Modelling and inference for complex survival data under interval censoring

Ma, Xiangmei

2012

Ma, X. (2012). Modelling and inference for complex survival data under interval censoring.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/50743>

<https://doi.org/10.32657/10356/50743>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**MODELLING AND INFERENCE
FOR COMPLEX SURVIVAL DATA
UNDER INTERVAL CENSORING**

MA XIANGMEI

**Division of Mathematical Sciences
School of Physical and Mathematical Sciences**

2012

MODELLING AND INFERENCE FOR COMPLEX SURVIVAL DATA UNDER INTERVAL CENSORING

MA XIANGMEI

Division of Mathematical Sciences
School of Physical and Mathematical Sciences

A thesis submitted to Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy in Mathematical Sciences

September 2012

Abstract

Interval-censored survival data commonly arise in many clinical and epidemiological studies, where the time to an event of interest cannot be observed exactly but is only known to lie between two adjacent clinical visit times. When data are collected from several clusters, such as clinical sites or geographical regions, clustered interval-censored data are then encountered. Due to advances in medical technology, some subjects are not susceptible to the event and defined as cured. It leads us to study clustered interval-censored data with the presence of a cured subgroup assumed.

For analyzing such complex survival data under interval censoring we propose a mixture cure modelling procedure in Chapter 2. This procedure integrates a logistic regression model for the proportion of cured subjects and proportional hazards (PH) model for those uncured subjects. To reflect the within-cluster correlation of clustered data, random effects are introduced in the manner of the GLMM method. We develop the REML estimation for regression parameters and variance component of random effects and propose an EM algorithm for its implementation, as well as an iterative algorithm to estimate the unknown baseline survival function. A simulation study is then conducted to evaluate the performance of the proposed method in various practical situations. To demonstrate its usefulness, we apply the proposed method to analysis of the interval-censored relapse time data from a smoking cessation study whose subjects were recruited from 51 zip code regions in the south-eastern corner of Minnesota.

Although the proposed estimation procedure in Chapter 2 is relatively easy to implement, the implementation may be cumbersome due to the complexity of mixture cure model framework and data structure. To some extent, it is necessary to adjust the cure fraction prior to the mixture cure

analysis. A score test for testing the presence of cured subjects in clustered interval-censored survival data is provided in Chapter 3. We then set up simulations to evaluate the sampling distribution and power behaviour of the score test. A bootstrap approach is further proposed, leading to more accurate significance levels and greater power in small sample situations. We illustrate applications of the test using data sets from a smoking cessation study and a retrospective study of early breast cancer patients, respectively.

To extend the existing modelling approach, we aim to introduce a general class of semiparametric mixture cure transformation models with random effects in Chapter 4, allowing for a wide range in model selection and avoiding prespecified model assumptions to interval-censored data. We develop the maximum likelihood estimation for the parameters of interest and propose a self-consistent iterative algorithm for estimating the non-parametric component. Under appropriate conditions, we investigate the model identifiability and establish asymptotic properties, including the consistency, asymptotic normality of the parameter estimators, and the consistency with a cube-root- n convergence rate of the estimator for the non-parametric component. We conduct intensive simulations to assess the performance of the estimators proposed and revisit the smoking cessation data to illustrate the methodology mentioned in this chapter.

As a concluding remark of my dissertation, major findings of current research and several open research topics will be framed in Chapter 5.

Acknowledgements

First and foremost, I would like to give my deepest thanks to my supervisor Prof. Xiang Liming for her personal guidance and encouraging engagement during my PhD study at Nanyang Technological University. I sincerely appreciate her for the continuing coaching and valuable discussion regarding my research. I have learnt numerous advanced statistics and earnest research attitude from her.

I really appreciate NTU for awarding me the Research Scholarship. I am also grateful to the faculty of the Division of Mathematical Sciences. I have benefited a lot from various courses, seminars and individual discussions. I would like to thank Prof. Wang Lilian and Prof. Oggier Elise Frdrique for kindly presenting my qualification examination. Many thanks go to Prof. Wang Desheng and Prof. Chen Xin who taught me how to be a good tutor. I wish to express my gratitude to the fellow students in the statistics group, Hong Zhaoping, Hu Yuao, Yang Yanrong, as well as my very dear friend of many years Gu Ying, for their unbounded help and invaluable friendships.

Finally, I am deeply grateful to my dear parents for their unwavering support on my studies and everything I did.

Contents

Abstract	iii
Acknowledgements	v
1 Introduction	1
1.1 Interval-censored survival data	1
1.2 Two examples	3
1.2.1 Breast cosmesis data	3
1.2.2 Smoking cessation data	4
1.3 Mixture cure model	5
1.4 Outline	7
2 Mixture cure model with random effects	9
2.1 Introduction	9
2.2 The model	11
2.3 Estimation procedure	14
2.4 Simulation study	16
2.5 Application to the smoking cessation data	19
2.6 Concluding remarks	23

2.7	Appendix	25
3	Testing for the presence of a cure fraction	29
3.1	Introduction	29
3.2	Modelling and estimation	32
3.3	Score test for the cure fraction	34
3.4	Simulation study	36
3.4.1	Sampling distribution	36
3.4.2	Power	37
3.4.3	Bootstrap approach	41
3.5	Applications	42
3.5.1	Smoking cessation data	42
3.5.2	Breast cosmesis data	45
3.6	Concluding remarks	48
3.7	Appendices	49
4	Semiparametric mixture cure transformation model with random effects	55
4.1	Introduction	55
4.2	Estimation procedure	59
4.3	Large sample properties	62
4.3.1	Identifiability	64
4.3.2	Information for θ	67
4.3.3	Asymptotic properties for the estimators proposed	69
4.4	Numerical analysis	71
4.4.1	Simulation study	71
4.4.2	Application to the smoking cessation study	73
4.5	Concluding remarks	80

4.6	Appendices	84
5	Conclusion	107
5.1	Major findings	107
5.2	Future works	109
	Bibliography	111
	List of Works	119

List of Figures

3.1	Q-Q plots of ordered score statistics against the quantiles of the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution based on 500 samples generated from the mixture cure model with random effects under the null hypothesis.	38
3.2	Estimates of the survival function for the breast cosmesis data using the nonparametric, single PH and mixture cure models.	47
4.1	Maximum log-likelihood of the transformation mixture cure model over values of parameters ρ and r in the Box-Cox transformation and the logarithm transformation, respectively.	81
4.2	Estimated survival probabilities of subjects resumed smoking with the special intervention (solid) and usual care (dotted) by the proportional hazards cure model.	83
4.3	Estimated survival probabilities for the overall population of smokers with the special intervention (solid) and usual care (dotted) by the proportional hazards cure model.	83

List of Tables

2.1	Simulation results for (i) the proposed mixture cure models with random effects and (ii) the naive model with random effects ignored, based on 500 replications with heavy censoring 60%-65%.	20
2.2	Simulation results for (i) the proposed mixture cure models with random effects and (ii) the naive model with random effects ignored, based on 500 replications with mild censoring 35%-40%.	21
2.3	Parameter estimates in mixture cure models for the smoking cessation data.	23
3.1	Empirical power of the score statistic based on 500 samples under the mixture cure model.	40
3.2	Empirical power of the bootstrap approach based on 500 bootstrap samples under the mixture cure model.	43
3.3	Summary of model fitting for the mixture cure model with random effects and single PH model with random effects to the smoking cessation data.	45

4.1	Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with known transformations $G(t) = [(1 + t)^\rho - 1]/\rho$	74
4.2	Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with known transformations $G(t) = \log(1 + rt)/r$	75
4.3	Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with estimated transformations $G(t) = [(1 + t)^\rho - 1]/\rho$	76
4.4	Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with estimated transformations $G(t) = \log(1 + rt)/r$	77
4.5	Parameter estimates for smoking relapse time data in mixture cure models with two types of transformations.	82

Chapter 1

Introduction

1.1 Interval-censored survival data

Survival analysis focuses on applying statistical techniques to make inferences of time to event. In many practical studies, the time of interest cannot be observed exactly. Naturally, censoring is present in survival data as a peculiar feature. There are various categories of censoring, such as right censoring, left censoring, and interval censoring. Right-censored survival data occurs when the time to event is lost or has not been observed by the end of the study. When the event has been experienced before the study, left-censored data appears. A general censoring type is interval censoring which happens when the event lies within a certain time interval. Interval-censored data includes left- or right-censored data as special cases with the time interval ends at zero or infinity, respectively.

Interval-censored data arise in many fields, especially in clinical trials and epidemiological studies, where periodic clinical visits of patients are commonly conducted to monitor for a non-fatal disease event, such as cancer-related cosmetic deterioration or Aids drug resistance (Lindsey

and Ryan, 1998). Analysis of interval-censored data has been attracting much research interest in the past three decades (Finkelstein and Wolfe (1985); Finkelstein (1986); Rabinowitz et al. (1995); Kim (1997); Goggins and Finkelstein (2000); Pan and Chappell (2002), and others). A nice summary for interval-censored data is given by Sun (2006). Current status data or case 1 interval-censored data are named for those survival data that consists of left- or right-censored data. Throughout this thesis, interval-censored data we are interested in is case 2 interval-censored data, in which at least one time interval ends at neither zero nor infinity.

Moreover, interval-censored survival data may be observed as repeated measurements or collected from several clusters, such as clinical sites or geographical regions. When the population is composed of heterogeneous clusters, clustered interval-censored data have been investigated by Bellamy et al. (2004) in a community-based study, Wong et al. (2005) for a multi-level analysis in dental research, and references therein. To model clustered survival data from the classical perspective, a class of survival models that allow random effects (also known as frailties) in the hazard function is required to accommodate the inherent correlation due to the clustered data structure. Generalized linear mixed models (GLMMs) have been widely recognized as powerful tools for the analysis of clustered data, because they can incorporate both within-cluster and between-cluster variations (McCulloch et al., 2008).

Due to advances in medical technology, there possibly exists some "cured" subjects who are not susceptible to the event of interest or have prolonged disease-free survival. As an effective means to account for such a portion of cured subjects in survival data analysis, the mixture cure modelling approach has received much attention in the statistical literature, see, for ex-

ample, Farewell (1982, 1986); Kuk and Chen (1992); Peng and Dear (2000); Sy and Taylor (2000), and Zhang and Peng (2007), among others. Recently, various models and estimation methods have been proposed in mixture cure model for case 1 or case 2 interval-censored data with a cure fraction including Taylor et al. (2003); Lam and Xue (2005); Cook et al. (2008), Ma (2009, 2010).

Throughout this thesis, we focus on modelling and inference of such complex survival data, which are clustered, interval-censored, and have a cured subgroup.

In this chapter, we first present two examples that motivated the research in Section 1.2. In Section 1.3, we summarize a common mixture cure model to handle survival data with a cure fraction. Finally the outline of this thesis is described in Section 1.4.

1.2 Two examples

In this section, we present two examples which are drawn from a retrospective study of early breast cancer patients and a smoking cessation study respectively.

1.2.1 Breast cosmesis data

Finkelstein and Wolfe (1985) provided an interval-censored data set from a retrospective study of early breast cancer patients conducted to compare two treatments with respect to the time to cosmetic deterioration. The study considered two groups of patients receiving radiation therapy alone (46 patients) or in combination with chemotherapy (48 patients). All patients were seen at clinic visits initially every 4 to 6 months with times and

times between visits varied from patient to patient. At the clinic visits, a physician determined the presence of breast retraction, which is regarded as cosmetic deterioration of the cancer treatment. The time to breast retraction was of interest in the study, but not directly observable. All observations were either right- or interval-censored. This data set is used in Chapter 3 as an application of the proposed tests.

1.2.2 Smoking cessation data

Interval-censored data may occur in more complicated settings with the presence of a cured subgroup assumed and having the clustered structure, such as relapse time data collected from a smoking cessation study. The event of interest in this study is whether individuals resume smoking (relapse). The relapse times are interval-censored because each subject in this study was observed once a year over the five-year follow-up, and if a subject resumed smoking after an initial attempt to quit, then only an approximate one-year time interval was observed from the previous observation to the current observation. The study includes 223 subjects, known to have quit smoking at least once during the study period, living in 51 zip code regions in the south-eastern corner of Minnesota. It was found that 158 of these subjects had not resumed smoking during the follow-up period, giving a cessation rate of around 71%. Meanwhile, these different geographical regions lead to a clustered data structure, and random cluster effects may have an impact on the risk of smoking resumption, given the possible dependence among subjects residing in the same zip code region. The study also includes former smokers who were successful in quitting smoking. They may be regarded as cured subjects, in contrast to failed quitters who may resume smoking soon after an initial attempt to quit.

These heterogeneous relapse times can therefore be modelled by a mixture cure survival model, with one component corresponding to the cured group and another to the sub-population at risk of smoking resumption. More details of the study can be found in Murray et al. (1998). These data are given in the Supplementary Materials of Banerjee and Carlin (2004). For each subject, covariate informations such as sex (0=male, 1=female), intervention type SI/UC (1=special intervention [SI], 0=usual care [UC]), number of cigarettes smoked per day, and duration of smoking habit in years are recorded. It is of interest to estimate the covariate effects on the success rate of smoking cessation as well as that on the smoking relapse time.

Chapter 2 applies the proposed method to the interval-censored smoking relapse time data for illustration purposes. Then the smoking cessation study is revisited to illustrate the application of the proposed tests in Chapter 3 for the presence of a cured subgroup in the population. At last, Chapter 4 makes a comparison among several transformed models and evaluates the suitability of the model selected by Chapter 2 for the smoking cessation data.

1.3 Mixture cure model

In many clinical trial studies, the mixture cure model provides an effective means of analysis of survival data with a cure fraction. This approach integrates the logistic regression model for the proportion of cured subjects and the proportional hazards (PH) for those uncured subjects. Throughout this thesis, we assumed that the censoring is non-informative and independent of the cure fraction and the event time for uncured subjects. Let T be a non-negative event time of interest with the d -dimensional vector of

covariates Z . To define the mixture cure model, we first define a latent binary variable Y with $Y = 1$ indicating that the subject will eventually experience a failure event (uncured). Then a marginal survival function of T for the whole population given Z , denoted by $S_{pop}(t|Z)$, satisfies the following mixture cure model:

$$\begin{aligned} S_{pop}(t|Z) &= P(T \geq t|Z) \\ &= P(T \geq t|Y = 1; Z)P(Y = 1|Z) + P(T \geq t|Y = 0; Z)P(Y = 0|Z) \\ &= 1 - \pi(Z) + \pi(Z)S(t|Z), \end{aligned} \tag{1.3.1}$$

where $\pi(Z) = P(Y = 1|Z)$ is the probability of a subject being uncured, and $S(t|Z) = P(T \geq t|Y = 1; Z)$ is a conditional survival function of the uncured subjects. The uncured probability of a subject with covariates Z can be modeled by a logistic regression (Farewell, 1982):

$$\pi(Z) = \frac{\exp(\xi)}{1 + \exp(\xi)}, \tag{1.3.2}$$

where ξ is a linear form of the covariates, $\xi = b^T W$ with $W = (1, Z^T)^T$, and b is a $(d+1)$ -dimensional vector representing the effects of the covariates on the uncured probability. We further assume that the conditional survival function $S(t|Z)$ for the uncured group takes the PH model form (Kuk and Chen, 1992):

$$S(t|Z) = S_0(t)^{\exp(\eta)}, \tag{1.3.3}$$

where $\eta = \beta^T Z$ is the linear predictor of the covariates, β is a d -dimensional vector representing the effects of the covariates on the survival model component, and $S_0(t)$ is the unknown baseline survival function of time t .

Interval-censored data have been widely investigated in PH model. The well-known Kaplan-Meier estimator, although applicable to right censored

data, cannot be applied to interval-censored data. An additional self-consistency algorithm or the nonparametric maximum likelihood estimation (NPMLE) must be employed to estimate the baseline distribution, which increases the complexity of interval-censored data analysis. More details and extensive reviews can be found in Lindsey and Ryan (1998) and Sun (2006). Recently, there has been growing interest in the development of mixture cure model for clustered interval-censored data.

