcome (which is survival time for each individual) the Cox proportional hazard model is applied. There are ten variables in the dataset and fitting advanced cure models with all 10 variables is complicated. Among all the covariates the two which have the highest correlation with the outcome variable are selected for our analysis of the data.

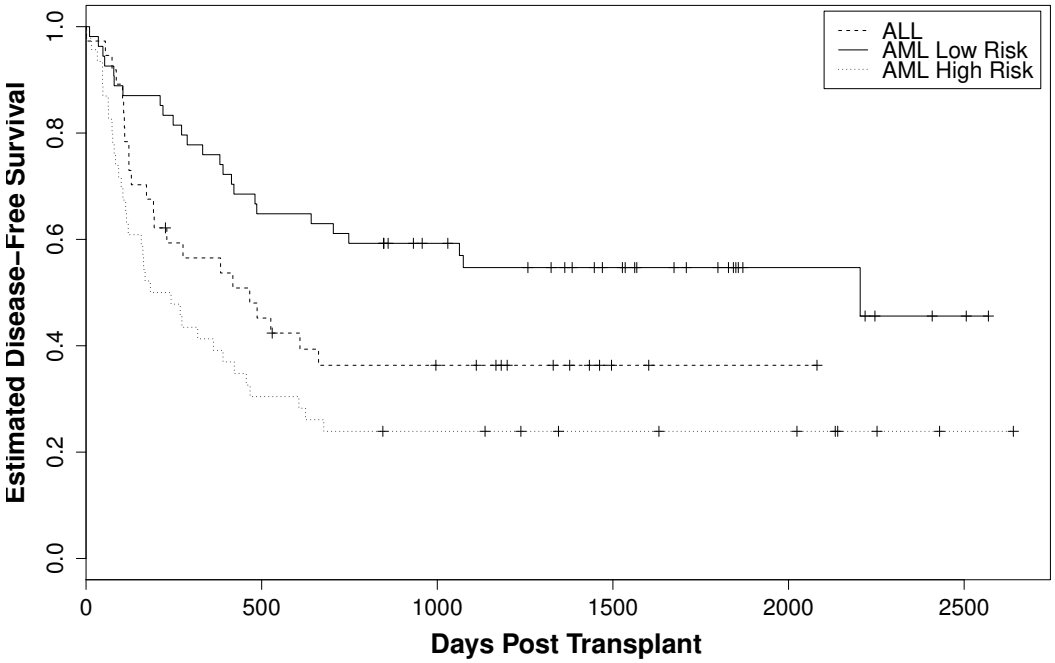

Figure 6.1 Kaplan-Meier estimate for each group of patients in BMT dataset.

Let $(z_1, \ldots z_{10})$ denote the vector of covariates where z_1 and z_2 are respectively the age of the patient and the age of the donor. The donor-recipient age has a clear effect on the survival time of recipient in transplantation surgeries (Douglas et al., 2004). We add another covariate to capture the interaction between donor age and recipient age. The z_{12} is the multiplicative variable of z_1 and z_2 which can possibly capture the interaction between these two variables. This multiplicative variable is also added to the Cox proportional model. The most significant variable that has smaller p-value in fitted model is z_{12}. Therefore, z_{12} has been considered as one of the covariates. Cox proportional model is fitted again, this time with all variables except z_{12}, z_1 and z_2. These three variables have been removed from Cox proportional model to find out another risk factor most correlated with the survival time. The variable with the second smallest p-value is z_8 which is an indicator for French-American-British (FAB) classifications of each patient. Therefore, z_8 and z_{12} were chosen to be the two covariates to contribute in the rest of the analysis. The Cox proportional hazard model with two covariates z_{12} and z_8 is

$$h(t|Z) \sim h_0(t) \exp(group + z_{12} + z_8). \tag{6.1}$$

Among the three variables, *group* was not significant; however, this variable has been kept

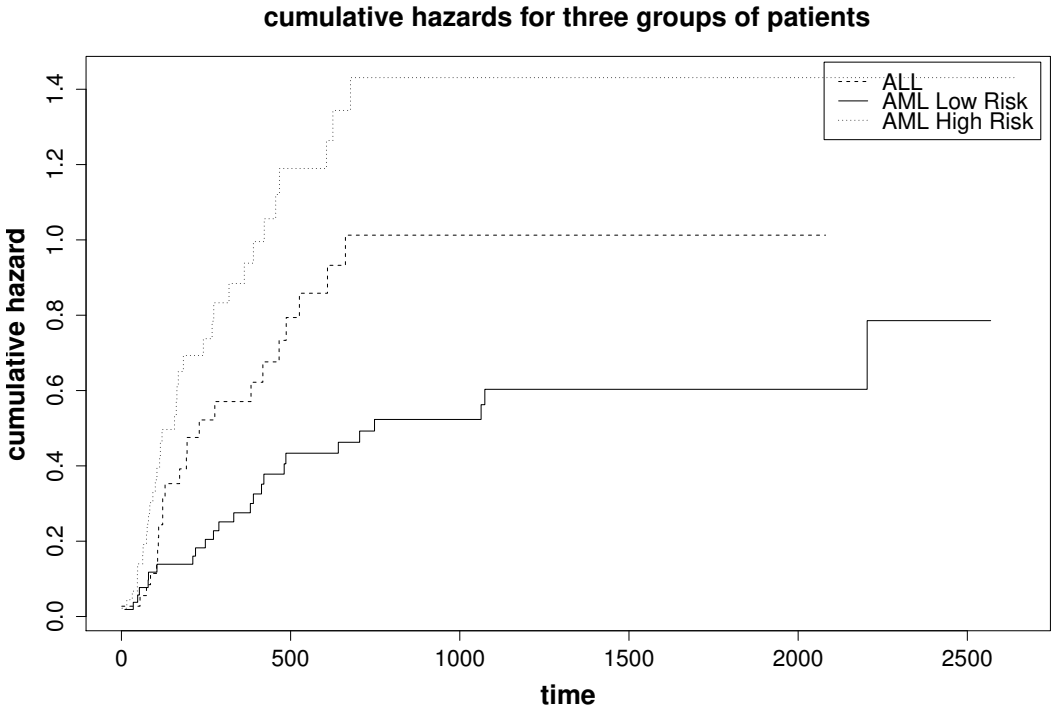

Figure 6.2 Cumulative hazard estimate for each group of patients in BMT dataset.

in the model for the future investigations.

6.3 Cure rate models

Cure rate models seem reasonable for these data, since we observe a long tail in Kaplan-Meier curve presented in Figure 6.1 for patients who are in group *ALL* and *AML high risk*. The first step in analysing cure rate models, as has been mentioned in Chapter 3, is to test the possibility of having a cured fraction in the dataset. In the next section preliminary tests are conducted for testing existence of a cure rate fraction.

6.3.1 Preliminary tests

There are two important questions that should be investigated before fitting any cure rate model on a dataset. As it has been explained comprehensively in Chapter 3, the first question is whether there is any cured fraction in the population, and the second question is whether the follow-up time is long enough.

Table 6.1 Estimating the effect of z_1, z_2, z_8 and z_{12} using Cox model.

	p-value	Confidence Interval
group	< 0.01	$(0.7, 1.4)$
z_{12}	0.0003	$(1.001, 1.005)$
z_8	0.01	$(1.1, 3.1)$

To answer the first question, $\hat{p}_n$ should be estimated. Largest observed survival time in each group of patients, $t_{(n)}$, are 2081, 2569 and 2640 days for group *AML low risk*, *ALL*, and *AML high risk* respectively. So, $\hat{p}_n$ for each group of patients is: 0.363, 0.456 and 0.24, and the number of observations in each group is 37, 54 and 46, respectively. From Maller and Zhou (1996), c_α for $\alpha = 5\%$ (or 10%) is always greater than 0.53, therefore the significance level is always below the cut-off value. Therefore, $\mathbf{H}_{01}$ of equation (3.4) is rejected and we conclude that there is a cured sub-population in all three groups of patients.

The answer to the second question has been discussed extensively in Section 3.2.2. The statistic that we used for the statistical test is q_n defined in equation (3.6). We computed q_n for each group of patients in the dataset. The results are $1, 0.018$ and 1 for group *AML low risk*, *ALL* and *AML high risk* respectively. Considering the number of patients in each group, the simulated thresholds for statistics q_n are above the 0.02 under significance level of 0.5% (or 0.10%) for group *ALL* and *AML high risk*, see Maller and Zhou (1996).
From the above discussion and with regards to discussion in Section 3.2.2, we accept $\mathbf{H}_{02}$ in equation (3.5) under significance level of 0.5% (or 0.10%) for group *ALL* and *AML high risk* of patients. Therefore, we conclude that the follow-up time is statistically long enough for group *ALL* and *AML high risk*, but not for group *AML low risk* of patients.

6.3.2 Mixture models

Different types of parametric and non-parametric mixture cure rate models were discussed in Chapter 3 and Chapter 4. In this section proportional hazard mixture model is fitted to the BMT dataset. Functions related to proportional hazard mixture model are available in packages `smcure` in `R`; see Cai et al. (2012). In addition, three of the more advanced mixture

models that have been investigated in simulation studies in Chapter 5 are applied to the BMT dataset.

In Section 3.5.1 theory and methodology of proportional hazard mixture cure rate model are discussed. Such models are fitted on the BMT dataset. Fitting proportional hazard mixture cure rate model on the dataset is easy and fast due to the availability of package `smcure` in `R`.

The variables in the model are z_{12} and z_8. The same variables have been used to model cured fraction and survival probability. The estimated results for coefficients and p-values are given in Table 6.2 and Table 6.3. Formulation of proportional hazard mixture model in package `smcure` does not include intercept for survival probability.

Table 6.2 Result of Cox proportional hazard cure rate model for *ALL* group, using package `smcure` in `R`.

	Survival parameters		Cure parameters	
Coefficient	$\text{Est}_{(\text{CV})}$	p-value	$\text{Est}_{(\text{CV})}$	p-value
Intercept	–	–	$1.20_{(1.89)}$	0.52
z_{12}	$0.27_{(0.40)}$	0.50	$0.24_{(0.36)}$	0.66
z_8	$0.36_{(1.0)}$	0.71	$-1.13_{(1.62)}$	0.48

Table 6.3 Result of Cox proportional hazard cure model for *AML high risk* group, using package `smcure` in `R`.

	Survival parameters		Cure parameters	
Coefficient	$\text{Est}_{(\text{CV})}$	p-value	$\text{Est}_{(\text{CV})}$	p-value
Intercept	–	–	$1.20_{(0.43)}$	0.01
z_{12}	$0.03_{(0.19)}$	0.88	$-0.02_{(0.42)}$	0.96
z_8	$-0.14_{(0.36)}$	0.70	$-0.17_{(0.90)}$	0.85

The null hypothesis cannot be rejected for neither group implying that z_{12} and z_8 have no effects. Given that there are only 38 patients in *ALL* group and 45 in *AML high risk*, such results are not reliable. In the next section we apply more advanced cure rate models. The

same covariates are used to see whether the selected covariates have any effects on the survival time.