1.4 Outline

The remainder of this thesis is organized into four chapters as follows. In Chapter 2, we propose a mixture cure modelling procedure for analyzing clustered survival data under interval censoring by incorporating random effects in both the logistic regression and the PH regression components. Under the GLMM framework, we develop the REML estimation for the parameters, as well as an iterative algorithm for the estimation of the baseline survival function in (1.3.3). A simulation study and a real data analysis are carried out to demonstrate the usefulness of the proposed method.

Chapter 3 focuses on the development of a score test for testing the presence of cured subjects in clustered and interval-censored survival data prior to the mixture cure analysis. Through simulations, we evaluate the sampling distribution and power behaviour of the score test, further develop a bootstrap version of the test for the level-adjustment. We illustrate applications using data sets from a smoking cessation study and a retrospective study of early breast cancer patients, respectively.

In Chapter 4, we extend the existing method to a general class of semi-parametric mixture cure transformation models with random effects. We

develop and implement the estimation procedure for estimating the parameters of interest and the unknown function involved in the proposed models by a proposed EM algorithm. Under appropriate conditions, we investigate the identifiability of the proposed models and establish the consistency, asymptotic normality and semiparametric efficiency of the estimators of parameters, and the consistency with a cube-root- n convergence rate of the nonparametric estimator for the unknown function. We conduct intensive simulations to assess the performance of the estimators proposed and revisit the smoking cessation data for illustration.

Chapter 5 concludes this thesis by major findings of current research and addresses some future works.

Chapter 2

Mixture cure model with random effects

2.1 Introduction

Mixture cure model for survival data has been defined in Section 1.3. When survival data are collected from several clusters, random effects should be introduced into the mixture cure model to accommodate the within cluster correlation. A number of researchers have recently applied the mixture cure model to such clustered survival data. Extending McGilchrist's method (McGilchrist, 1993) to analyze survival data with long-term survivors, Yau and Ng (2001) proposed a long-term survivor model that introduces random effects in both the logistic regression and the hazard regression components. This method was then modified for the finite mixture survival model in Ng et al. (2004). More recently, a further extension was made by Lai and Yau (2008), where correlated random effects in both the logistic regression model and the survival model have been considered within the GLMM setting. However, the aforementioned works are mainly applicable to survival data

with exact failure/censoring times.

Researchers have developed several methods in recent years to deal with such complex relapse time data, which are clustered, interval-censored, and have a cured subgroup. For example, Kim and Jhun (2008) established the cure rate model for multivariate interval-censored survival data based on a piecewise exponential distribution and introduced common random effects in their models for the uncured survival times and the latent proportion of cured subjects. From a Bayesian point of view, Banerjee and Carlin (2004) specified spatial correlation to fit geographically clustered survival data using a parametric cure rate model. Yu and Peng (2008) proposed a marginal approach to the mixture cure model with a parametrically specified, say Weibull, distribution for the data. These approaches rely on pre-specified parametric model assumptions. In practice, however, it is often difficult to find an appropriate parametric distribution to fit the data, which gives rise to the risk of model mis-specification, and an incorrect parametric choice for the distribution of the survival times may lead to erroneous inference.

In this chapter, we propose a new modelling approach to analyze such clustered and interval-censored survival data using mixture cure models under the GLMM framework. To accommodate the possible dependence of subjects within the same cluster, random effects are introduced respectively to the linear predictors of the logistic regression for the cure proportion and the PH model for uncured survival times. Adopting the GLMM approach, we develop the REML estimation of parameters through an EM algorithm, embedding an iterative procedure to find a self-consistent estimator of the baseline survival function for the interval-censored data.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the mixture cure model with normally distributed random effects

for the observed interval-censored data. The REML estimation procedure is then developed via an EM algorithm in Section 2.3. In Section 2.4, a simulation study is carried out to evaluate the performance of the proposed method in various practical situations. For illustration purposes, the proposed method is employed to analyze interval-censored relapse time data obtained from a smoking cessation study in Section 2.5. Finally, Section 2.6 provides concluding remarks and some technical details can be found in Appendix.

2.2 The model

Denote the observed data by $\{(L_{ij}, R_{ij}], Z_{ij}, \delta_{ij}) : j = 1, \dots, n_i; i = 1, \dots, n; \sum_{i=1}^n n_i = m\}$, where $(L_{ij}, R_{ij}]$ is the interval during which the response of subject j in cluster i occurs, Z_{ij} is the d -dimensional vector of covariates, and δ_{ij} is the following interval censoring indicator: $\delta_{ij} = 1_{[R_{ij} < \infty]}$. For the special case in which the survival time is right-(left-) censored, $R_{ij} = +\infty$ ($L_{ij} = 0$). We assume that $0 \leq L_{ij} < R_{ij} \leq \infty$ for all $j = 1, \dots, n_i; i = 1, \dots, n$. That means there is no exact observed time. Let $S(t|Z)$ denote the survival function for a subject with survival time t given covariates Z . Then, the likelihood function for interval-censored data is proportional to

$$\prod_{i=1}^n \prod_{j=1}^{n_i} \{S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})\}. \quad (2.2.1)$$

We review the mixture cure model in the same line with Section 1.3. $S_{pop}(t|Z_{ij})$ is the marginal survival function defined in (1.3.1) for the population given covariates Z_{ij} ,

$$S_{pop}(t|Z_{ij}) = 1 - \pi(Z_{ij}) + \pi(Z_{ij})S(t|Z_{ij}).$$

The uncured probability $\pi(Z_{ij})$ is modelled by a logistic regression in (1.3.2) with the linear form of the covariates ξ_{ij} ,

$$\pi(Z_{ij}) = \frac{\exp(\xi_{ij})}{1 + \exp(\xi_{ij})}.$$

The conditional survival function $S(t|Z_{ij})$ for the uncured group takes the proportional hazards model form with the linear predictor of the covariates η_{ij} ,

$$S(t|Z_{ij}) = S_0(t)^{\exp(\eta_{ij})}.$$

When data are grouped into clusters, certain within-cluster correlation is anticipated. It can be modelled explicitly with random effects attached to each observation in the manner of the GLMM method (McGilchrist, 1993). More specifically, random effects are introduced into both the logistic regression and survival function components through linear predictor expressions

$$\begin{aligned}\xi_{ij} &= b^T W_{ij} + v_i \\ \eta_{ij} &= \beta^T Z_{ij} + \omega_i\end{aligned}$$

where v_i and ω_i are the unobservable random effects of cluster i . A larger value of v_i implies that the subjects being treated in cluster i tend to have a lower cure probability, while a larger value of ω_i indicates that the subjects in this cluster are at a higher risk of failure. Let $\omega = (\omega_1, \omega_2, \dots, \omega_n)^T$, $v = (v_1, v_2, \dots, v_n)^T$. In addition, we assume that ω and v are independent and distributed as $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively.

We define a latent random variable Y_{ij} as $Y_{ij} = 1$ if the j th subject in cluster i will eventually experience a failure event (uncured) and as $Y_{ij} = 0$ if the subject will never experience such an event (cured). Then the likelihood

function for interval-censored data in the mixture cure model is given by

$$\begin{aligned}
Lk &= \prod_{i=1}^n \prod_{j=1}^{n_i} [S_{pop}(L_{ij}|Z_{ij}) - S_{pop}(R_{ij}|Z_{ij})]^{\delta_{ij}} S_{pop}(L_{ij}|Z_{ij})^{1-\delta_{ij}} \\
&= \prod_{i=1}^n \prod_{j=1}^{n_i} \{\pi(Z_{ij})[S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})]\}^{\delta_{ij}} \\
&\quad \{1 - \pi(Z_{ij}) + \pi(Z_{ij})S(L_{ij}|Z_{ij})\}^{1-\delta_{ij}}.
\end{aligned} \tag{2.2.2}$$

With the latent random variable denoted above, it is easy to see that $Y_{ij} = 1$ if $\delta_{ij} = 1$, and Y_{ij} is unknown if $\delta_{ij} = 0$. So $(1 - Y_{ij})(1 - \delta_{ij}) = 1 - Y_{ij}$. The likelihood then is reformatted as follows,

$$\begin{aligned}
Lk &= \prod_{i=1}^n \prod_{j=1}^{n_i} \{\pi(Z_{ij})[S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})]\}^{\delta_{ij}y_{ij}} \\
&\quad \{[1 - \pi(Z_{ij})]^{1-y_{ij}}[\pi(Z_{ij})S(L_{ij}|Z_{ij})]^{y_{ij}}\}^{1-\delta_{ij}} \\
&= \prod_{i=1}^n \prod_{j=1}^{n_i} [\pi(Z_{ij})]^{y_{ij}} [S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})]^{\delta_{ij}y_{ij}} \\
&\quad [1 - \pi(Z_{ij})]^{1-y_{ij}} [S(L_{ij}|Z_{ij})]^{y_{ij}(1-\delta_{ij})} \\
&= \prod_{i=1}^n \prod_{j=1}^{n_i} [\pi(Z_{ij})]^{y_{ij}} [S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})]^{\delta_{ij}y_{ij}} \\
&\quad [1 - \pi(Z_{ij})]^{1-y_{ij}} [S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})]^{y_{ij}(1-\delta_{ij})} \\
&= \prod_{i=1}^n \prod_{j=1}^{n_i} [\pi(Z_{ij})]^{y_{ij}} [1 - \pi(Z_{ij})]^{1-y_{ij}} [S(L_{ij}|Z_{ij}) - S(R_{ij}|Z_{ij})]^{y_{ij}}.
\end{aligned} \tag{2.2.3}$$

The third equation is derived from the fact that if $\delta_{ij} = 0$ then $S(R_{ij}|Z_{ij}) = S(\infty|Z_{ij}) = 0$. Thus, the log-likelihood of (2.2.3) can be written as

$$\begin{aligned}
\log Lk &= \sum_{i=1}^n \sum_{j=1}^{n_i} \{y_{ij} \log \pi(Z_{ij}) + (1 - y_{ij}) \log [1 - \pi(Z_{ij})] \\
&\quad + y_{ij} \log [S_0(L_{ij})^{\exp(\eta_{ij})} - S_0(R_{ij})^{\exp(\eta_{ij})}]\}
\end{aligned} \tag{2.2.4}$$

in terms of the regression parameters β, b and the baseline survival function $S_0(t)$ with $S_0(0) = 1$ and $S_0(\infty) = 0$.

2.3 Estimation procedure

To handle interval-censored data using model (1.3.3), we note that the log-likelihood function (2.2.4) depends on $S_0(t)$ only through its values at the different observation time points. Hence, we only need to focus on estimating the values of $S_0(t)$ at these time points. A self-consistent estimator of baseline survival $S_0(t)$ can be found through an iterative procedure known as the self-consistency algorithm, which was initially proposed by Turnbull (1976). Let $0 = t_0 < t_1 < \cdots < t_Q = \infty$ denote the ordered distinct time points of all observed interval end points $\{L_{ij}, R_{ij}; j = 1, \dots, n_i, i = 1, \dots, n\}$, and define $\alpha_{ijq} = 1_{[(t_{q-1}, t_q] \subset (L_{ij}, R_{ij}]}$. The weight α_{ijq} indicates whether the event that the j th subject in cluster i experiences within the interval $(L_{ij}, R_{ij}]$ could have occurred at time t_q , $q = 1, \dots, Q$, $j = 1, \dots, n_i$, $i = 1, \dots, n$. To avoid range restrictions on the parameters for the survival function, we denote $\gamma_q = \log[\log S_0(t_{q-1}) - \log S_0(t_q)]$, representing a logarithmic increase of the logarithm of the survival function within time interval $(t_{q-1}, t_q]$. The baseline survival at each time t_q is thus parameterized via γ_q

$$S_0(t_q) = \prod_{k=1}^q e^{-\exp(\gamma_k)} = \exp\left\{-\sum_{k=1}^q \exp(\gamma_k)\right\}, \quad q = 1, \dots, Q,$$

where $S_0(t_0) = 1$ when $\gamma_0 = -\infty$, and $S_0(t_Q) = 0$ when $\gamma_Q = \infty$. In terms of the parameters β, b and $\gamma = (\gamma_1, \dots, \gamma_{Q-1})^T$, the log-likelihood function (2.2.4) can then be rewritten as follows for given random effects ω and v .

$$\begin{aligned} \log Lk = & \sum_{i=1}^n \sum_{j=1}^{n_i} \{y_{ij} \log \pi_{ij} + (1 - y_{ij}) \log(1 - \pi_{ij}) \\ & + y_{ij} \log \sum_{q=1}^Q \alpha_{ijq} [e^{-\sum_{k=1}^{q-1} \exp(\gamma_k + \eta_{ij})} - e^{-\sum_{k=1}^q \exp(\gamma_k + \eta_{ij})}]\} \end{aligned} \quad (2.3.1)$$

where $\pi_{ij} = \pi(Z_{ij})$.

Following the GLMM method (McGilchrist, 1993), the best linear unbiased prediction (BLUP) type log-likelihood under the proposed model is given by $l = l_1 + l_2$, where l_1 is the log-likelihood given in (2.3.1) and

$$l_2 = -\frac{1}{2}\{n \log(2\pi\sigma_1^2) + \frac{1}{\sigma_1^2}\omega^T\omega\} - \frac{1}{2}\{n \log(2\pi\sigma_2^2) + \frac{1}{\sigma_2^2}v^T v\}.$$

Here l can be viewed as a penalized log-likelihood function with l_2 being the penalty for the conditional log-likelihood l_1 when the random effects are conditionally fixed.

Consequently, the estimators of parameters $\beta, b, \gamma, \omega, v$ are obtained by maximizing the BLUP-type loglikelihood l first and then extending them to find the REML estimators of $\beta, b, \gamma, \omega, v, \sigma_1^2$, and σ_2^2 . However, it is often challenging to maximize l directly, as the latent variable y_{ij} is included in (2.3.1). Instead, the EM algorithm is employed to achieve the estimation. In fact, the estimation procedure can be interpreted as an incomplete data problem. We first consider the complete-data log-likelihood which is constructed as $l_C = l_\beta + l_b$, where

$$\begin{aligned} l_\beta &= \sum_{i=1}^n \sum_{j=1}^{n_i} \left(E(y_{ij}) \log \sum_{q=1}^Q \alpha_{ijq} [e^{-\sum_{k=1}^{q-1} \exp(\gamma_k + \eta_{ij})} - e^{-\sum_{k=1}^q \exp(\gamma_k + \eta_{ij})}] \right) \\ &\quad - \frac{1}{2}\{n \log(2\pi\sigma_1^2) + \frac{1}{\sigma_1^2}\omega^T\omega\}, \\ l_b &= \sum_{i=1}^n \sum_{j=1}^{n_i} \{E(y_{ij}) \log \pi_{ij} + (1 - E(y_{ij})) \log(1 - \pi_{ij})\} \\ &\quad - \frac{1}{2}\{n \log(2\pi\sigma_2^2) + \frac{1}{\sigma_2^2}v^T v\}. \end{aligned} \tag{2.3.2}$$

In the E-step, with given values of $\beta, b, \gamma, \omega, v, \sigma_1^2$ and σ_2^2 in the $(r-1)$ th iteration, the conditional expectation $E(y_{ij})$ for the j th subject from cluster

i in the r th iteration is computed by

$$\delta_{ij} + \frac{(1 - \delta_{ij})\pi_{ij} \left(\sum_{q=1}^Q \alpha_{ijq} [e^{-\sum_{k=1}^{q-1} \exp(\gamma_k + \eta_{ij})} - e^{-\sum_{k=1}^q \exp(\gamma_k + \eta_{ij})}] \right)}{1 - \pi_{ij} + \pi_{ij} \left(\sum_{q=1}^Q \alpha_{ijq} [e^{-\sum_{k=1}^{q-1} \exp(\gamma_k + \eta_{ij})} - e^{-\sum_{k=1}^q \exp(\gamma_k + \eta_{ij})}] \right)}$$

With a fixed $E(y_{ij})$ in the r th iteration, the M-step of the EM algorithm is to maximize the expected complete-data log-likelihood l_C with respect to the parameters. Particularly, under the GLMM framework, updated estimators $(\hat{\beta}^{(r)}, \hat{\omega}^{(r)}, \hat{\gamma}^{(r)})$ and $(\hat{b}^{(r)}, \hat{v}^{(r)})$ in the r th iteration can be obtained by maximizing l_β and l_b separately using the Newton-Raphson method. It is equivalent to solving estimating equations (A1) and (A2) in the Appendix with initial values of $(\beta, b, \omega, v, \gamma)$ replaced by their updated estimates in each iteration for given values of σ_1^2 and σ_2^2 . When estimates $(\beta, b, \omega, v, \gamma)$ are obtained, the REML estimates of σ_1^2 and σ_2^2 are determined by equation (A3) based on the updated values of $\hat{\omega}^{(r)}, \hat{v}^{(r)}$ and the inverse of the observed information matrices corresponding to l_β and l_b . Given these updated values of σ_1^2 and σ_2^2 , the Newton-Raphson iteration starts again until the estimation procedure for all parameters is convergent. Finally, the asymptotic variances of estimators $\hat{\beta}, \hat{b}, \hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$ can be obtained from equations (A4). Further details of the M-step are given in the Appendix.