The cure rate quantile regression model in 4.2.1 and semi-parametric cure model for two groups trials in 4.2.2 have been fitted on bone marrow transplantation (BMT) dataset in this section. The covariates of interest are z_{12} and z_8 which have been used in section 6.3.2. These covariates are used both for survival probability and cure rate estimations. Patients in groups *ALL* and *AML high risk* are involved in this study. We have generated 200 bootstrap samples for estimating variance and p-value.

The coefficient of variation (CV) is used instead of standard error in the following tables because standard errors for most of the parameter estimations are large, and estimations are not significant in these cases. Then we used CV to be able to compare relative standard error to parameter estimation.

Table 6.4 $\beta_0(\tau)$: Parameter estimation using cure rate quantile regression for group *ALL*, and *AML high risk* of patients.

	Model parameters *ALL*		Model parameters *AML high risk*	
	$\text{Est}_{(\text{CV})}$	p-value	$\text{Est}_{(\text{CV})}$	p-value
$\tau = 0.2$	$3.83_{(0.12)}$	0.12	$3.54_{(0.24)}$	0.33
$\tau = 0.4$	$5.23_{(0.13)}$	0.38	$3.79_{(0.1)}$	0.28
$\tau = 0.6$	$5.38_{(0.1)}$	0.2	$4.28_{(0.08)}$	0.22
γ_0	$13.37_{(0.57)}$	0.68	$20_{(0.54)}$	0.14

6.3.3 Non-mixture model

Results in Table 6.7 show that semi-parametric two groups trial method fits the data well. It seems semi-parametric two groups trials is a good model for small datasets. While β has been estimated equal to zero which implies that log ratio of two density functions for each group, according to model described in Section 4.2.2, is constant. The estimated π's are close to the failure rate in each group (failure rates are 63%, and 75% for groups *ALL* and *AML high risk* of patients respectively). Since the scope of study is small, it sounds reasonable to repeat the study with larger dataset to validate the findings.

Table 6.5 $\beta_1(\tau)$: Parameter estimation using cure rate quantile regression for group *ALL*, and *AML high risk* of patients. The SE/μ is coefficient of variance.

	Model parameters *ALL*		Model parameters *AML high risk*	
	$\text{Est}_{(\text{CV})}$	p-value	$\text{Est}_{(\text{CV})}$	p-value
$\tau = 0.2$	$-1.2_{(-0.88)}$	0.3	$-0.13_{(-113.21)}$	0.2
$\tau = 0.4$	$-0.26_{(-1.04)}$	0.4	$-0.1_{(-9.77)}$	0.2
$\tau = 0.6$	$-0.33_{(-1.05)}$	0.22	$-0.09_{(-9.9)}$	0.24
γ_1	$-0.76_{(2.82)}$	0.78	$12.5_{(-7.12)}$	0.12

Table 6.6 $\beta_2(\tau)$: Parameter estimation using cure rate quantile regression for group *ALL* and *AML high risk* of patients. The CV is coefficient of variation.

	Model parameters *ALL*		Model parameters *AML high risk*	
	$\text{Est}_{(\text{CV})}$	p-value	$\text{Est}_{(\text{CV})}$	p-value
$\tau = 0.2$	$0.47_{(9.3)}$	0.18	$1.13_{(0.6)}$	0.31
$\tau = 0.4$	$-0.61_{(9.91)}$	0.68	$0.95_{(0.51)}$	0.28
$\tau = 0.6$	$-0.72_{(6.47)}$	1	$0.5_{(0.66)}$	0.3
γ_2	$0.6_{(0.66)}$	0.65	$8.94_{(0.73)}$	0.11

Table 6.7 Results of the analysis using semi-parametric two groups mixture cure model for groups *ALL* and *AML high risk* group.

Parameter	$\text{Est}_{(\text{CV})}$	p-value
α	$0.54_{(0.95)}$	0.03
β	$0.00_{(-0.56)}$	0.04
p_1	$0.65_{(0.12)}$	0.001
p_2	$0.76_{(0.08)}$	0.001

Table 6.8 Results of the analysis using the non-mixture cure rate model for *ALL* and *AML high risk* of patients.

	Model parameters group *ALL*		Model parameters group *AML high risk*	
Parameter	$\text{Est}_{(\text{CV})}$	p-value	$\text{Est}_{(\text{CV})}$	p-value
	$0.1_{(2.26)}$	0.25	$0.28_{(0.87)}$	0.36
β_1	$-0.55_{(-0.8)}$	0.49	$-0.22_{(-1.99)}$	0.55
β_2	$0.39_{(0.72)}$	0.28	$-0.07_{(-2.61)}$	0.34
α	$-0.17_{(-1.03)}$	0.38	$-0.08_{(-1.11)}$	0.53
λ	$-5.59_{(-0.03)}$	0.00	$-5.32_{(-0.03)}$	0.00

CHAPTER 7 CONCLUSION

In this chapter, we present a synthesis of the study of cure rate models developed throughout the thesis. We also recapitulate on the simulation studies that we have conducted, and finally we conclude about what could be done as future research.

7.1 Synthesis of cure rate models

Although cure rate models have been broadly studied for decades and many applications have been reported, the scientific literature lacks studies that collect and analyse multiple research studies or papers related to the topic. The current thesis provides an overview for modelling the time-to-event data with cured fraction. We presented a summary on the early approaches as well as some of recent works on cure rate models, which allow the reader to have an overall understanding about the topic and different methods in cure rate models.

We discussed both mixture cure rate models and non-mixture cure rate models. The models explained under mixture cure rate comprised proportional hazard mixture cure rate model, quantile regression mixture cure rate model, and two groups trial mixture cure rate model. The non-mixture cure rate model consists of an application of cure rate model with bayesian approach for cancer data. All these models help to investigate different methods to adding covariates and the impacts of these covariates in fitting mixture cure rate and non-mixture cure rate models.

Each proposed model has different advantages and disadvantages. The proportional hazard mixture cure rate model keeps all the advantages of regular proportional odds model for survival analysis. For example we can still fit proportional hazard mixture cure rate model on survival data with cured fraction without assuming a parametric model for survival time of subjects. While most of the cure rate models provide mean-based assessment for analysing survival time and effects of covariates on survival time, the quantile regression mixture cure rate model provides assessment for different quantiles of survival time. Therefore, with quantile regression mixture models we can have a better understanding of trend of covariates effects on survival times. We discussed another mixture cure rate model which provides a semi-parametric model for survival data with cured fraction when we have two

groups of patients. This model assumes a semi-parametric model for the ratio of the density functions of survival times without considering a specific parametric model for each density function. The non-mixture cure rate model that we discussed is an alternative to mixture cure rate model, and provides a nice and more meaningful biological interpretation, especially for cancer dataset.

7.2 Summary of simulation results and future work

The results of simulation studies in Chapter 4 show that all the models have smaller standard error for estimation of parameters when sample size gets larger. This result implies consistency of parameter estimates in each model. The application of proposed models on bone marrow transplant dataset in Chapter 6 does not show significant estimates for parameters except for the semi-parametric mixture cure model of Section 4.2.2 designed for dataset with two groups of patients. This is mostly due to the fact that the dataset is relatively small. As future work, if larger datasets are available, they could be used to validate the findings for each model and perform appropriate comparisons of the results.

REFERENCES

J. Berkson et R. P. Gage, "Survival curve for cancer patients following treatment", *Journal of the American Statistical Association*, vol. 47, no. 259, pp. 501–515, 1952.

J. W. Boag, "Maximum likelihood estimates of the proportion of patients cured by cancer therapy", *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 11, no. 1, pp. 15–53, 1949.

C. Cai, Y. Zou, Y. Peng, et J. Zhang, "smcure: An r-package for estimating semiparametric mixture cure models", *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1255 – 1260, 2012.

M.-H. Chen, J. G. Ibrahim, et S. R. Lipsitz, "Bayesian methods for missing covariates in cure rate models", *Lifetime Data Analysis*, vol. 8, no. 2, pp. 117–146, 2002.

S. C. Cheng, L. J. Wei, et Z. Ying, "Analysis of transformation models with censored data", *Biometrika*, vol. 82, no. 4, pp. 835–845, 1995.

E. Copelan, J. Biggs, J. Thompson, P. Crilley, J. Szer, J. Klein, N. Kapoor, B. Avalos, I. Cunningham, et K. Atkinson, "Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with bucy2", *Blood*, vol. 78, no. 3, pp. 838–843, 1991.

R. D. Cox, "Regression models and life tables", *Journal of the Royal Statistic Society*, vol. B, no. 34, pp. 187–202, 1972.

A. P. Dempster, N. M. Laird, et D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

S. K. Douglas, D. Angelo, G. Muralikrishna, P. Jonathan, et N. Douglas, "Effect of donor recipient age match on survival after first deceased donor renal transplantation", *Journal of the American Society of Nephrology*, vol. 15, no. 4, pp. 1086–1091, April 2004.

V. T. Farewell, "The use of mixture models for the analysis of survival data with long-term survivors", *Biometrics*, vol. 38, no. 4, pp. 1041+, Déc. 1982.

——, "Mixture models in survival analysis: are they worth the risk?" *Canad. J. Statist.*, vol. 14, no. 3, pp. 257–262, 1986.

A. I. Goldman, "Survivorship analysis when cure is a possibility: A monte carlo study", *Statistics in Medicine*, vol. 3, no. 2, pp. 153–163, 1984.

J. Ibrahim, M. Chen, et D. Sinha, *Bayesian Survival Analysis*. Springer, 2001.

M. L. M. John P. Klein, Shih-Chang Lee, "A partially parametric estimator of survival in the presence of randomly censored data", *Biometrics*, vol. 46, no. 3, pp. 795–811, 1990.

J. Klein et M. Moeschberger, *Survival analysis: Techniques for censored and truncated regression*. Springer, 1997.

A. Y. C. KUK et C.-H. CHEN, "A mixture model combining logistic regression with proportional hazards regression", *Biometrika*, vol. 79, no. 3, pp. 531–541, 1992.

R. A. Maller et X. Zhou, *Survival analysis with long-term survivors*. Wiley, New York, 1996.

M. Mao et J.-L. Wang, "Semiparametric efficient estimation for a class of generalized proportional odds cure models", *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 302–311, 2010.

D. S. Ming-Hui Chen, Joseph G. Ibrahim, "A new Bayesian model for survival data with a surviving fraction", *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 909–919, 1999.

L. Peng et Y. Huang, "Survival analysis with quantile regression models", *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 637–649, 2008.

Y. Peng et K. B. G. Dear, "A nonparametric mixture model for cure rate estimation", *Biometrics*, vol. 56, no. 1, pp. 237–243, 2000.

J. Qin et B. Zhang, "A goodness-of-fit test for logistic regression models based on case-control data", *Biometrika*, vol. 84, no. 3, pp. 609–618, 1997.

Y. Shen, J. Qin, et J. P. Costantino, “Inference of tamoxifen’s effects on prevention of breast cancer from a randomized controlled trial”, *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1235–1244, 2007.

J. M. G. Taylor, “Semi-parametric estimation in failure time mixture models”, *Biometrics*, vol. 51, no. 3, pp. 899–907, 1995.