2.4 Simulation study

We carry out a simulation study in this section to examine the performance of the proposed method. The interval-censored survival times (T_L, T_R, δ) with the cure fraction are generated within each cluster in a manner similar to that employed by Pan (2000) and Kim and Jhun (2008).

First, we generate binary latent variable Y_{ij} under a logistic regression $P(Y_{ij} = 1) = 1/[1 + \exp\{-(b_1 + b_2 Z_{ij} + v_i)\}]$ for the j th subject in the i th

cluster, $j = 1, \dots, n_i, i = 1, \dots, n$, where covariates Z_{ij} are generated as m independent *Bernoulli*(0.5) random variables, i.e., Z_{ij} is equal to zero or one with probability 0.5. Interval censored data $(T_{Lij}, T_{Rij}, \delta_{ij})$ are then generated as follows.

Step 1. If $Y_{ij} = 0$, then we sample $T_{ij} = T_{Lij}$ from the exponential distribution with mean 10 and censoring indicator $\delta_{ij} = 0$.

Step 2. If $Y_{ij} = 1$, then we generate T_{ij} under the conditional hazard function $\exp(\beta Z_{ij} + \omega_i)$ and censoring times C_{ij} from the exponential distribution with mean 10. $\delta_{ij} = 1$ if $T_{ij} \leq C_{ij}$ and $\delta_{ij} = 0$ otherwise.

Step 3. For $\delta_{ij} = 0$, let $T_{Lij} < T_{Rij} = \infty$.

Step 4. For $\delta_{ij} = 1$, we create len_{ij} from distribution $U(0.2, 0.7)$ and l_{ij} from $U(0, 1)$. Then, from $(0, l_{ij}], (l_{ij}, l_{ij} + len_{ij}], \dots, (l_{ij} + k * len_{ij}, \infty)$, $k = 1, 2, \dots$, we choose $(T_{Lij}, T_{Rij}]$ as that satisfying $T_{Lij} < T_{ij} \leq T_{Rij}$.

For subjects in cluster i , random effects u_i and v_i are generated from normal distributions $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively, where standard deviation $\sigma_1 = \sigma_2 \in \{0, 0.3, 0.7, 1\}$. We set $n=10$ clusters with $n_i = 10$ subjects per cluster. Referring to Tables 2.1 and 2.2, different values of β , b_1 and b_2 are used to yield two different levels of the censoring proportion, mild censoring 35%-40% and heavy censoring 60%-65%.

We simulate 500 repeated samples for each combination of parameter configurations. For comparison purposes, each simulated sample is fitted separately with the proposed mixture cure model using the REML estimation procedure given in Section 2.3 and a naive mixture cure model (ignor-

ing random effects) using the maximum likelihood (ML) estimation. The standard error of estimators $\hat{\beta}$, $\hat{b} = (\hat{b}_1, \hat{b})$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$ in each simulated sample are obtained by taking square root of their asymptotic variances from equations (A4) in Appendix. To evaluate the performance of the parameter estimates, the average bias (Bias), average standard errors (SE), root mean square errors (RMSE) and the convergence probability (C.P.) are calculated and summarized in Table 2.1 and Table 2.2 with two different censoring levels respectively.

When $\sigma_1, \sigma_2 \neq 0$, that is, heterogeneity exists among clusters, the proposed method yields estimated coefficients closer to the true values than those obtained under the naive model for almost all simulation settings. This indicates that the proposed approach effectively corrects the bias of the estimated coefficients by incorporating random effects in both the logistic regression and the PH models. The naive model with the presence of cluster variation ignored, in contrast, results in considerably biased estimates. The good performance of the proposed method is even more apparent as the level of censoring and/or the variances of the random effects increase.

When $\sigma_1 = \sigma_2 = 0$, i.e. there is actually no random cluster effect in the data, the proposed method provides regression coefficient estimates very close to those obtained from the naive model in terms of the bias, SE, and RMSE. This shows the stability of the proposed method even if the presence of random effects has been wrongly assumed in the model.

Regarding the effect of censoring, we observe that both the bias and the standard errors of most parameter estimates tend to be smaller when the censoring proportion decreases.

The simulation results indicate that the proposed estimation method can obtain convergence in most cases. Out of the 500 simulated data sets (C.P.

ranges from 93.8% to 100%), only around 5%-6% displayed divergence, and they were mainly found in settings in which the random effect variance was large.

2.5 Application to the smoking cessation data

We now apply the proposed method to the interval-censored smoking relapse time data. The event of interest in this smoking cessation study is whether subjects resume smoking (relapse). Four covariates are considered in the study such as sex (0=male, 1=female), intervention type SI/UC (1=special intervention [SI], 0=usual care [UC]), number of cigarettes smoked per day (Cigarettes), and duration of smoking habit in years (Duration). All recorded covariates are considered in both the logistic regression and the survival model components of the mixture cure model. The REML estimates for the parameters obtained from the proposed model with random effects and the ML estimates with the model ignoring random effects are presented in Table 2.3. The estimated regression coefficients in both cases appear to have a similar inference conclusion at the 5% level of significance.

In the logistic regression used to assess the risk of smoking resumption, it is found that the following covariates are significant at the 0.05 level: the number of cigarettes smoked per day and duration of the smoking habit. In other words, subjects with a higher level of cigarette consumption have a significantly higher likelihood of resuming smoking, and subjects who have smoked for a longer duration have a significantly higher cure rate (are more likely to quit smoking). The estimated coefficient (-0.5154) of SI/UC indicates that subjects with special intervention are more likely to quit smoking than those with usual care. However, the effect of special

Table 2.1: *Simulation results for (i) the proposed mixture cure models with random effects and (ii) the naive model with random effects ignored, based on 500 replications with heavy censoring 60%-65%.*

		Model with random effects			Model ignoring random effects		
True		Bias	SE	RMSE	Bias	SE	RMSE
C.P. = 94.7%					C.P. = 100.0%		
β	0.3	-0.0436	0.5246	0.5268	-0.0824	0.5086	0.5156
b_1	-0.5	-0.0231	0.6784	0.6661	0.1015	0.5562	0.5792
b_2	0.3	0.0210	0.7887	0.7760	-0.0312	0.6779	0.7298
σ_1	1.0	-0.3022	0.3668	0.4753			
σ_2	1.0	-0.2142	0.4278	0.4787			
C.P. = 98.0%					C.P. = 100.0%		
β	0.3	0.0661	0.5379	0.5429	0.0290	0.4939	0.4957
b_1	-0.5	-0.0202	0.6456	0.6464	-0.0536	0.5887	0.6021
b_2	0.3	-0.0015	0.7386	0.7276	-0.0189	0.6806	0.7086
σ_1	0.7	-0.1717	0.2897	0.3368			
σ_2	0.7	-0.1457	0.3320	0.3627			
C.P. = 100.0%					C.P. = 100.0%		
β	0.3	0.0722	0.5346	0.5403	0.0526	0.4979	0.5015
b_1	-0.5	-0.0085	0.6246	0.6261	0.0232	0.5984	0.6045
b_2	0.3	-0.0021	0.7374	0.7325	-0.0090	0.7109	0.7235
σ_1	0.3	0.0347	0.1532	0.1571			
σ_2	0.3	0.0146	0.1933	0.1941			
C.P. = 100.0%					C.P. = 100.0%		
β	0.3	0.0960	0.5070	0.5161	0.0923	0.5034	0.5121
b_1	-0.5	-0.0054	0.6315	0.6204	-0.0042	0.6308	0.6197
b_2	0.3	0.0166	0.7363	0.7336	0.0167	0.7357	0.7338
σ_1	0	0.0945	0.0020	0.0945			
σ_2	0	0.0980	0.0077	0.0983			

Table 2.2: *Simulation results for (i) the proposed mixture cure models with random effects and (ii) the naive model with random effects ignored, based on 500 replications with mild censoring 35%-40%.*

		Model with random effects			Model ignoring random effects		
	True	Bias	SE	RMSE	Bias	SE	RMSE
		C.P. = 93.8%			C.P. = 100.0%		
β	0.2	0.0122	0.3556	0.3559	−0.0322	0.3377	0.3397
b_1	0.7	0.1020	0.5575	0.5415	0.1049	0.5920	0.5630
b_2	0.2	−0.0357	0.6959	0.6521	−0.0571	0.7111	0.6667
σ_1	1.0	−0.1813	0.2926	0.3441			
σ_2	1.0	−0.2505	0.3996	0.4717			
		C.P. = 98.4%			C.P. = 100.0%		
β	0.2	0.0260	0.3616	0.3628	0.0037	0.3529	0.3530
b_1	0.7	0.0332	0.4798	0.4745	0.0470	0.4941	0.4816
b_2	0.2	−0.0090	0.6535	0.6400	−0.0233	0.6634	0.6432
σ_1	0.7	−0.1024	0.2350	0.2564			
σ_2	0.7	−0.1606	0.3411	0.3772			
		C.P. = 100.0%			C.P. =100.0%		
β	0.2	0.0332	0.3331	0.3349	0.0291	0.3182	0.3196
b_1	0.7	0.0023	0.4409	0.4498	0.0187	0.4488	0.4557
b_2	0.2	0.0160	0.6047	0.6036	0.0119	0.6079	0.6000
σ_1	0.3	0.0565	0.1460	0.1565			
σ_2	0.3	0.0078	0.1858	0.1863			
		C.P. = 100.0%			C.P. = 100.0%		
β	0.2	0.0280	0.3217	0.3230	0.0265	0.3191	0.3202
b_1	0.7	0.0178	0.4467	0.4498	0.0189	0.4476	0.4508
b_2	0.2	0.0067	0.6006	0.5937	0.0068	0.6010	0.5937
σ_1	0	0.0963	0.0030	0.0963			
σ_2	0	0.0980	0.0049	0.0982			

intervention was not significant at the 0.05 level. This result coincides with Banerjee and Carlin (2004)'s findings. The standard deviation of random effects in this component is estimated to be 0.116, which appears to be not very apparent.

In the survival model for relapse time, except for covariate SI/UC, all of the covariates have significant effects on the hazard of the relapse time. The intervention type SI/UC is also not significant in the survival regression component, which coincides with the results of Banerjee and Carlin (2004) and Yu and Peng (2008). With estimated coefficient -0.1741 , special intervention decreases the risk of a smoking relapse, but not to the extent of statistical significance. In addition, our estimation results show that female smokers and smokers with a lengthier smoking habit are at a significantly greater risk of relapse, and the number of cigarettes smoked per day has a significant negative effect on the risk of relapse. We note that the effects of smoking duration and the number of cigarettes smoked per day differ from those found by Banerjee and Carlin (2004). This discrepancy may be due to the imposition of a parametric Weibull survival function in their study, whereas the proposed model with PH component does not rely on the parametric assumption.

The estimated standard deviation $\hat{\sigma}_1$ of random effects is 0.258, which indicates that there is considerable heterogeneity among the clusters. Compared with $\hat{\sigma}_2$, the random cluster effects in the survival model are quite different from those in the logistic model. Thus, the assumption that common random effects are shared by both the survival and the logistic models, as given in Kim and Jhun (2008), may not be appropriate in this example, which may explain why the parameter estimates and covariate effects obtained in Kim and Jhun (2008) are different from our results and those

Table 2.3: *Parameter estimates in mixture cure models for the smoking cessation data.*

Parameter	Model with random effects		Model ignoring random effects	
	Survival	Logistic	Survival	Logistic
	model	regression	model	regression
	$Estimate_{(SE)}$	$Estimate_{(SE)}$	$Estimate_{(SE)}$	$Estimate_{(SE)}$
Intercept		1.6817 _(0.8156)		1.0573 _(0.7811)
Sex	0.5329* _(0.2713)	0.0921 _(0.3286)	0.4206 _(0.2713)	0.2463 _(0.3232)
SI/UC	-0.1741 _(0.2926)	-0.5154 _(0.3729)	-0.0630 _(0.2889)	-0.6071 _(0.3630)
Cigarettes	-0.0583* _(0.0137)	0.0904* _(0.0169)	-0.0654* _(0.0139)	0.0836* _(0.0162)
Duration	0.0762* _(0.0151)	-0.1281* _(0.0245)	0.0721* _(0.0151)	-0.1072* _(0.0233)
σ_1	0.2581 _(0.3292)			
σ_2		0.1162 _(0.4081)		

* p -value < 0.05.

provided in Banerjee and Carlin (2004) and Yu and Peng (2008).

2.6 Concluding remarks

Analyzing clustered and interval-censored survival time data with a cure fraction is a challenging problem in survival analysis. The existing work to tackle this problem includes the use of Bayesian or marginal approaches to the mixture cure model. However, these approaches require a parametric distribution assumption for the survival function of the uncured subjects. Unlike the marginal method, the GLMM methodology provides estimation of and inferences about the variance component parameters in the models,

thus allowing it to reveal more information about the heterogeneity between clusters and the inherent associations within clusters. Analysis of clustered interval-censored data with a cure fraction has not yet been considered under the GLMM framework in the literature. In this chapter, we propose a flexible class of PH-based cure models in the GLMM setting. Our approach requires no specification of the parametric survival function form, and thus eliminates the potential problems due to the misspecification of the survival distribution. Moreover, random effects with a particular variance structure can be explicitly modelled via the linear predictor in both the logistic regression and the PH models. The implementation of the proposed method is, in principle, straightforward via an EM algorithm, integrating the REML estimation with a self-consistency algorithm to estimate the survival function for interval-censored data in the M-step. The simulation study shows that the proposed method provides reasonable estimation results, and the smoking cessation data analysis demonstrates its applicability.

Unlike the usual EM algorithm employed to obtain ML estimates in the GLMMs, where random effects are generally regarded as missing values and handled in the E-step, the proposed REML procedure estimates random effects ω and ν in the M-step. Unbalanced clusters and/or large cluster sizes have little effect on the estimation procedure. However, if the number of clusters or the number of distinct interval end points grows much larger, then the number of unknown parameters increases. In this case, computing the Hessian matrices used in the Newton-Raphson iteration in the Appendix would be very time-consuming and requires a high level of computing capacity. For large-scale numerical problems with many parameters, a quasi-Newton technique with the BFGS update would be far more efficient.

In addition, it is noted from the simulation study that the proposed method reduces the bias of the estimates to some extent, but tends slightly overestimate the standard error and hence the RMSE, especially when the censoring proportion is large. This weakness might be due to the increase in number of parameters to be estimated in the proposed model, when compared with the model ignoring random effects. In our real data example, standard errors of the estimates in the proposed model are similar to those obtained from the model with random effects ignored, which enlightens us that the weakness might be relieved with an increasing number of clusters. However, a large number of clusters may involve considerable computation burden and computer capability requirements as aforementioned. This dilemma is often encountered in cluster/longitudinal data analysis using the GLMM approach. For the cases with a small number of clusters, standard errors obtained from the bootstrap method may provide a way to resolve the overestimation problem for the standard errors (Jiang and Lahiri, 2006).