A. D. Tsodikov, J. G. Ibrahim, et A. Y. Yakovlev, “Estimating cure rates from survival data: An alternative to two-component mixture models”, *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 1063–1078, 2003.

Y. Wu et G. Yin, “Cure rate quantile regression for censored data with a survival fraction”, *Journal of the American Statistical Association*, vol. 108, no. 504, pp. 1517–1531, 2013.

A. Y. Yakovlev, B. Asselain, V. Bardou, A. Fourquet, T. Hoang, A. Rochefediere, et A. Tsodikov, “A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer”, *Biometrie et analyse de donnees spatio-temporelles*, vol. 12, pp. 66–82, 1993.

APPENDIX A R IMPLEMENTATION OF MIXTURE QUANTILE REGRESSION.

```
# computing H_gamma(u|W)

   H_gam<-function(dat, u, gam){
    n=nrow(dat)
    gamW=apply(matrix(rep(gam,n),nrow=n,byrow=T)*dat[,colW],1,sum)
    out=-log(1-logitf(gamW)*u)
    return(out)
     }

# computing the intergral

  int_0<-function(dat, gam){
    out=H_gam(dat,grd[1],gam)
     return(out)
    }

# computing the intergral

int_j<-function(dat, j, gam, betQS){
   n=nrow(dat)
   sum_j_0=rep(0,n)
   sum_j_1=sum_j_0
   tldZ=cbind(dat[,ncol(dat)],dat[,colZ])
   for(k in (1:j)-1){
    if(k==0){
       sum_j_0=H_gam(dat,grd[1],gam)}
if(k!=0){
    betQ=betQS[k,]
    betQZ=exp(apply(matrix(rep(betQ,n),nrow=n,byrow=T)*tldZ,1,sum))
    sum_j_1=sum_j_1+(dat[,1]>=betQZ)*(H_gam(dat,grd[k+1],gam) -
    H_gam(dat,grd[k],gam))
```

```
}
   }
   sum_j=sum_j_0+sum_j_1
   out=tldZ*matrix(rep(sum_j,p1+1),nrow=n,byrow=F)
    return(2*out)
   }

# computing l_j(h)

  l_j<-function(h,dat,j,gam,betQS){
    n=nrow(dat)
    bf_h=matrix(rep(h,n),nrow=n,byrow=T)
    tldZ=cbind(dat[,ncol(dat)],dat[,colZ])
     DltZ=tldZ*matrix(rep(dat[,2],p1+1),nrow=n,byrow=F)
     bf_hDltZ=apply(bf_h*DltZ,1,sum)
     term1=sum(abs(log(dat[,1])*dat[,2]-bf_hDltZ))
     term2=abs(R_star+sum(bf_hDltZ))
     term3=abs(R_star-sum(bf_h*int_j(dat, j, gam, betQS)))
     out=term1+term2+term3
     return(out)
   }

# use L_1fit to find minimum value of l_j(h)
# computing response and covariate in l_j(h)

  lad<-function(dat,j,gam,betQS){
   n=nrow(dat)
   tldZ=cbind(dat[,ncol(dat)],dat[,colZ])
   DltZ=tldZ*matrix(rep(dat[,2],p1+1),nrow=n,byrow=F)
   sumDltZ=-apply(DltZ,2,sum)
   tmp1=int_j(dat, j, gam, betQS)
   sumQuan=apply(tmp1,2,sum)

   Y_resp=c(log(dat[,1])*dat[,2],R_star,R_star)
   X_cov=rbind(DltZ,sumDltZ,sumQuan)
   fitrq<-rq.fit(X_cov,Y_resp, tau=0.5, method="br", interp=FALSE)
```

```
   tmp2=fitrq$coeff
   return(tmp2)
   }

# replaceNA converts vector A to a matrix with first row equal to A  and the
# rest of rows equal to NA.

replaceNA<-function(A){
   for(j in 1:ncol(A)){
   idx_nan=(is.nan(A[,j]) | is.na(A[,j]))*(1:nrow(A))
   if(sum(idx_nan)>0){
   idx1=idx_nan[idx_nan!=0]
   A[idx1,j]=rep(A[min(idx1)-1,j],length(idx1))
    }
  }
  return(A)
}

# use L_1fit to find minimum value of l_j(h)

l1fit_bet<-function(dat,gam,betQS){
    n=nrow(dat)
   for(jj in 1:lgrd){
    betQS=replaceNA(betQS)
betQS[jj,]=lad(dat,jj,gam,betQS)
   }
   return(betQS)
}

# Sn_gam_rev1 function is a function that demonstrates equation S_n.

Sn_gam_rev1<-function(dat, j, gam, betQS){
   n=nrow(dat)
   gamW=apply(matrix(rep(gam,n),nrow=n,byrow=T)*dat[,colW],1,sum)
   pigamW=logitf(gamW)
   dpigamW=pigamW*(1-pigamW)