2.7 Appendix

The REML estimation used in the M-step of the EM algorithm can be obtained as follows.

The linear predictors can be written in matrix form as $\eta = \mathbf{Z}\beta + \mathbf{I}\omega$ and $\xi = \mathbf{W}b + \mathbf{I}v$, where $\mathbf{Z} = (Z_{11}, \dots, Z_{n,m_n})^T$ and $\mathbf{W} = (W_{11}, \dots, W_{n,m_n})^T$, $W_{ij} = (1, Z_{ij}^T)^T$, are $m \times d$ and $m \times (d + 1)$ design matrices of β and b , respectively; $\mathbf{I} = (e_1, \dots, e_n)$ is an $m \times n$ matrix with column vector e_i having elements equal to 1 for subjects from cluster i , and 0 otherwise, $i = 1, \dots, n$. For a given value of $E(y)$, the Newton-Raphson iteration procedure is performed to maximize the log-likelihood l_C with respect to

$\beta, \omega, \gamma, b, v$.

$$\begin{bmatrix} \hat{\beta} \\ \hat{\omega} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \omega_0 \\ \gamma_0 \end{bmatrix} + B_\beta^{-1} \left\{ \begin{bmatrix} \mathbf{Z}^T & 0 \\ \mathbf{I}^T & 0 \\ 0 & \mathbf{1} \end{bmatrix} \begin{bmatrix} \partial l_\beta / \partial \eta \\ \partial l_\beta / \partial \gamma \end{bmatrix} - \begin{bmatrix} 0 \\ \omega_0 / \sigma_1^2 \\ 0 \end{bmatrix} \right\} \quad (\text{A1})$$

$$\begin{bmatrix} \hat{b} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} b_0 \\ v_0 \end{bmatrix} + B_b^{-1} \left\{ \begin{bmatrix} \mathbf{W}^T \\ \mathbf{I}^T \end{bmatrix} \begin{bmatrix} \partial l_b / \partial \xi \end{bmatrix} - \begin{bmatrix} 0 \\ v_0 / \sigma_2^2 \end{bmatrix} \right\} \quad (\text{A2})$$

where $\mathbf{1}$ is a vector with all elements be 1, $\beta_0, b_0, \omega_0, v_0, \gamma_0$ are initial values of $\beta, b, \omega, v, \gamma$, respectively, and are replaced by their updated estimates in each iteration for given values of σ_1^2, σ_2^2 . Matrices B_β and B_b are the negative second derivatives of l_β and l_b with respect to (β, ω, γ) and (b, v) , respectively.

$$B_\beta = \begin{bmatrix} \mathbf{Z}^T & 0 \\ \mathbf{I}^T & 0 \\ 0 & \mathbf{1} \end{bmatrix} \begin{bmatrix} -\frac{\partial^2 l_\beta}{\partial \eta \partial \eta^T} & -\frac{\partial^2 l_\beta}{\partial \eta \partial \gamma^T} \\ -\frac{\partial^2 l_\beta}{\partial \gamma \partial \eta^T} & -\frac{\partial^2 l_\beta}{\partial \gamma \partial \gamma^T} \end{bmatrix} \begin{bmatrix} \mathbf{Z} & \mathbf{I} & 0 \\ 0 & 0 & \mathbf{1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \omega_0 / \sigma_1^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$B_b = \begin{bmatrix} \mathbf{W}^T \\ \mathbf{I}^T \end{bmatrix} \begin{bmatrix} -\frac{\partial^2 l_b}{\partial \xi \partial \xi^T} \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \mathbf{I} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & v_0 / \sigma_2^2 \end{bmatrix}$$

Similar to the arguments in Sun (2006), we denote

$$\begin{aligned}
S(t_q|Z_{ij}; \omega_i) &= \exp\left\{-\sum_{k=1}^q \exp(\mu_{ijk})\right\}, \text{ where } \mu_{ijk} = \exp(\gamma_k + \eta_{ij}) \\
g_{ij} &= \sum_{q=1}^Q \alpha_{ijq} \{S(t_{q-1}|Z_{ij}; \omega_i) - S(t_q|Z_{ij}; \omega_i)\} \\
c_{ijq} &= \sum_{l=q}^Q (\alpha_{ijl} - \alpha_{ij(l+1)}) S(t_l|Z_{ij}; \omega_i) \\
d_{ijq} &= \sum_{l=q}^Q (\alpha_{ijl} - \alpha_{ij(l+1)}) S(t_l|Z_{ij}; \omega_i) \log S(t_l|Z_{ij}; \omega_i) \\
f_{ijq} &= S(t_q|Z_{ij}) \log S(t_q|Z_{ij}; \omega_i), \quad f_{ij0} = f_{ijQ} = 0 \\
h_{ijq} &= f_{ijq} \log S(t_q|Z_{ij}; \omega_i) + f_{ijq}, \quad h_{ij0} = h_{ijQ} = 0.
\end{aligned}$$

Then, the derivatives $\frac{\partial l_\beta}{\partial \eta}$, $\frac{\partial l_b}{\partial \xi}$, $\frac{\partial l_\beta}{\partial \gamma}$, $\frac{\partial^2 l_\beta}{\partial \eta \partial \eta^T}$, $\frac{\partial^2 l_b}{\partial \xi \partial \xi^T}$, $\frac{\partial^2 l_\beta}{\partial \gamma \partial \gamma^T}$, and $\frac{\partial^2 l_\beta}{\partial \eta \partial \gamma^T}$ have components

$$\begin{aligned}
\frac{\partial l_\beta}{\partial \eta_{ij}} &= y_{ij}/g_{ij} \sum_{q=1}^Q \alpha_{ijq} (f_{ij(q-1)} - f_{ijq}) \\
\frac{\partial l_b}{\partial \xi_{ij}} &= y_{ij} - \pi_{ij} \\
\frac{\partial l_\beta}{\partial \gamma_q} &= \sum_{i,j} y_{ij} \mu_{ijq} c_{ijq} / g_{ij} \\
-\frac{\partial^2 l_\beta}{\partial \eta_{ij} \partial \eta_{ij}} &= y_{ij} \left[\left\{ \frac{\sum_{q=1}^Q \alpha_{ijq} (f_{ij(q-1)} - f_{ijq})}{g_{ij}} \right\}^2 - \frac{\sum_{q=1}^Q \alpha_{ijq} (h_{ij(q-1)} - h_{ijq})}{g_{ij}} \right] \\
-\frac{\partial^2 l_b}{\partial \xi_{ij} \partial \xi_{ij}} &= \pi_{ij} (1 - \pi_{ij}) \\
-\frac{\partial^2 l_\beta}{\partial \gamma_q^2} &= \sum_{i=1}^n \sum_{j=1}^{n_i} y_{ij} \mu_{ijq} c_{ijq} / g_{ij} \left\{ \frac{\mu_{ijq} c_{ijq}}{g_{ij}} - (1 - \mu_{ijq}) \right\}
\end{aligned}$$

$$\begin{aligned}
-\frac{\partial^2 l_\beta}{\partial \gamma_q \gamma_k} &= \sum_{i=1}^n \sum_{j=1}^{n_i} y_{ij} \left(\frac{\mu_{ijq} \mu_{ijk} c_{ijq} c_{ijk}}{g_{ij}^2} + \frac{\mu_{ijq} \mu_{ijk} c_{ijk}}{g_{ij}} \right), \quad \text{for } q < k \\
-\frac{\partial^2 l_\beta}{\partial \gamma_q \eta_{ij}} &= y_{ij} \mu_{ijq} \left\{ \frac{c_{ijq}}{g_{ij}^2} \sum_{l=1}^Q \alpha_{ijl} (f_{ij(l-1)} - f_{ijl}) - \frac{c_{ijq} + d_{ijq}}{g_{ij}} \right\} \\
-\partial^2 l_\beta / \partial \eta_{ij} \partial \eta_{kl} &= 0, \quad -\partial^2 l_\beta / \partial \xi_{ij} \partial \xi_{kl} = 0, \quad \text{for } (i, j) \neq (k, l).
\end{aligned}$$

Suppose that B_β is partitioned conformally to $\beta|\omega|\gamma$ as

$$B_\beta^{-1} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

and B_b is partitioned conformally to $b|v$ as

$$B_b^{-1} = \begin{bmatrix} A_{44} & A_{45} \\ A_{54} & A_{55} \end{bmatrix}$$

Applying McGilchrist (1993)'s results, we obtain the REML estimators of σ_1^2 and σ_2^2 ,

$$\begin{aligned}
\hat{\sigma}_1^2 &= n^{-1} \{tr(A_{22}) + \hat{\omega}^T \hat{\omega}\} \\
\hat{\sigma}_2^2 &= n^{-1} \{tr(A_{55}) + \hat{v}^T \hat{v}\}
\end{aligned} \tag{A3}$$

and the asymptotic variances of $(\hat{\beta}, \hat{b}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$,

$$\begin{aligned}
var(\hat{\beta}) &= A_{11} \\
var(\hat{b}) &= A_{44} \\
var(\hat{\sigma}_1^2) &= 2(\sigma_1^2)^2 \{n - 2(\sigma_1^2)^{-1} tr(A_{22}) + (\sigma_1^2)^{-2} tr(A_{22}^2)\}^{-1} \\
var(\hat{\sigma}_2^2) &= 2(\sigma_2^2)^2 \{n - 2(\sigma_2^2)^{-1} tr(A_{55}) + (\sigma_2^2)^{-2} tr(A_{55}^2)\}^{-1}
\end{aligned} \tag{A4}$$

where $tr(A)$ represents the trace of matrix A .

Chapter 3

Testing for the presence of a cure fraction

3.1 Introduction

With rapid progress in medical and health sciences, researchers often encounter a portion of patients who are possible to be "cured" or have prolonged disease-free survival. As an effective means to account for such a portion of cured subjects in survival data analysis, the mixture cure model allows to simultaneously estimate the cure rate of subjects and the survival function of those noncured subjects using two separate regression components. Recently, there has been growing interest in the development of mixture cure models for clustered interval-censored data (Banerjee and Carlin, 2004; Kim and Jhun, 2008). In Chapter 2, we have established a semiparametric mixture cure model where the clustered nature of the data was accommodated through random effects in both the logistic regression and proportional hazards (PH) model components in a manner similar to the generalized linear mixed model (GLMM) approach (McGilchrist, 1993).

However, these mixture cure methods often involve rather cumbersome estimation procedures and may not be readily implemented to some extent.

Another challenging problem is that right-censoring, a special case of interval-censoring, may occur for either cured or uncured subjects when fitting the mixture cure model for interval-censored data. It is vague and difficult to determine whether or not a subject having a right-censored event time belongs to the cured subgroup. Particularly, when covariates appear in both the latent cure fraction and event time components, identifiability can be a problematic issue in the mixture models of this sort. As pointed by Farewell (1986) and Kuk and Chen (1992), the mixture cure models should only be used when there is compelling biological evidence suggesting a cure fraction. However, such strong evidence may not be always available in practice. Prior to applying the complex mixture analysis, it is therefore essential to develop a statistical method to test whether a cured proportion is indeed present in the population. If this is not the case, misspecification of a cumbersome mixture cure model can be avoided. While some methods have been studied for this purpose under parametric mixture cure models based on the likelihood ratio test (Goldman, 1984; Maller and Zhou, 1995; Peng et al., 2001) and the score test Zhao et al. (2009) with right-censored event times, tests for settings where data are interval-censored and correlated within clusters have so far not been developed.

In this chapter, we aim to develop formal tests for the possible cure fraction in clustered and interval-censored survival data. We extend the parametric survival models considered in the existing work to a semiparametric Cox model for the component of uncured subjects. Thus our method appears to be more flexible. To account for the within-cluster dependence of subjects, we incorporate random effects in the linear predictors of the

survival model component so that parameter estimation can be established using the GLMM approach. Although Zhao et al. (2009) provided a score test for testing cure fraction in the mixture cure model with clustered right-censored survival data, their test cannot be applied straightforwardly to the current testing problem because of the complex nature of interval-censored data, which requires an additional self-consistent nonparametric maximum likelihood estimation for the nonparametric baseline hazard. In addition, the finite sample performance of their test seems unclear for clustered data as random effects were ignored in their simulation study.

The rest of the chapter is structured as follows. We first review the mixture cure model with random effects for clustered interval-censored data and present briefly the REML estimation procedure. We then derive a Likelihood Ratio Test statistic which asymptotically follows the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution under the null hypothesis when the boundary case is involved in Section 3.3. Section 3.4 further develops the score test for cure fraction under the semiparametric mixture cure model framework. The sampling distribution and power performance of the score test are evaluated through simulation studies in Section 3.5, while a bootstrap approach is further proposed in Section 3.6 for adjusting the significance level of the score test in finite sample situations. Section 3.7 provides two illustrative examples based on data from a smoking cessation study and a breast cancer study, respectively. We conclude in Section 3.8 and leave some technical details to the Appendices.

3.2 Modelling and estimation

The survival function of event time T for the whole population given covariate vector Z defined as (1.3.1) satisfies the following mixture cure model:

$$S_{pop}(t|Z) = 1 - \pi + \pi S(t|Z), \quad (3.2.1)$$

where $S(t|Z) = S_0(t)^{\exp(\eta)}$ takes the PH model for the uncured group as Chapter 2. The probability of a subject being uncured π is modelled by a logistic regression described in Chapter 2, but we fix it as a constant on $(0, 1]$ here for the purpose of testing for the presence of a cure fraction. To address the intra-cluster correlation for clustered data, similar to the work in Chapter 2, an unobserved random effect ω attached to each observation within a certain cluster is introduced into the survival component through linear predictor $\eta = \beta^T Z + \omega$. We assume that ω follows the normal distribution $N(0, \sigma^2)$ independently across clusters. The baseline survival function $S_0(t)$ and d -dimensional vector β are unknown parameters involved in the mixture cure model.

Now we consider a set of interval-censored data collected from m subjects within n clusters with the same notations as Section 2.2,

$\{(L_{ij}, R_{ij}], Z_{ij}, \delta_{ij}) : j = 1, \dots, n_i; i = 1, \dots, n\}$, and $\sum_{i=1}^n n_i = m$. $\omega = (\omega_1, \omega_2, \dots, \omega_n)^T$, and $\eta_{ij} = \beta^T Z_{ij} + \omega_i$. The likelihood function for the observed interval-censored data based on model (3.2.1) is given by

$$Lk = \prod_{i=1}^n \prod_{j=1}^{n_i} [\pi(S(L_{ij}|Z_{ij}; \omega_i) - S(R_{ij}|Z_{ij}; \omega_i))]^{\delta_{ij}} [1 - \pi + \pi S(L_{ij}|Z_{ij}; \omega_i)]^{(1-\delta_{ij})}, \quad (3.2.2)$$

in terms of the unknown parameters β , π and the baseline survival function $S_0(t)$.

Following the estimation procedure proposed in Chapter 2, the REML

estimators of parameters can be achieved via an EM algorithm, embedding a self-consistency algorithm for estimating $S_0(t)$ non-parametrically. In particular, we denote t_{Q-1} being the observed longest follow-up time. The baseline survival at each time t_q is then parameterized via γ_q ,

$$S_0(t_q) = \exp\left(-\sum_{k=1}^q \exp(\gamma_k)\right), \quad q = 1, \dots, Q,$$

where $S_0(t_0) = 1$ when $\gamma_0 = -\infty$, and $S_0(t_Q) = 0$ when $\gamma_Q = \infty$. Those subjects who have $L_{ij} > t_{Q-1}$ are regarded as the cured subjects to contribute to the likelihood. Given random effects ω , the logarithm of likelihood (3.2.2) can thus be rewritten as follows:

$$\begin{aligned} l_1 = \log Lk = \sum_{i=1}^n \sum_{j=1}^{n_i} \log \left\{ \sum_{q=1}^{Q-1} \alpha_{ijq} [\pi(S(t_{q-1}|Z_{ij}; \omega_i) - S(t_q|Z_{ij}; \omega_i))] \right. \\ \left. + \alpha_{ijQ} [1 - \pi + \pi S(t_{Q-1}|Z_{ij}; \omega_i)] \right\}, \end{aligned} \quad (3.2.3)$$

where $S(t_q|Z_{ij}; \omega_i) = S_0(t_q)^{\exp(\eta_{ij})} = e^{-\sum_{k=1}^q \exp(\gamma_k + \eta_{ij})}$, and π is a constant on $(0, 1]$.

In the manner of the GLMM method, the best linear unbiased prediction (BLUP) type log-likelihood function under the model is given by $l = l_1 + l_2$, with l_1 defined in (3.2.3), and $l_2 = -(1/2)[n \log(2\pi\sigma^2) + \sigma^{-2}\omega^T\omega]$. The estimators of parameters $\beta, \gamma = (\gamma_1, \dots, \gamma_{Q-1})^T$, and ω are obtained by maximizing the BLUP-type log-likelihood l first and then extending them to find the REML estimates of β, γ, ω , and σ^2 .

Our interest throughout this chapter is to justify whether it is necessary to employ the complicated mixture analysis, that is, to test for null hypothesis $H_0 : \pi = 1$ against alternative $H_1 : \pi < 1$ in the context of clustered interval-censored survival data.

Under the null hypothesis, the REML estimators of β, γ, ω , and σ^2 are obtained from the log-likelihood function l for $\pi = 1$ corresponding to a single PH model. Under the alternative hypothesis, a mixture model, the REML estimates of β, γ, ω , and σ^2 are obtained from the log-likelihood function l with $\pi = \hat{\pi}$ constrained in $(0, 1)$, which can be obtained by modelling a logistic regression without covariates.

3.3 Score test for the cure fraction

In the related context of mixture models, score tests have been investigated for zero-inflation and over-dispersion for count data (Van den Broek, 1995; Xiang et al., 2006; Yang et al., 2009). The score test becomes appealing for the current testing problem due to its computational simplicity in comparison with the burdensome computational costs of the likelihood ratio tests. For simplicity, denote $\vartheta = (1 - \pi)/\pi$, then $0 \leq \vartheta < \infty$ as $0 \leq 1 - \pi < 1$. We are interested in testing the null hypothesis $H_0 : \pi = 1$ against the alternative hypothesis $H_1 : \pi < 1$, which is equivalent to testing $H_0^* : \vartheta = 0$ versus $H_1^* : \vartheta > 0$. Using notation ϑ , the conditional log-likelihood function l_1 can be rewritten as

$$l_1 = \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ \log \left[\sum_{q=1}^{Q-1} \alpha_{ijq} (S(t_{q-1}|Z_{ij}; \omega_i) - S(t_q|Z_{ij}; \omega_i)) \right. \right. \\ \left. \left. + \alpha_{ijQ} (\vartheta + S(t_{Q-1}|Z_{ij}; \omega_i)) \right] - \log(1 + \vartheta) \right\}, \quad (3.3.1)$$

and $l_2 = -(1/2)[n \log(2\pi\sigma^2) + \sigma^{-2}\omega^T\omega]$. Under the null hypothesis, we have the first derivative of $l = l_1 + l_2$ with respect to ϑ ,

$$\begin{aligned} & \left. \frac{\partial l}{\partial \vartheta} \right|_{\hat{\beta}; \hat{\omega}} \\ &= \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\alpha_{ijQ}}{\sum_{q=1}^{Q-1} \alpha_{ijq} (\hat{S}(t_{q-1}|Z_{ij}; \omega_i) - \hat{S}(t_q|Z_{ij}; \omega_i)) + \alpha_{ijQ}(\vartheta + \hat{S}(t_{Q-1}|Z_{ij}; \omega_i))} \\ & \quad - \frac{m}{1 + \vartheta}, \end{aligned}$$

and the score function

$$Uc(\hat{\beta}, \hat{\omega}, \vartheta = 0) = (0, \dots, 0, \left. \frac{\partial l}{\partial \vartheta} \right|_{\hat{\beta}; \hat{\omega}; \vartheta=0})^T,$$

where the survival function $\hat{S}(t|Z_{ij})$ is evaluated at REML estimates $\hat{\beta}, \hat{\omega}, \hat{\sigma}^2$, and $\vartheta = 0$ under a single semiparametric PH model with random effects, while γ are regarded as nuisance parameters here.

To derive the score test statistic, we partition the Fisher information matrix $\Gamma(\hat{\beta}, \hat{\omega}, \hat{\vartheta})$ as follows:

$$\Gamma(\hat{\beta}, \hat{\omega}, \hat{\vartheta}) = \begin{pmatrix} \Gamma_{\hat{\beta}\hat{\beta}} & \Gamma_{\hat{\beta}\hat{\omega}} & \Gamma_{\hat{\beta}\hat{\vartheta}} \\ & \Gamma_{\hat{\omega}\hat{\omega}} & \Gamma_{\hat{\omega}\hat{\vartheta}} \\ & & \Gamma_{\hat{\vartheta}\hat{\vartheta}} \end{pmatrix} = \begin{pmatrix} \hat{\Gamma}_{11} & \hat{\Gamma}_{12} \\ \hat{\Gamma}_{21} & \hat{\Gamma}_{22} \end{pmatrix},$$

where $\hat{\Gamma}_{22} = \Gamma_{\hat{\vartheta}\hat{\vartheta}}$ is a scalar, $\hat{\Gamma}_{11} = \begin{pmatrix} \Gamma_{\hat{\beta}\hat{\beta}} & \Gamma_{\hat{\beta}\hat{\omega}} \\ & \Gamma_{\hat{\omega}\hat{\omega}} \end{pmatrix}$ and $\hat{\Gamma}_{12} = \hat{\Gamma}_{21}^T =$

$\begin{pmatrix} \Gamma_{\hat{\beta}\hat{\vartheta}} \\ \Gamma_{\hat{\omega}\hat{\vartheta}} \end{pmatrix}$ with all entries obtained from the second derivatives of l evaluated at $\hat{\beta}, \hat{\omega}, \hat{\sigma}^2$ and $\vartheta = 0$; see Appendix A for formulas of derivatives.

Therefore, the score test statistic for testing $H_0^* : \vartheta = 0$ is given by

$$\begin{aligned} Sc &= Uc^T(\hat{\beta}, \hat{\omega}, \vartheta = 0) \Gamma^{-1}(\hat{\beta}, \hat{\omega}, 0) Uc(\hat{\beta}, \hat{\omega}, \vartheta = 0) \\ &= (\hat{\Gamma}_{22} - \hat{\Gamma}_{21} \hat{\Gamma}_{11}^{-1} \hat{\Gamma}_{12})^{-1} \left(\left. \frac{\partial l}{\partial \vartheta} \right|_{\hat{\beta}; \hat{\omega}; \vartheta=0} \right)^2. \end{aligned} \tag{3.3.2}$$

One can see that when the cure proportion is close to 1, the score statistic will increase indefinitely, but will approach 0 as the cure proportion goes to 0. If the test does not involve the boundary of the parameter space, the test statistic Sc exhibits a quadratic form so has an asymptotic χ_1^2 distribution under the null hypothesis H_0^* . When the boundary case is involved, Maller and Zhou (1995) has studied the LRT for independently and exponentially distribution event times with right censoring. To extend this existing work, we show that the LRT statistic also asymptotically follows the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution under the null hypothesis for interval-censored data when the boundary case is involved. The proof of this is sketched in Appendix B. Note that the score test is asymptotically equivalent to the likelihood ratio test. The finite sample behaviour of the score test statistic Sc in (3.3.2) will be examined through simulations in the next section, compared with the asymptotic null distribution of the LRT statistic.

3.4 Simulation study

We conduct Monte Carlo simulations to evaluate the performance of the score test statistic Sc in terms of the sampling distribution under the null hypothesis and the power.

3.4.1 Sampling distribution

We consider a setting with a single covariate Z_{ij} , which is generated from $Bernoulli(0.5)$ for the j th subject in the i th cluster. Under the null hypothesis $H_0 : \pi = 1$ and given values of Z_{ij} , the interval-censored survival data $(T_{Lij}, T_{Rij}, \delta_{ij})$ for $j = 1, \dots, n_i$, $i = 1, \dots, n$ are generated in the absence of a cure fraction. We first generate a random variate T_{ij} from the

exponential distribution with conditional hazard rate $\exp(\beta x_{ij} + \omega_i)$ and a censoring time C_{ij} from the exponential distribution with mean 10. Then specify $\delta_{ij} = 1$ if $T_{ij} \leq C_{ij}$ and $\delta_{ij} = 0$ otherwise. For those $\delta_{ij} = 1$, T_{Lij}, T_{Rij} are generated as Step 4 in Section 2.4.

We consider three choices of number of clusters $n=5, 10$, and 20 with $n_i=10$ and 20 subjects per cluster and $\beta = 0.2$. Random effect ω_i for subjects in cluster i is generated from a normal distribution $N(0, 0.2^2)$. Empirical ordered statistics of Sc are calculated based on 500 replications for each simulation setting, and compared with the corresponding quantiles of reference distribution $0.5\chi_0^2 + 0.5\chi_1^2$. The resulting Q-Q plots are presented in Figure 3.1. The figure shows that the sampling distribution of the score statistic Sc is close to the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution for most of chosen scenarios with clustered interval-censored data.

3.4.2 Power

To investigate the power performance of the test, we generate clustered interval-censored data $(T_{Lij}, T_{Rij}, \delta_{ij})$ under model (3.2.1). To ensure the presence of a cured subgroup among data, we generate a binary latent variable Y_{ij} with $P(Y_{ij} = 1) = \pi$ for the j th subject in the i th cluster. Then $(T_{Lij}, T_{Rij}, \delta_{ij})$ are generated by Steps 1-4 in Section 2.4.

Let $\varsigma(1-a)$ denote the $(1-a)$ th quantile of the distribution $0.5\chi_0^2 + 0.5\chi_1^2$. Since

$$1 - a = P\{0.5\chi_0^2 + 0.5\chi_1^2 < \varsigma(1 - a)\} = 0.5 + 0.5P\{\chi_1^2 < \varsigma(1 - a)\},$$

we have $\varsigma(1 - a) = \chi_1^2(1 - 2a)$. Given values of the uncured probability π and nominal level a , the empirical power of the test is then calculated as the proportion of 500 random samples in which the statistic Sc is greater

Figure 3.1: *Q-Q plots of ordered score statistics against the quantiles of the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution based on 500 samples generated from the mixture cure model with random effects under the null hypothesis.*

than $\chi_1^2(1 - 2a)$ under the alternative hypothesis $H_1 : \pi < 1$,

$$\frac{1}{500} \sum_{k=1}^{500} 1_{[S_{C_k} > \chi_1^2(1-2a)|H_1]},$$

where S_{C_k} is the observed score statistic based on the k th random sample, $k = 1, \dots, 500$.

We set the nominal level $a = 0.1, 0.05$, and 0.01 , and the uncured proportion $\pi = 1, 0.95, 0.85$, and 0.65 corresponding to cases with no cure fraction, small, medium and large fractions of cured subjects, respectively. The values for other parameters are the same as those in the previous subsection. Simulation results are summarized in Table 3.1.

When $\pi \neq 1$, results indicate that under the alternative $H_1 : \pi < 1$ the test is reasonably powerful in rejecting the null hypothesis, given power frequently above 60% in most of the settings considered. The power of the test increases with increasing number of clusters and/or cluster size. The greater values of cure fraction $(1 - \pi)$ improves the empirical power of the test.

When $\pi = 1$, in other words, there is no cure fraction, this corresponds to the significance level (or empirical probability of a Type I error) of the test. The estimated significance levels are small in general, but not very close to the nominal level a in most cases. The actual significance levels are slightly underestimated for small sample cases, and overestimated for cases with moderate number of clusters. This result differs from that with right-censored data reported by Zhao et al. (2009), in which their score test always underestimated the significance level. It could be mainly because their simulated power were calculated using χ_1^2 , instead of $0.5\chi_0^2 + 0.5\chi_1^2$, as the distribution under the null hypothesis was on the boundary of the parameter space.

Table 3.1: *Empirical power of the score statistic based on 500 samples under the mixture cure model.*

n_i	n	$\pi=1.00$			$\pi=0.95$		
		Censoring level (0%-15%)			Censoring level (10%-15%)		
		$a = 0.1$	$a = 0.05$	$a = 0.01$	$a = 0.1$	$a = 0.05$	$a = 0.01$
10	5	0.062	0.030	0.008	0.198	0.128	0.066
	10	0.102	0.048	0.016	0.382	0.282	0.178
	20	0.152	0.086	0.016	0.590	0.534	0.416
20	5	0.078	0.050	0.020	0.388	0.298	0.184
	10	0.154	0.070	0.020	0.618	0.554	0.432
	20	0.142	0.092	0.026	0.758	0.740	0.688
n_i	n	$\pi=0.85$			$\pi=0.65$		
		Censoring level (20%-25%)			Censoring level (40%-45%)		
		$a = 0.1$	$a = 0.05$	$a = 0.01$	$a = 0.1$	$a = 0.05$	$a = 0.01$
10	5	0.592	0.440	0.248	0.960	0.906	0.654
	10	0.890	0.846	0.632	1.000	0.998	0.986
	20	0.966	0.962	0.934	1.000	1.000	1.000
20	5	0.862	0.840	0.698	0.996	0.996	0.980
	10	0.966	0.962	0.940	1.000	1.000	1.000
	20	0.996	0.992	0.992	1.000	1.000	1.000

Our simulation results indicate that the score test based on $0.5\chi_0^2 + 0.5\chi_1^2$ may lead to incorrect estimates of the significance level in finite sample inference, even though the asymptotic null distribution can be derived as such a 50-50 mixture of chi-squares. This phenomenon has been also addressed by several authors when score tests based on mixtures of chi-squares are used for inference about variance components in generalized linear mixed models. See e.g., (Fitzmaurice et al., 2007) and (Sinha, 2009). To overcome the aforementioned problem, we will perform bootstrap inference in the next section to adjust the estimated significance level of the test.

3.4.3 Bootstrap approach

Bootstrap is commonly used for level-adjustment of a test in practice. In this section, we propose and explore a parametric bootstrap version of the score test to account for finite-sample bias in estimated significance levels, which arise from the test Sc with the null distribution $0.5\chi_0^2 + 0.5\chi_1^2$ assumed.

Due to the complex structure of the censoring mechanism, the inference procedure for clustered interval-censored data is often complicated and computationally expensive. We hereby develop a bootstrap test procedure based on similar ideas as in the work of Davidson and MacKinnon (2006). Its implementation details are given as follows.

- (i) Generate sample $(T_{Lij}, T_{Rij}, \delta_{ij})$ for $j = 1, \dots, n_i$, $i = 1, \dots, n$, then derive the REML estimates $(\hat{\beta}, \hat{\omega}, \hat{\sigma}^2)$ and the score statistic Sc .
- (ii) Based on the values of $(\hat{\beta}, \hat{\omega}, \hat{\sigma}^2)$, generate a bootstrap sample under the null hypothesis $H_0 : \pi = 1$. Then compute the corresponding bootstrap estimates $(\hat{\beta}^*, \hat{\omega}^*, \hat{\sigma}^{2*})$ and the bootstrap score statistic Sc^* .

- (iii) Repeat the above process, then obtain the corresponding score statistics Sc_k and bootstrap score statistics Sc_k^* , $k = 1, \dots, K$.
- (iv) Calculate the $(1 - a)$ th quantile of the $\{Sc_k^*\}$, denoted by $Q^*(1 - a)$. Then the approximate rejection probability (p -value) is $\frac{1}{K} \sum_{k=1}^K 1_{[Sc_k > Q^*(1-a)]}$ at the a level of significance.