```

```
   tmpa_0=tmpa_1=tmpb_0=tmpb_1=rep(0,n)
   dtmpa_0=dtmpa_1=dtmpb_0=dtmpb_1=rep(0,n)
   tldZ=cbind(dat[,ncol(dat)],dat[,colZ])

 for(k in 0:(j-1)){
    if(k==0){
 betQ_0=betQS[1,]
     betQZ_1=exp(apply(matrix(rep(betQ_0,n),nrow=n,byrow=T)*tldZ,1,sum))
 C0_1=1-pigamW
 aa_0=grd[1]/2
 C0_2=1-pigamW*aa_0
 C0_3=dat[,2]*(dat[,1]<betQZ_1)
 C0_4a=1-pigamW*grd[1]
 C0_4b=1-pigamW*0
 tmpa_0=C0_1*(C0_2^-1)*C0_3
 tmpb_0=C0_1*(C0_4a^-1-(C0_4b^-1))

 dC0_1=-dpigamW
 dC0_2=-dpigamW*aa_0
 dC0_4a=-dpigamW*grd[1]
 dtmpa_0=(dC0_1*C0_2-C0_1*dC0_2)*(C0_2^-2)*C0_3
     dtmpb_0=dC0_1*(C0_4a^-1-(C0_4b^-1))-C0_1*dC0_4a*(C0_4a^-2)
}
if(k!=0){
     betQ_k=betQS[k,]
     betQZ_k=exp(apply(matrix(rep(betQ_k,n),nrow=n,byrow=T)*tldZ,1,sum))
     betQ_k1=betQS[k+1,]
     betQZ_k1=exp(apply(matrix(rep(betQ_k1,n),nrow=n,byrow=T)*tldZ,1,sum))

 C1_1=1-pigamW
 aa_k=grd[k]/2+grd[k+1]/2
 C1_2=1-pigamW*aa_k
 C1_3=dat[,2]*(dat[,1]>=betQZ_k)*(dat[,1]<betQZ_k1)
 C1_4a=1-pigamW*grd[k+1]
 C1_4b=1-pigamW*grd[k]
 C1_5=(dat[,1]>=betQZ_k)*1
```

```
 a_1=C1_1*(C1_2^-1)*C1_3
 b_1=C1_1*((C1_4a^-1)-(C1_4b^-1))*C1_5
 tmpa_1=tmpa_1+a_1
 tmpb_1=tmpb_1+b_1

 dC1_1=-dpigamW
 dC1_2=-dpigamW*aa_k
 dC1_4a=-dpigamW*grd[k+1]
 dC1_4b=-dpigamW*grd[k]
 da_1=(dC1_1*C1_2-C1_1*dC1_2)*(C1_2^-2)*C1_3
 db_1=((dC1_1*C1_4a-C1_1*dC1_4a)*(C1_4a^-2)-
       (dC1_1*C1_4b-C1_1*dC1_4b)*(C1_4b^-2))*C1_5
 dtmpa_1=dtmpa_1+da_1
 dtmpb_1=dtmpb_1+db_1
}
 }

 tmp=tmpa_0+tmpa_1-(tmpb_0+tmpb_1)
 dtmp=dtmpa_0+dtmpa_1-(dtmpb_0+dtmpb_1)

 Mtmp_W=dat[,colW]*matrix(rep(tmp,p2),nrow=n,byrow=F)
 Mdtmp_W=dat[,colW]*matrix(rep(dtmp,p2),nrow=n,byrow=F)
 out=apply(Mtmp_W,2,sum)
 dout=t(dat[,colW])%*%Mdtmp_W
 sdgam=matrix(1,ncol=p2,nrow=p2)
 ans=list(out=out,dout=dout,sdgam=sdgam)
 return(ans)
}

# NREEgam is a function that find gamma for rest of computations.

NREEgam<-function(dat,gam,betQS){
 flag=1
 ans=optim(gam,Sn_gam_nleqslv,
           method="L-BFGS-B",lower=c( -20,-20,-20), upper = c( 20, 20,20))
 gams=ans$par
```

```
 return(list(gam=gams,flag=flag))
}

# The final results for gamma and Beta, is obtained with following function.

iter2<-function(num,eps,dat,gam,betQS){
 n=nrow(dat)

 for(it in 1:num){
    Flag0=Flag1=Flag2=Flag3=0
COND0=(sum(is.nan(betQS)) || sum(is.na(betQS)))
if(COND0){return(list(Flag0=Flag0,Flag1=Flag1,Flag2=Flag2))}
Flag0=1
    ans1=l1fit_bet(dat,gam,betQS)
COND1=(sum(is.nan(ans1)) || sum(is.na(ans1)))
if(COND1){return(list(Flag0=Flag0,Flag1=Flag1,Flag2=Flag2))}
    Flag1=1
ans_gam=NREEgam(dat,gam,ans1)
COND2=((!ans_gam$flag) || sum(is.nan(ans_gam$gam)))
if(COND2){return(list(Flag0=Flag0,Flag1=Flag1,Flag2=Flag2))}
Flag2=1
ans2=ans_gam$gam
    if((max(abs(ans1-betQS))<=eps) && (max(abs(ans2-gam))<=eps)){
      Flag3=1
     return(list(Ebet=ans1,Egam=ans2,it=it,Flag0=Flag0,Flag1=Flag1,
     Flag2=Flag2,Flag3=Flag3))
    }

    betQS<<-ans1
gam<<-ans2
 }
 return(list(Ebet=ans1,Egam=ans2,it=it,Flag0=Flag0,Flag1=Flag1,
 Flag2=Flag2,Flag3=Flag3))
}

# link function to get cure probability for each individual.
```

```
logitf<-function(gamW)
  {
  out<-exp(gamW)/(1+exp(gamW))
  return(out)
  }

# Function that has been used in NREEgam to be optimized.

Sn_gam_nleqslv<-function(x){
  g<-c(x[1],x[2],x[3])
  r<-Sn_gam_rev1(dat, j=lgrd, g, betQS)$out
  out<-r[1]^2+r[2]^2+r[3]^2
  return(out)
}
```