We now carry out a new simulation study to examine the performance of this bootstrap test. Results of the estimated significance level and power of the bootstrap test for nominal levels $a = 0.1, 0.05, 0.01$ are presented in Table 3.2 based on $K = 500$ bootstrap samples. We observe that using the above bootstrap approach clearly maintains the correct significance level a under the null hypothesis. In the meantime, it improves the power by a large margin under the alternative hypothesis for all cases regardless of the number of clusters and cluster sizes.

3.5 Applications

3.5.1 Smoking cessation data

We first revisit the smoking cessation study to illustrate the application of the proposed tests for the presence of a cured subgroup in the population. The interval-censored smoking relapse time data were collected from a total of 223 smokers among which 158 smokers never resumed smoking, regarded as cured subjects, in contrast to the remaining failed quitters who might resume smoking soon after an initial attempt to quit. This gives a cessation rate of around 71%.

Given the clustered structure arising from the different geographical regions, the possible dependence among subjects residing in the same zip

Table 3.2: *Empirical power of the bootstrap approach based on 500 bootstrap samples under the mixture cure model.*

		$\pi=1.00$			$\pi=0.95$		
		Censoring level (0%-15%)			Censoring level (10%-15%)		
n_i	n	$a = 0.1$	$a = 0.05$	$a = 0.01$	$a = 0.1$	$a = 0.05$	$a = 0.01$
10	5	0.088	0.040	0.010	0.604	0.436	0.272
	10	0.114	0.050	0.008	0.836	0.724	0.480
	20	0.098	0.046	0.008	0.968	0.924	0.872
20	5	0.092	0.050	0.010	0.856	0.784	0.516
	10	0.086	0.046	0.012	0.964	0.940	0.874
	20	0.112	0.058	0.018	1.000	1.000	0.966
		$\pi=0.85$			$\pi=0.65$		
		Censoring level (20%-25%)			Censoring level (40%-45%)		
n_i	n	$a = 0.1$	$a = 0.05$	$a = 0.01$	$a = 0.1$	$a = 0.05$	$a = 0.01$
10	5	0.980	0.942	0.756	1.000	0.998	0.986
	10	0.998	0.998	0.984	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000
20	5	1.000	1.000	0.964	1.000	1.000	0.998
	10	1.000	1.000	1.000	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000

code region is anticipated. In Chapter 2, we have applied a semiparametric mixture cure model in the GLMM framework to accommodate the heterogeneity among clusters, which resulted in substantial improvement in interpreting the inter-cluster variation and impacts of covariates in the risk of smoking relapse. However, the statistical significance of the apparent cure fraction among the data still remains to be confirmed.

To test the cure fraction, we consider all recorded covariates and include random effects via the linear predictor in the survival component of the mixture cure model. Using the estimation procedure provided in Section 3.2, the REML estimates of parameters are obtained for this mixture cure model with random effects. Meanwhile, the REML estimates of the survival model for uncured subjects, a single PH regression with random effects, are also calculated for comparison. Fitting results are summarized in Table 3.3. Based on the mixture cure model with random effects, the cured probability $1 - \pi$ is estimated to be 65%. Some estimated regression coefficients in both models appear to have different inference conclusions. After incorporating the cure fraction, variable sex becomes a significant risk factor for smoking relapse. The intervention type SI/UC is found not significant in both the mixture cure and single PH models. It is shown that the special intervention helps to reduce the risk of relapse by 14% (relative risk $\exp(-0.1524) = 0.86$) only when adjusted by the cured proportion, which is considerably lower than 35% (relative risk $\exp(-0.4298) = 0.65$) given by the single PH model. In other words, the single PH model is likely to overestimate the survival difference among uncured subjects between the two intervention arms.

The score test statistic Sc is computed by 17.143 with p -value < 0.001 under the approximate null distribution $0.5\chi_0^2 + 0.5\chi_1^2$. Meanwhile, the

Table 3.3: *Summary of model fitting for the mixture cure model with random effects and single PH model with random effects to the smoking cessation data.*

Parameter	Mixture cure model		Single PH model	
	with random effects		with random effects	
	Estimate	SE	Estimate	SE
Uncured probability	0.3497*	0.0319		
Sex	0.6990*	0.2712	0.3813	0.2261
SI/UC	−0.1524	0.2958	−0.4298	0.2353
Cigarettes	0.0039	0.0116	0.0116	0.0091
Duration	−0.0302*	0.0128	−0.0277*	0.0100
σ	0.2778	0.3113	0.2661	0.2815

* p -value < 0.05.

bootstrap test based on $K = 1000$ bootstrap samples also produces $p < 0.001$. Both results confirm the existence of a nonnegligible cure fraction in the population.

3.5.2 Breast cosmesis data

This application is concerned with a retrospective study of early breast cancer subjects conducted to compare two treatments with respect to the time to cosmetic deterioration. The detail of the study is provided in Chapter 1. A single covariate considered is treatment type with value 0 for the patients given the radiation therapy (RT) only and 1 for those given the radiation therapy plus adjuvant chemotherapy (RCT). Fitting a single PH model for this data set gives the estimated coefficient $\hat{\beta} = 0.95$ with standard error 0.237. This indicates that the treatment given to patients significantly af-

fects the risk of breast retraction at the 0.05 level, and patients receiving the RCT have 2.6 ($\exp(0.95)$) times higher hazard to cosmetic deterioration than those with the RT. During the study, it was found that 38 patients did not experience breast retraction. To formally test whether there was indeed a significant proportion of patient who were not at risk of deterioration, we apply the score test with test statistic $Sc = 0.009$ and $p = 0.462$ based on the null distribution $0.5\chi_0^2 + 0.5\chi_1^2$, providing insufficient evidence against the null hypothesis that there is no cure fraction. The bootstrap test based on $K = 1000$ bootstrap samples is then applied, which gives $p = 0.731$, also suggesting little evidence of a cure fraction. Thus it is not necessary for incorporating a cure fraction in modelling these data. This conclusion can be verified by Figure 3.2, in which estimated survival functions from both the single PH and mixture cure models tend to agree very well and are close to the empirical estimate.

Moreover, note that the difference in p -values of the two test methods indicates that the 50-50 mixture of chi-square distributions for the score test may result in incorrect inference in certain situations. To illustrate this, we conduct both tests for a subset of the data with a small sample size. The subset consists of 23 out of 46 patients from the RT group and 24 out of 48 subjects from the RCT group. In particular, all the right-censored observations are selected into the subset first, while the rest of the data in the subset are then randomly chosen from those observations with interval censoring in each group. Based on this subset, the estimated regression coefficient is $\hat{\beta} = 0.936$ with standard error 0.364, still keeping the significant effect of the treatment. The score statistic for the subset produces a value of 3.467 with $p = 0.032$, indicating a significant cure fraction at the prespecified 5% level. On the other hand, the bootstrap

Figure 3.2: *Estimates of the survival function for the breast cosmesis data using the nonparametric, single PH and mixture cure models.*

method does not detect any cure fraction given $p = 0.614$. In summary, the two test procedures provide different conclusions, and the bootstrap method should be preferable for a small sample.

3.6 Concluding remarks

In this chapter, we developed a score test for the presence of the cure fraction, which involves the boundary of the parameter space, in interval-censored data within a mixture cure model framework. The advantage of the score test lies in its computational convenience in that it requires only a single model be estimated under the null hypothesis, so avoiding the more cumbersome mixture analysis.

To investigate the null distribution of the score statistic for the boundary case with interval censoring, we first show that the LRT statistic asymptotically follows a 50-50 mixture of a chi-squared distribution and the probability mass at zero under the null hypothesis with interval-censored data independently from exponential distribution, i.e., $0.5\chi_0^2 + 0.5\chi_1^2$. Then we build the finite sample behaviour of the score statistic compared with the reference distribution $0.5\chi_0^2 + 0.5\chi_1^2$, as the score test and LRT are asymptotically equivalent through Monte Carlo simulations. In our simulation studies, the score test statistic based on the null distribution $0.5\chi_0^2 + 0.5\chi_1^2$ showed a reasonably good power, with values greater than 60% in most of the simulation settings considered. However, it inflated significance levels, i.e., probabilities of committing Type I errors, jeopardizing the interpretation of the results. Therefore, we proposed a parametric bootstrap version of the test, which clearly reduced the Type I error rates and preserved them close to the nominal levels.

For assessing the necessity of adjusting the cured proportion in clustered interval-censored data, the score test should be recommended in median and large sample situations in view of its simple implementation and satisfactory power performance, while the parametric bootstrap test could be suggested whenever an accurate assessment of the significance level of the test is required.

Our study extends the existing work on testing for cure fraction with right-censored data to a more general data structure, in which event times may be clustered, interval-censored and free of parametric distribution assumptions. The methods developed can be applied as well for survival data under other censoring schemes such as recurrent data, doubly censoring and mixed case interval censoring.

3.7 Appendices

Appendix A: Derivatives in the score test

Given the denotation of $S(t_q|Z_{ij})$, g_{ij} , f_{ijq} , and h_{ijq} the same as Appendix in Chapter 2,

The first derivatives of l are given by

$$\begin{aligned}\frac{\partial l}{\partial \vartheta} &= \sum_{i=1}^n \sum_{j=1}^{n_i} \left[\frac{\alpha_{ijQ}}{g_{ij} + \alpha_{ijQ}(\vartheta + S(t_{Q-1}; x_{ij}))} - \frac{1}{1 + \vartheta} \right], \\ \frac{\partial l}{\partial \eta_{ij}} &= \frac{\sum_{q=1}^{Q-1} \alpha_{ijq}(f_{ij(q-1)} - f_{ijq}) + \alpha_{ijQ}f_{ij(Q-1)}}{g_{ij} + \alpha_{ijQ}(\vartheta + S(t_{Q-1}; x_{ij}))}, \\ \frac{\partial l}{\partial \beta} &= \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\partial l}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \beta}, \\ \frac{\partial l}{\partial \omega} &= \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\partial l}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \omega} - \frac{\omega}{\sigma^2}.\end{aligned}$$

Then, entries of the Fisher information matrix $\Gamma(\beta, \omega, \vartheta)$ under the null hypothesis $H_0^* : \vartheta = 0$ can be obtained as follows:

$$\begin{aligned}\Gamma_{\beta\beta} &= -\frac{\partial^2 l}{\partial\beta\partial\beta^T} = \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\partial\eta_{ij}}{\partial\beta^T} \left(-\frac{\partial^2 l}{\partial\eta_{ij}\partial\eta_{ij}}\right) \frac{\partial\eta_{ij}}{\partial\beta}, \\ \Gamma_{\omega\omega} &= -\frac{\partial^2 l}{\partial\omega\partial\omega^T} = \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\partial\eta_{ij}}{\partial\omega^T} \left(-\frac{\partial^2 l}{\partial\eta_{ij}\partial\eta_{ij}}\right) \frac{\partial\eta_{ij}}{\partial\omega} + \frac{\mathbf{I}_n}{\sigma^2}, \\ \Gamma_{\beta\omega} &= -\frac{\partial^2 l}{\partial\beta\partial\omega^T} = \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\partial\eta_{ij}}{\partial\omega^T} \left(-\frac{\partial^2 l}{\partial\eta_{ij}\partial\eta_{ij}}\right) \frac{\partial\eta_{ij}}{\partial\beta}, \\ \Gamma_{\beta\vartheta} &= -\frac{\partial^2 l}{\partial\beta\partial\vartheta} = \frac{[\sum_{q=1}^{Q-1} \alpha_{ijq}(f_{ij(q-1)} - f_{ijq}) + \alpha_{ijQ}f_{ij(Q-1)}]\alpha_{ijQ}}{[g_{ij} + \alpha_{ijQ}S(t_{Q-1}|Z_{ij}; \omega_i)]^2} \frac{\partial\eta_{ij}}{\partial\beta}, \\ \Gamma_{\omega\vartheta} &= -\frac{\partial^2 l}{\partial\omega\partial\vartheta} = \frac{[\sum_{q=1}^{Q-1} \alpha_{ijq}(f_{ij(q-1)} - f_{ijq}) + \alpha_{ijQ}f_{ij(Q-1)}]\alpha_{ijQ}}{[g_{ij} + \alpha_{ijQ}S(t_{Q-1}|Z_{ij}; \omega_i)]^2} \frac{\partial\eta_{ij}}{\partial\omega}, \\ \Gamma_{\vartheta\vartheta} &= -\frac{\partial^2 l}{\partial\vartheta\partial\vartheta} = \sum_{i=1}^n \sum_{j=1}^{n_i} \left[\frac{\alpha_{ijQ}}{(g_{ij} + \alpha_{ijQ}S(t_{Q-1}|Z_{ij}; \omega_i))^2} - 1 \right],\end{aligned}$$

where the negative second derivative of l

$$\begin{aligned}-\frac{\partial^2 l}{\partial\eta_{ij}\partial\eta_{ij}} &= \left\{ \frac{\sum_{q=1}^{Q-1} \alpha_{ijq}(f_{ij(q-1)} - f_{ijq}) + \alpha_{ijQ}f_{ij(Q-1)}}{g_{ij} + \alpha_{ijQ}S(t_{Q-1}|Z_{ij}; \omega_i)} \right\}^2 \\ &\quad - \frac{\sum_{q=1}^{Q-1} \alpha_{ijq}(h_{ij(q-1)} - h_{ijq}) + \alpha_{ijQ}h_{ij(Q-1)}}{g_{ij} + \alpha_{ijQ}S(t_{Q-1}|Z_{ij}; \omega_i)}.\end{aligned}$$

Appendix B: Likelihood ratio test for the cure fraction

For inference later, we first derive a likelihood ratio test (LRT) of the cure fraction for independently and exponentially distributed survival data under interval censoring. We notice that this test involves testing at the boundary of the parameter space: $H_0 : \pi = 1$. The boundary case was studied by Maller and Zhou (1995) based on the LRT for independently and exponentially distributed event times under right censoring. They showed that the asymptotic null distribution of the LRT statistic follows a 50-50

mixture of a chi-squared distribution and the probability mass at zero, i.e., $0.5\chi_0^2 + 0.5\chi_1^2$.

In line with the work of Maller and Zhou (1995), we now extend to the context of interval-censored survival data from the same distribution $\exp(\lambda)$ as Maller and Zhou (1995) in which no covariates are considered. Denote the observed data by $\{(L_i, R_i], \delta_i; i = 1, 2, \dots, n\}$, $\delta_i = 1$ indicates the interval-censored data and $\delta_i = 0$ indicates those with right censoring. Followed the likelihood function in (3.2.2) with $S(t) = e^{-\lambda t}$, the log likelihood function of the parameter π for the mixture cure model is derived as,

$$l_n(\pi) = \sum_{i=1}^n \{\delta_i [\log \pi + \log(e^{-\lambda L_i} - e^{-\lambda R_i})] + (1 - \delta_i) \log(1 - \pi + \pi e^{-\lambda L_i})\},$$

where λ is a known hazard rate of the exponential distribution.

The likelihood ratio statistic for H_0 versus H_1 is then obtained by $lD_n = -2[l_n(1) - l_n(\hat{\pi})]$ with $\hat{\pi}$ being the unique MLE of π in the parameter space $(0,1]$. Next we derive the asymptotic distribution of the statistic lD_n under the null hypothesis $H_0 : \pi = 1$.

Take partial derivative of $l_n(\pi)$ with respect to π , then we consider the function $\pi \frac{\partial l_n(\pi)}{\partial \pi}$ which is derived as follows,

$$\begin{aligned} \pi \frac{\partial l_n(\pi)}{\partial \pi} &= \pi \sum_{i=1}^n \frac{\delta_i}{\pi} + (1 - \delta_i) \frac{-1 + e^{-\lambda L_i}}{e^{1-\pi+\pi\lambda L_i}}, \\ &= \sum_{i=1}^n \frac{\delta_i - \pi + \pi e^{-\lambda L_i}}{1 + \pi + \pi e^{\lambda L_i}}. \end{aligned}$$

It is a non-increasing function of π on $(0,1]$ and its value at $\pi = 0$ is $\sum_{i=1}^n \delta_i$, which is greater than zero with probability approaching 1 as n goes to infinity. Equivalently, denote lG_n as the value of $\pi \frac{\partial l_n(\pi)}{\partial \pi}$ at $\pi = 1$,

$$lG_n = n - \sum_{i=1}^n (1 - \delta_i) e^{\lambda L_i}.$$

Then we have $\hat{\pi} = 1$ if and only if $lG_n \geq 0$, and when $lG_n < 0$, $\hat{\pi}$ is the estimator obtained by maximizing the likelihood function in the parameter space $(0, 1)$.

Given a value $x \geq 0$, we then compute

$$\begin{aligned} P(lD_n \leq x) &= P(lD_n \leq x | \hat{\pi} < 1)P(\hat{\pi} < 1) + P(lD_n \leq x | \hat{\pi} = 1)P(\hat{\pi} = 1), \\ &= P(lD_n \leq x | \hat{\pi} < 1)P(lG_n < 0) + P(lG_n \geq 0). \end{aligned}$$

By Lemma 2.1 of Zhou and Maller (1995), for any function of time T , say $\zeta(T)$, $E[(1-\delta)\zeta(T)] = E[E[\zeta(L)1_{[L < T < \infty]} | L]] = E\{\zeta(L)(1-\pi + \pi e^{-\lambda L})\}$. Under the condition $E(e^{\lambda L})$ is finite, denoted by $\sigma_0^2 + 1$, we can derive

$$E\left[\pi \frac{\partial l_n(\pi)}{\partial \pi}\right] = n - \sum_{i=1}^n E[(1 - \delta_i) \frac{1}{1 - \pi + \pi e^{-\lambda L}}] = 0,$$

and the corresponding variance should be $n\sigma_0^2$ for the n samples. Based these facts, $\frac{lG_n}{\sqrt{n\sigma_0^2}}$ converge to the standard normal distribution as $n \rightarrow \infty$ by the central limit theorem. Thus, we conclude $P(lG_n < 0) = P(lG_n \geq 0) = 0.5$.

When $\hat{\pi} < 1$, a Taylor expansion shows that

$$l_n(1) - l_n(\hat{\pi}) = \frac{1}{2}(1 - \hat{\pi})^2 \frac{\partial^2 l_n(\tilde{\pi})}{\partial \pi^2}$$

for some $\tilde{\pi}$ between $\hat{\pi}$ and 1. Applying the weak law of large numbers we can show that $\frac{\partial^2 l_n(\tilde{\pi})}{\partial \pi^2}$ convergent to $n\sigma_0^2$. Thus we have

$$\begin{aligned} P(lD_n \leq x | \hat{\pi} < 1) &= P((1 - \hat{\pi})^2 n\sigma_0^2 \leq x | \hat{\pi} < 1) \\ &\rightarrow P(\chi_1^2 \leq x), \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Therefore,

$$P(lD_n \leq x) = \frac{1}{2}P(\chi_1^2 \leq x) + \frac{1}{2}.$$

The asymptotic distribution of the likelihood ratio statistic lD_n is a 50-50 mixture of a chi-square random variable with one degree of freedom and a

point mass at zero, namely, the $0.5\chi_0^2 + 0.5\chi_1^2$ distribution. Moreover, note that this proof works for any transformation $\zeta(T)$ of time T , so the above result still applies for interval-censored event times from any distribution for which the survival function S satisfies $E\{S(L)^{-1}\} < \infty$.

Chapter 4

Semiparametric mixture cure transformation model with random effects

4.1 Introduction

The mixture cure model is a popular model to analyze survival data with a non-negligible cure fraction by integrating a logistic regression on the cure proportion and a survival model on the uncured subgroup. Suppose T is a non-negative event time of interest, and Z is the corresponding d -dimensional covariate vector which may affect the survival function of T . With possibility of cure, the survival function of T given covariates $Z \in \mathbb{R}^d$ for the whole population is assumed to take a mixture cure model as (1.3.1) $S_{pop}(t|Z) = 1 - \pi(Z) + \pi(Z)S(t|Z)$, where $\pi(Z)$ is the probability of a subject being uncured, and can be modelled by a logistic regression (Farewell, 1982; Kuk and Chen, 1992), $\pi(Z) = [1 + \exp(-b^T W)]^{-1}$ with $W = (1, Z^T)^T$ and b is an unknown $(d + 1)$ -dimensional regression parameter representing the

effects of covariates on the probability of being uncured.

When event times are observed as repeated measurements or collected from several clusters due to geographic exposures, an unobserved random effect ω is incorporated into the model to account for the intracluster dependence. Then the mixture cure model with random effects takes the form

$$S_{pop}(t|Z; \omega) = 1 - \pi(Z) + \pi(Z)S(t|Z; \omega), \quad (4.1.1)$$

where $S(t|Z; \omega)$ is a conditional survival function of the uncured subjects associated with random effect ω . Prevailing modelling techniques have been advocated in the literature for the conditional survival function and successfully utilized for interval-censored independent data, such as the proportional hazards (PH) model introduced by Finkelstein (1986), the proportional odds model by Shen (1998), the partly linear accelerated failure time (AFT) model by Lam and Xue (2005), the semiparametric PH models by Ma (2009, 2010), and so forth.

Limited research has been done recently on multivariate generalizations of the aforementioned models for the conditional survival function to take into account the intracluster correlation. Banerjee and Carlin (2004) studied the parametric AFT model using a Bayesian approach for spatially correlated interval-censored data. Kim and Jhun (2008) provided the semiparametric PH model for the conditional survival function, sharing a common random effect with the logistic regression for the cure proportion. They applied an approximate likelihood approach for estimation. Our Chapter 2 consider a similar semiparametric model by incorporating different random effects in both components of the mixture cure model. A penalized likelihood approach together with the REML method were adopted to estimate

parameters and the nonparametric baseline hazard. Li and Ma (2010) introduced location-scale parametric models to the conditional survival, allowing two random effects to separately account for the varying cure rate and survival over repeated measurements and using maximum likelihood approach for estimation of parameters. A common feature of all these methods is the lack of a sound theoretical justification behind them.

The main purpose of this chapter is to introduce a semiparametric transformation model for the conditional survival function $S(t|Z; \omega)$ in (4.1.1), develop the maximum likelihood estimation procedure and fully investigate the theoretical property of the proposed estimator. The semiparametric transformation model is becoming an attractive alternative to the PH model because it provides a more general framework for studying covariate effects on survival. In the presence of a cure fraction in right censored data, Lu and Ying (2004) and Zeng et al. (2006b) extended the mixture cure model and nonmixture promotion time cure model respectively to semiparametric transformation cure models. Zeng and Lin (2007) further studied the semiparametric transformation models with random effects accounting for the dependence of recurrent times within the same subject, but ignoring the presence of cured individuals. Motivated by the work of Zeng and Lin (2007), we define a semiparametric transformation model for the conditional survival function as follows:

$$S(t|Z; \omega) = \exp(-G\{\Lambda(t) \exp(\beta^T Z + \omega)\}), \quad (4.1.2)$$

where β is an unknown parameter vector representing the effects of covariates on the survival model component, $\Lambda(t)$ is a positive and increasing function of T with $\Lambda(0) = 0$, and function $G(\cdot)$ is thrice continuously differentiable and increasing with $G(0) = 0$ and $G(\infty) = \infty$. These assumptions

ensure that $S(t|Z; \omega)$ is a proper survival function. To a certain extent, the transformation involved in $S(t|Z; \omega)$ provides a wide range in selecting models. Two commonly used transformations are the classical Box-Cox transformation $G(t) = \{(1+t)^\rho - 1\}/\rho$ for $\rho > 0$ and the logarithmic transformation $G(t) = \log(1+rt)/r$ for $r > 0$. When $\rho = 1$, model (4.1.2) corresponds to the well-known PH model. When $r = 1$, it reduces to the proportional odds model. The proposed modelling approach connecting transformations to the mixture cure model with random effects provides a broader class of candidate models for the analysis of clustered and interval-censored data.

Our interest focuses on the maximum likelihood estimators (MLEs) for parameter vectors β, b , the variance component associated with ω , and non-parametric function $\Lambda(\cdot)$. There are a few works so far exploring the theoretical properties of MLEs with interval-censored data. Huang (1996) established the consistency, efficiency and asymptotic normality of the MLE for regression parameter and $n^{1/3}$ convergence rate of the MLE for the baseline cumulative hazard function in the proportional hazards model with case 1 interval-censored data. His results were extended by Ma (2009) to a semiparametric cure model for case 1 interval-censored data with a cured subgroup. For case 2 interval-censored data with a cure fraction, Liu and Shen (2009) showed the strong consistency of the MLEs under the Hellinger distance with a non-mixture cure model.

In the presence of a cure fraction in clustered and interval-censored data, to obtain the asymptotic properties of the estimators proposed is more challenging. This relies on the facts that 1) interval censoring involves two correlated response variables rather than a single variable under right censoring, and 2) the unspecified function $\Lambda(\cdot)$ and random effects make the

score function for all parameters and Λ have no explicit form thus leading to difficulty in obtaining their information matrix. In this chapter, we use empirical process techniques to establish theoretical properties for the proposed transformation model, providing valuable theoretical insights into the analysis of such complex structured data.

The rest of the chapter is organized as follows. Section 4.2 provides the maximum likelihood estimation procedure and its implementation via an EM algorithm. In Section 4.3, we establish the model identifiability and asymptotic properties of the resulting estimators. Section 4.4 demonstrates the finite-sample performance of the proposed estimators via simulation studies and illustrates the application of our method by analyzing relapse time data from a smoking cessation study. We conclude the chapter in Section 4.5 with a few remarks. The proofs of theoretical results are given in Appendices.

4.2 Estimation procedure

We assume that the censoring in this work is non-informative and independent of the cure fraction and the event time for uncured subjects. Define the density function of unobserved random effects ω as $\phi(\omega; \sigma)$ with mean 0 and scale parameter σ . There are considerable flexibilities in choosing the transformation G and the density $\phi(\omega; \sigma)$. In general, we could choose any continuous distribution density for $\phi(\omega; \sigma)$. The normal or gamma distribution is typically used in the literature (e.g. Nelson et al., 2006; Zeng et al., 2008).

Suppose that there are n independent clusters unless otherwise mentioned. Each cluster has potentially different sizes. The observations are

denoted as $\{X_{ij} = (L_{ij}, R_{ij}, Z_{ij}, \delta_{ij}) : j = 1, \dots, n_i; i = 1, \dots, n\}$ with the same definition in Section 2.2. $\sum_{i=1}^n n_i$ is the total sample size which has the same order as the number of clusters n (Li et al., 2003; Zeng et al., 2008). Let $\theta = (\beta^T, b^T, \sigma)^T$ which is the unknown parameter set. Our interest centers on the estimation of (θ, Λ) in the proposed model. Given both the covariates and random effects, the likelihood of (θ, Λ) is in the form of

$$L(\theta, \Lambda | X_{ij}; \omega_i) = \{\pi(Z_{ij})[S(L_{ij} | Z_{ij}; \omega_i) - S(R_{ij} | Z_{ij}; \omega_i)]\}^{\delta_{ij}} \\ \{1 - \pi(Z_{ij}) + \pi(Z_{ij})S(L_{ij} | Z_{ij}; \omega_i)\}^{1-\delta_{ij}}.$$

As is the same with any random effects model, integrating over the distribution of the random effects gives the marginal likelihood function:

$$L(\theta, \Lambda) = \prod_{i=1}^n \int \prod_{j=1}^{n_i} L(\theta, \Lambda | X_{ij}; \omega_i) \phi(\omega_i; \sigma) d\omega_i. \quad (4.2.1)$$

We then obtain the estimators of θ and Λ , denoted by $\hat{\theta}_n$ and $\hat{\Lambda}_n$, through maximizing the above marginal likelihood function (4.2.1) over θ and Λ , and treating Λ as an increasing step function with possible jumps only at the interval endpoints and $\Lambda(0) = 0$.

The proposed estimation can be implemented using an EM algorithm. Particularly, to account for the cure possibility, we introduce a latent binary variable Y_{ij} as Section 2.2, where $Y_{ij} = 1$ if the j th subject in cluster i is uncured and $Y_{ij} = 0$ otherwise. Obviously, $Y_{ij} = 1$ if $\delta_{ij} = 1$. Thus, the complete-data marginal likelihood function can be written as

$$L_c(\theta, \Lambda) = \prod_{i=1}^n \left\{ \prod_{j=1}^{n_i} \pi(Z_{ij})^{y_{ij}} (1 - \pi(Z_{ij}))^{1-y_{ij}} \right\} \\ \left\{ \int \prod_{j=1}^{n_i} [S(L_{ij} | Z_{ij}; \omega_i) - \delta_{ij} S(R_{ij} | Z_{ij}; \omega_i)]^{y_{ij}} \phi(\omega_i; \sigma) d\omega_i \right\},$$

where $S(T|\cdot)$ is a proper survival function, $S(0|\cdot) = 1$ and $S(\infty|\cdot) = 0$. Let $\hat{E}[\cdot]$ denote the conditional expectation given observable data and current

parameter estimates. Treating the latent variable Y_{ij} and random effect ω_i as missing data, the EM algorithm proceeds as follows.

E-step. We calculate the expectation $\hat{E}_\omega[\zeta(\omega_i)]$ for some function $\zeta(\omega_i)$ as

$$\hat{E}_\omega[\zeta(\omega_i)] = \frac{\int \zeta(\omega_i) \prod_{j=1}^{n_i} L(\theta, \Lambda | X_{ij}; \omega_i) \phi(\omega_i; \sigma) d\omega_i}{\int \prod_{j=1}^{n_i} L(\theta, \Lambda | X_{ij}; \omega_i) \phi(\omega_i; \sigma) d\omega_i},$$

where integrations are evaluated by a numerical approximation such as Gauss Quadrature. The conditional expectation of Y_{ij} for the j th individual within cluster i is computed by

$$\hat{E}_Y[Y_{ij}] = \delta_{ij} + \frac{(1 - \delta_{ij})\pi(Z_{ij})\hat{E}_\omega[S(L_{ij}|Z_{ij}; \omega_i)]}{1 - \pi(Z_{ij}) + \pi(Z_{ij})\hat{E}_\omega[S(L_{ij}|Z_{ij}; \omega_i)]}.$$

M-step. We maximize the complete-data log-likelihood function $l_c = \log L_c(\theta, \Lambda)$, which can be constructed additively as $l_c = l_1(\beta, \Lambda) + l_2(b) + l_3(\sigma)$, where

$$\begin{aligned} l_1(\beta, \Lambda) &= \sum_{i=1}^n \sum_{j=1}^{n_i} \hat{E}_Y[Y_{ij}] \hat{E}_\omega[\log(S(L_{ij}|Z_{ij}; \omega_i) - \delta_{ij}S(R_{ij}|Z_{ij}; \omega_i))], \\ l_2(b) &= \sum_{i=1}^n \sum_{j=1}^{n_i} \hat{E}_Y[Y_{ij}] \log \pi(Z_{ij}) + (1 - \hat{E}_Y[Y_{ij}]) \log(1 - \pi(Z_{ij})), \\ l_3(\sigma) &= \sum_{i=1}^n \hat{E}_\omega[\log \phi(\omega_i; \sigma)]. \end{aligned}$$

The parameters b and σ can be updated in the E-step by their corresponding estimators obtained through maximizing $l_2(b)$ and $l_3(\sigma)$ in the M-step respectively. To obtain the updated values of (β, Λ) , we need to handle the unknown function Λ in the maximization procedure of $l_1(\beta, \Lambda)$.

We employ an iterative procedure to find a self-consistent estimator Λ by treating it as a step function parallel to the method of Turnbull (1976). We further assume, similar to the work of Liu and Shen (2009), that Λ has jumps only at the distinct interval end points defined as $0 < t_1 < t_2 < \dots < t_Q <$

$t_{Q+1} = \infty$, in which $t_q \in \{R_{ij} : \delta_{ij} = 1\}$ and the subinterval (t_q, t_{q+1}) must contain at least one member of $\{L_{ij}\}$. Denoting the strictly positive jump-size of Λ at t_q by e^{γ_q} for $q = 1, \dots, Q$, we then have $\Lambda(t) = \sum_{q=1}^Q e^{\gamma_q} 1_{[t_q \leq t]}$. Consequently, estimating Λ is equivalent to estimating $\gamma = (\gamma_1, \dots, \gamma_Q)$ in the corresponding log-likelihood function. The Newton-Raphson method can be used to compute the updated values of θ and γ numerically in the M-step.