APPENDIX B R IMPLEMENTATION OF TWO GROUPS MIXTURE MODEL.

```
# p_i and q_i estimation

prob.estimate<-function(opt,psi,pla.event,tam.event,h)
  {
     hat_p<-c()
     hat_q<-c()
     a<-pla.event
     b<-tam.event
     ro<-length(b)/length(a)
     hat_p<-(1/length(a))*(1/(1+ro*exp(opt$par[1]+
                                       opt$par[2]*h+psi)))

     hat_q<-exp(opt$par[1]+opt$par[2]*h+psi)*hat_p
     return(list(Hp=hat_p,Hq=hat_q))
}

#parameter estimation

estimate<-function(h,tam.event,pla.event,ppsi=0,initial=c(0,0))
 {
    a<-tam.event
    b<-pla.event
    ro<-length(a)/length(b)

 # Computing score function for Alpha.
    f<-function(x)
  {
    length(a)-sum(1/
                    (1+(1/ro*exp(x[1]+x[2]*h))))
  }
```

```
# Computing score function for Beta.

  g<-function(x)
  {
    sum(h[(length(a)+1):(length(b)+length(a))])-
      sum(h/(1+1/(ro*exp(x[1]+x[2]*h))))

  }

  fg<-function(x)
    {
    f(x)^2+g(x)^2
    }

  myf<-fg(initial)
  opt<-optim(par=initial,fn=fg,hessian=FALSE,method="Nelder-Mead")

  return(list(opt=opt,myf.initial=myf))
}

# Function Psi.

ppsi<-function(X0,delta0,X1,delta1,y)
{
  #\hat{H_0}
  surv.obj.cen<-Surv(time=X0,event=(1-delta0))
  kaplan.cen<-survfit(surv.obj.cen~1)
  a.cen<-summary(kaplan.cen)
  step.a.cen<-stepfun(x=a.cen$time,y=c(a.cen$surv[1],a.cen$surv))

  #\hat{H_1}

  surv.obj.tcen<-Surv(time=X1,event=(1-delta1))
  kaplan.tcen<-survfit(surv.obj.tcen~1)
  b.cen<-summary(kaplan.tcen)
  step.b.cen<-stepfun(x=b.cen$time,y=c(b.cen$surv[1],b.cen$surv))
```

```
  # estimator for psi

  ppsi<-c()
  for(i in 1:length(y))
  {
    ppsi[i]<-log(step.b.cen(y[i])+0.0001)-log(step.a.cen(y[i])+0.0001)
  }
  return(ppsi)
}

# survival object for two groups of patients
surv.obj<-function(X0,delta0,X1,delta1)
{
  surv.obj.pla<-Surv(time=X0,event=delta0)
  kaplan.pla<-survfit(surv.obj.pla~1)
  surv.obj.tam<-Surv(time=X1,event=delta1)
  kaplan.tam<-survfit(surv.obj.tam~1)
  y<-c(summary(kaplan.pla)$time,summary(kaplan.tam)$time)
  return(list(su.pla=surv.obj.pla,ka.pla=kaplan.pla,su.tam=surv.obj.tam,
  ka.tam=kaplan.tam,failur.times=y))
}
```

APPENDIX C R IMPLEMENTATION OF NON-MIXTURE MODEL.

```
# simulation parameters:

# n, is number of individuals.
# p, is number of covariates.
# t.follow, is follow up time.
# a, is shape parameter.
# l, is scale parameter.
# X, is covariate matrix.
# tb, is the true coefficient functions.
# d, is the censoring indicator and probability of censoring is 0.06.
# t.theta, is true theta which has the formula theta=exp((x)^T B).
# P, is the cure rate which is different for each individual.

X<-scale(matrix(rnorm(p*n,mean=1),ncol=p))
t.theta=exp(X%*%tb)
d<-rbinom(n,size=1,prob=0.3)
P<-exp(-t.theta)
U<-runif(n,0,1)
quant<--log(U)/t.theta
tt<-qweibull(quant[-which(U<=P)], shape=a, scale=exp(-l))
t<-rep(NA,n)
t[which(U<=P)]<-t.follow
d[which(U<=P)]<-0
t[-which(U<=P)]<-tt
si<-pweibull(t,shape=a,scale=exp(l),lower.tail=FALSE)
N<-rpois(n,lambda=si*t.theta)+d

# Related functions.

# Computing Beta posterior function.

post.beta<-function(b,y,X,alpha,lambda,d)
{
```

```
  e<-10^-300
  sh<-exp(-alpha)
  si<-pweibull(y,shape=sh,scale=exp(-lambda),lower.tail=FALSE)+e
  Xb<- X%*%b
  loglike<-d*Xb-exp(Xb)*(1-si)
  return(sum(loglike))
}

post.beta.mcmc1<-function(myb1)
{
  return(post.beta(c(myb1,bb[2],bb[3]),t,X,alpha,lambda,d))
}

post.beta.mcmc2<-function(myb2)
{
  return(post.beta(c(bb[1],myb2,bb[3]),t,X,alpha,lambda,d))
}

post.beta.mcmc3<-function(myb3)
{
  return(post.beta(c(bb[1],bb[2],myb3),t,X,alpha,lambda,d))
}

# Computing  psi=(alpha,lambda) posterior.

post.psi<-function(alpha,lambda,N,b,t,X,d)
{
  e<-10^-300
  sh<-exp(-alpha)
  si<-pweibull(t,shape=sh,scale=exp(-lambda),lower.tail=FALSE)+e
  fi<-dweibull(t,shape=sh,scale=exp(-lambda))+e
  loglike<-(N-d)*log(si)+d*log(fi)
  return(sum(loglike))
}

post.psi.mcmc1<-function(psi1)
```

```
{
  return(post.psi(alpha=psi1,lambda=lambda,N,bb,t,X,d))
}

post.psi.mcmc2<-function(psi2)
{
  return(post.psi(alpha=alpha,lambda=psi2,N,bb,t,X,d))
}

prod.N<-function(X,t,tb,a,l,d)
  {
  t.theta=exp(X%*%tb)
  si<-pweibull(t,shape=0.5,scale=1,lower.tail=FALSE)
  N<-rpois(length(t),lambda=si*t.theta)+d
  return(N)
  }

# Computing gibb's sampling.

gibbs<-function(tb,a=1,l=-1.5,Lim1,Lim2,data)
{
  X<-data[,c("1","z1","w1")]
  d<-data[,"d"]
  t<-data[,"X"]

  assign("X",X,envir = .GlobalEnv)
  assign("d",d,envir = .GlobalEnv)
  assign("t",t,envir = .GlobalEnv)

  v.alpha<-c()
  v.lambda<-c()
  v.alpha[1]<-rgamma(1,shape=1,rate=1)
  v.lambda[1]<-rnorm(1,sd=1)
  beta<-matrix(ncol=length(tb),nrow=1)
  beta[1,]<-tb
  alpha<<-v.alpha[1]
```

```
  lambda<<-v.lambda[1]
  bb<<-beta[1,]
  N<<-prod.N(X,t,tb,a,l,d)

  for(i in 1:Lim2)
    {
   out_beta1<-metrop(post.beta.mcmc1,initial=bb[1],1,scale=1)
   bb[1]<<-out_beta1$batch
   out_beta2<-metrop(post.beta.mcmc2,initial=bb[2],1,scale=1)
   bb[2]<<-out_beta2$batch
   out_beta3<-metrop(post.beta.mcmc3,initial=bb[3],1,scale=1)
   bb[3]<<-out_beta3$batch
   out_beta<-c(out_beta1$batch,out_beta2$batch,out_beta3$batch)
   beta<-rbind(beta,out_beta)

   out_psi1<-metrop(post.psi.mcmc1,initial=alpha,1,scale=1)
   v.alpha[i+1]<-out_psi1$batch
   assign("alpha",as.numeric(v.alpha[i+1]),envir = .GlobalEnv)

   out_psi2<-metrop(post.psi.mcmc2,initial=lambda,1,scale=1)
   v.lambda[i+1]<-out_psi2$batch
   assign("lambda",as.numeric(v.lambda[i+1]),envir = .GlobalEnv)

    }

   mB0<-round(mean(beta[Lim1:Lim2,1]),digits=2)
   mB1<-round(mean(beta[Lim1:Lim2,2]),digits=2)
   mB2<-round(mean(beta[Lim1:Lim2,3]),digits=2)
   mlam<-round(mean(v.lambda[Lim1:Lim2]),digits=2)
   malp<-round(mean(v.alpha[Lim1:Lim2]),digits=2)

   return(list(Alpha=malp,Lambda=mlam,Beta=c(mB0,mB1,mB2)
   ,V.beta=beta,V.lambda=v.lambda,V.alpha=v.alpha))
  }

# Running the non-mixture model and Running bootstrap for estimating
```

```
# parameter variance and p-value.

result<-function(Data=data,Iter,lim1=5,lim2=10,bootnum=200,B0=c(0,0,0))
{
n<-nrow(Data$dat)
mB0<-c()
mB1<-c()
mB2<-c()
mlam<-c()
malp<-c()
mydata<-Data$dat

res<-gibbs(tb=B0,a=1,l=-1.5,Lim1=lim1,Lim2=lim2,data=mydata)

bootresult<-vector(mod="list")
for(k in 1:bootnum)
{
  newsample<-sample(n,replace=TRUE)
  dat<-mydata[newsample,]
  assign("dat",dat,envir = .GlobalEnv)
  myit<-gibbs(tb=c(0,0,0),a=1,l=-1.5,Lim1=lim1,Lim2=lim2,data=dat)
  bootresult[[k]]<-myit
}
Ebeta<-c()
Ealpha<-c()
Elambda<-c()
mbB<-c()
vbB<-c()
P_valB<-c()
ConfIntB<-c()

for(k in 1:bootnum)
{
  b<-bootresult[[k]]$Beta
  Ebeta<-rbind(b,Ebeta)
  Ealpha[k]<-bootresult[[k]]$Alpha
```

```
  Elambda[k]<-bootresult[[k]]$Lambda
}

  mbB<-apply(Ebeta,2,mean)
  vbB<-apply(Ebeta,2,var)
  matrix1<-abs(Ebeta - matrix(rep(mbB,each=bootnum),nrow=bootnum) )>
    abs( res$Beta)
  P_valB <-apply(matrix1,2,mean)
  ConfIntB <- apply(Ebeta,2,function(x) quantile(x, c(.025, 0.975)))

mbAlpha<-mean(Ealpha)
vbAlpha<-var(Ealpha)
matrix2<-abs(Ealpha - rep(mbAlpha,times=bootnum) )> abs( res$Alpha)
P_valAlpha <-mean(matrix2)
ConfIntAlpha <- quantile(Ealpha, c(.025, 0.975))

mbLambda<-mean(Elambda)
vbLambda<-var(Elambda)
matrix3<-abs(Elambda - rep(mbLambda,times=bootnum) )> abs( res$Lambda)
P_valLambda <-mean(matrix3)
ConfIntLambda <- quantile(Elambda, c(.025, 0.975))

return(list(result0=res,Bootresult=bootresult,
BBeta=list(mean=mbB,var=vbB,P=P_valB,conf=ConfIntB),
BAlpha=list(mean=mbAlpha,var=vbAlpha,P=P_valAlpha,conf=ConfIntAlpha),
 BLambda=list(mean=mbLambda,var=vbLambda,P=P_valLambda,conf=
 ConfIntLambda)))

}
```