Given the updated values of θ and γ , the iteration stops until all estimators converge. Let $\hat{\theta}$ and $\hat{\gamma}$ be the resulting estimators. Finally, the asymptotic variances of $\hat{\theta}$ and $\hat{\gamma}$ can be estimated by the inverse of observed information matrix. See Appendix A for details of the observed information matrix.

4.3 Large sample properties

The technical challenges of theoretically justifying the estimators proposed arise from the mixture model framework, the random effects involved in the likelihood, and the complex nature of interval censoring mechanism that involves two related variables U and V .

We start by giving some notation and assumptions that required in this section. Assume that the unobservable event time T is independent of (U, V) given covariate vector Z , and the joint distribution of (U, V, Z) does not depend on the parameters of interest (θ, Λ) . This independence assumption has been commonly set up in many studies with interval-censored data. Denote true values of θ and Λ by θ_0 and Λ_0 respectively. To derive theoretical results of the estimators proposed, we impose the following conditions:

- Condition 1. The covariate vector Z has a bounded support and its distribution can not be concentrated on R^{d_0} with dimension $d_0 < d$.
- Condition 2. (a) There exists a positive constant c such that $P(V - U \geq c) = 1$. (b) The union of the support for U and V is contained in a closed interval $[\tau_0, \tau_1]$, where $0 < \tau_0 < \tau_1 < \infty$.
- Condition 3. The function Λ_0 has strictly positive and bounded continuous derivative on $[\tau_0, \tau_1]$, and θ_0 lies in the interior of a compact set $\Theta \subset R^d$, which is a finite dimensional parameter space of θ .
- Condition 4. The conditional density $\psi(U, V|Z; \omega)$ of (U, V) given Z and random effects ω has bounded partial derivatives with respect to U and V . The bounds of these partial derivatives do not depend on Z or ω , and Z has a density $p(Z)$.
- Condition 5. (a) If $\phi(\omega; \sigma) = \phi(\omega; \sigma_0)$ for almost every ω , then $\sigma = \sigma_0$. (b) $\phi'(\omega; \sigma_0)^T \mathbf{a}_\sigma = 0$ almost everywhere for ω implies $\mathbf{a}_\sigma = \mathbf{0}$, where $\phi'(\omega; \sigma)$ is the derivative of ϕ with respect to σ .
- Condition 6. For any finite constant C_0 , $\sup_\sigma \int e^{C_0|\omega|} |\phi^{(k)}(\omega; \sigma)| d\omega < \infty$ for $k = 0, 1, 2$, where $\phi^{(k)}$ is the k th derivative of ϕ with respect to σ .

Conditions 1-4 are standard in the proofs of the consistency and weak convergence for interval-censored data (Huang, 1996; Zeng et al., 2006a). Condition 5 ensures the unique parametrization of the random-effects distribution and a nonsingular information matrix. See Zeng and Lin (2007)

for details. Condition 6 refers to the random-effects distribution. It is similar to C.8 in Zeng et al. (2008) and satisfied by the normal distribution and other distributions with tails less heavy than the exponential distribution.

Furthermore, Condition 2 (a) tell us that the examination times are away from zero and ∞ . The observed interval $(L_{ij}, R_{ij}]$ in the previous section and chapters has one of the following forms: $(0, U]$, $(U, V]$, (V, ∞) , corresponding to the left-censored case with indicator $\delta^1 = 1_{[T \leq U]}$, the interval-censored case with indicator $\delta^2 = 1_{[U < T \leq V]}$, and right-censored case with indicator $1 - \delta^1 - \delta^2$, respectively.

4.3.1 Identifiability

Identifiability is the foundation for parameter estimation in mixture models. The lack of identifiability makes inferences in the proposed model meaningless. Therefore, we consider the identifiability issue first. Suppose that $(T_1, T_2, \dots, T_N)$ are random samples in a common cluster with corresponding covariates $Z_1, \dots, Z_N$. The joint survival function of $(T_1, T_2, \dots, T_N)$ is given by

$$\int \prod_{j=1}^N [1 - \pi(Z_j) + \pi(Z_j)S(t_j|Z_j; \omega)] \phi(\omega; \sigma) d\omega.$$

Let $\pi_0(Z)$ and $S_0(t|Z; \omega)$ be the uncured probability and survival function under true parameters θ_0 and Λ_0 . Then the identifiability of the proposed model is meant that the equation

$$\begin{aligned} & \int \prod_{j=1}^N [1 - \pi(Z_j) + \pi(Z_j)S(t_j|Z_j; \omega)] \phi(\omega; \sigma) d\omega \\ &= \int \prod_{j=1}^N [1 - \pi_0(Z_j) + \pi_0(Z_j)S_0(t_j|Z_j; \omega)] \phi(\omega; \sigma_0) d\omega \end{aligned} \tag{4.3.1}$$

can imply $\beta = \beta_0, b = b_0, \sigma = \sigma_0$ and $\Lambda = \Lambda_0$. To show the identifiability of the model, we have the following theorem.

Theorem 4.3.1. *Under Conditions 1 and 5, the proposed mixture cure transformation model with random effects is identifiable.*

Proof. Note that the result $b = b_0$ will be obtained if we can show $b^T W = b_0^T W$ for all $W = (1, Z^T)^T$. Suppose $b \neq b_0$. There exists at least one covariate vector Z such that $b^T W \neq b_0^T W$. In other words, $\pi(Z) \neq \pi_0(Z)$ from the definition of π . On the other hand, when each $t_j \rightarrow \infty$, equation (4.3.1) becomes

$$\prod_{j=1}^N [1 - \pi(Z_j)] = \prod_{j=1}^N [1 - \pi_0(Z_j)] \quad (4.3.2)$$

as $\int \phi(\omega; \sigma) d\omega = \int \phi(\omega; \sigma_0) d\omega = 1$ and $S(\infty|Z; \omega) = 0$. Note that equation (4.3.1) is satisfied for any time sequence $\{t_j\}_{j=1}^N$ and any covariate vector Z_j . Now we choose N covariate vectors $Z_1, \dots, Z_N$, among which only one covariate vector, say Z^1 satisfies $\pi(Z^1) \neq \pi_0(Z^1)$ and $\pi(Z^j) = \pi_0(Z^j)$ for $j = 2, \dots, N$. This is in contradiction to equation (4.3.2). Thus, we obtain $\pi(Z) = \pi_0(Z)$ for any Z , which implies $b = b_0$ from Condition 1.

Next we show the results about β , Λ and σ . Given $b = b_0$, equation (4.3.1) is rewritten as

$$\begin{aligned} & \int \prod_{j=1}^N [1 - \pi(Z_j) + \pi(Z_j) S(t_j|Z_j; \omega)] \phi(\omega; \sigma) d\omega \\ &= \int \prod_{j=1}^N [1 - \pi(Z_j) + \pi(Z_j) S_0(t_j|Z_j; \omega)] \phi(\omega; \sigma_0) d\omega \end{aligned} \quad (4.3.3)$$

We denote a set of size 2^N by Δ consisting all subsets of $R_N = \{1, 2, \dots, N\}$, and expand the products on both sides of equation (4.3.3). For example, each term on the left-hand side can be written in the form $\prod_{k \in R_N - J} [1 -$

$\pi(Z_k)] \prod_{j \in J} \pi(Z_j) S(t|Z; \omega)$, $J \in \Delta$. Comparing terms on both sides corresponding to covariate Z_j , we have

$$\begin{aligned} & \int \prod_{j \in J} \exp(-G\{\Lambda(t_j) \exp(\beta^T Z_j + \omega)\}) \phi(\omega; \sigma) d\omega \\ &= \int \prod_{j \in J} \exp(-G\{\Lambda_0(t_j) \exp(\beta_0^T Z_j + \omega_0)\}) \phi(\omega; \sigma_0) d\omega \end{aligned} \quad (4.3.4)$$

Let $g(t) = G\{\Lambda(t) \exp(\beta^T Z + \omega)\}$ be a function of t conditional on Z and ω . When event time T has survival function $S(t|Z; \omega) = \exp(-G\{\Lambda(t_j) \exp(\beta^T Z_j + \omega)\})$, then $g(T)$ also has survival function $S(\Lambda^{-1}\{G^{-1}(g) \exp(-\beta^T Z - \omega)\}) = \exp(-g)$. So $\int \prod_{j \in J} \exp(-g_j) \phi(\omega; \sigma) d\omega$ is the joint survival function of $\{g_j\}_{j \in J}$. From equation (4.3.4), we have that $\{G\{\Lambda(t_j) \exp(\beta^T Z_j + \omega)\}\}_{j \in J}$ and $\{G\{\Lambda_0(t_j) \exp(\beta_0^T Z_j + \omega_0)\}\}_{j \in J}$ share the same joint survival function, where $\omega \sim \phi(\omega; \sigma)$ and $\omega_0 \sim \phi(\omega_0; \sigma_0)$. Under the same transformation function $\log G^{-1}(\cdot)$, it follows that the joint survival function of $\{\log \Lambda(t_j) + \beta^T Z_j + \omega\}_{j \in J}$ is the same as that of $\{\log \Lambda_0(t_j) + \beta_0^T Z_j + \omega_0\}_{j \in J}$. Since the mean of random effects is assumed to be 0, we have $\log \Lambda(t) + \beta^T Z = \log \Lambda_0(t) + \beta_0^T Z$, which is independent of random effect ω . Thus, $\beta = \beta_0$ and $\Lambda(t) = \Lambda_0(t)$ can be obtained from Condition 1. Consequently, ω has the same distribution as ω_0 . That is, $\phi(\omega; \sigma) = \phi(\omega_0; \sigma_0)$ for almost every ω . Therefore, $\sigma = \sigma_0$ from Condition 5.

□

The result in this theorem is more general than that for a cure survival model provided by Ma (2010) with independent interval-censored data since his semiparametric two-part model is included in our model class as a special case.

4.3.2 Information for θ

The main difficulty of obtaining the information for θ lies in the fact that random effects and a nonparametric baseline hazard function are assumed in the survival model associated with uncured subjects. Here we consider the log-likelihood function for samples from a cluster of size N . Note that any two responses in the cluster are independent conditional on random effects. Given observation $X = (U, V, Z, \delta^1, \delta^2)$, the log-likelihood $l(\theta, \Lambda|X)$ is given by

$$\begin{aligned} l(\theta, \Lambda|X) &= \log L(\theta, \Lambda|X) \\ &= \log \int \prod_{j=1}^N L(\theta, \Lambda|X_j; \omega) \phi(\omega; \sigma) d\omega, \end{aligned} \quad (4.3.5)$$

where

$$\begin{aligned} L(\theta, \Lambda|X; \omega) &= [\pi(Z)(1 - S(U|Z; \omega))]^{\delta^1} [\pi(Z)(S(U|Z; \omega) - S(V|Z; \omega))]^{\delta^2} \\ &\quad [1 - \pi(Z) + \pi(Z)S(V|Z; \omega)]^{1-\delta^1-\delta^2}. \end{aligned}$$

The score functions for θ , $\dot{l}_\theta = (\dot{l}_\beta^T, \dot{l}_b^T, \dot{l}_\sigma^T)^T$, are then obtained as

$$\begin{aligned} \dot{l}_\beta &= \frac{\partial}{\partial \beta} l(\theta, \Lambda|X) \\ &= \sum_{j=1}^N Z_j e^{\beta Z_j} \{ \delta_j^1 A_{1j}(X) \Lambda(U_j) - \delta_j^2 [A_{2j}(X) \Lambda(U_j) - A_{3j}(X) \Lambda(V_j)] \\ &\quad - (1 - \delta_j^1 - \delta_j^2) A_{4j}(X) \Lambda(V_j) \}, \\ \dot{l}_b &= \frac{\partial}{\partial b} l(\theta, \Lambda|X) = \sum_{j=1}^N W_j [\delta_j^1 + \delta_j^2 - \pi(Z_j) + (1 - \delta_j^1 - \delta_j^2) A_{5j}(X)], \\ \dot{l}_\sigma &= \frac{\partial}{\partial \sigma} l(\theta, \Lambda|X) = L^{-1}(\theta, \Lambda|X) \int \prod_{j=1}^N L(\theta, \Lambda|X_j; \omega) \frac{\partial \phi(\omega; \sigma)}{\partial \sigma} d\omega, \end{aligned}$$

where the formulas for $\{A_{rj}(X)\}_{r=1}^5$ are

$$A_{1j}(X) = \frac{\int \prod_{k=1}^N L(\theta, \Lambda|X_k; \omega) \frac{S(U_j|Z_j; \omega) G' \{ \Lambda(U_j) e^{\beta^T Z_j + \omega} \}}{1 - S(U_j|Z_j; \omega)} e^{\omega} \phi(\omega; \sigma) d\omega}{L(\theta, \Lambda|X)},$$

$$A_{2j}(X) = \frac{\int \prod_{k=1}^N L(\theta, \Lambda|X_k; \omega) \frac{S(U_j|Z_j; \omega) G' \{ \Lambda(U_j) e^{\beta^T Z_j + \omega} \}}{S(U_j|Z; \omega) - S(V_j|Z_j; \omega)} e^{\omega} \phi(\omega; \sigma) d\omega}{L(\theta, \Lambda|X)},$$

$$A_{3j}(X) = \frac{\int \prod_{k=1}^N L(\theta, \Lambda|X_k; \omega) \frac{S(V_j|Z_j; \omega) G' \{ \Lambda(V_j) e^{\beta^T Z_j + \omega} \}}{S(U_j|Z; \omega) - S(V_j|Z_j; \omega)} e^{\omega} \phi(\omega; \sigma) d\omega}{L(\theta, \Lambda|X)},$$

$$A_{4j}(X) = \frac{\int \prod_{k=1}^N L(\theta, \Lambda|X_k; \omega) \frac{\pi(Z_j) S(V_j|Z_j; \omega) G' \{ \Lambda(V_j) e^{\beta^T Z_j + \omega} \}}{1 - \pi(Z_j) + \pi(Z_j) S(V_j|Z_j; \omega)} e^{\omega} \phi(\omega; \sigma) d\omega}{L(\theta, \Lambda|X)},$$

$$A_{5j}(X) = \frac{\int \prod_{k=1}^N L(\theta, \Lambda|X_k; \omega) \frac{\pi(Z_j) S(V_j|Z_j; \omega)}{1 - \pi(Z_j) + \pi(Z_j) S(V_j|Z_j; \omega)} \phi(\omega; \sigma) d\omega}{L(\theta, \Lambda|X)}.$$

By Condition 2, $\{A_{rj}\}_{r=1}^5$ are positive for $j = 1, \dots, N$.

To obtain the information for θ only, we treat Λ as a nuisance parameter. Since Λ is a function, the derivative of $l(\theta, \Lambda|X)$ with respect to Λ can be derived as follows. Let F be the distribution function corresponding to Λ , and $L_2^0(F) \equiv \{f : \int f dF = 0 \text{ and } \int f^2 dF < \infty\}$. It follows from Hardy's inequality that $h = \frac{\partial}{\partial a} \Lambda_a \Big|_{a=0} \in L_2(F)$ (Huang and Wellner, 1995). Then the score operator for Λ is

$$\begin{aligned} i_\Lambda(h) &= \frac{\partial}{\partial a} l(\theta, \Lambda_a|X) \Big|_{a=0} \\ &= \sum_{j=1}^N e^{\beta^T Z_j} \{ \delta_j^1 A_{1j}(X) h(U_j) - \delta_j^2 [A_{2j}(X) h(U_j) - A_{3j}(X) h(V_j)] \\ &\quad - (1 - \delta_j^1 - \delta_j^2) A_{4j}(X) h(V_j) \}. \end{aligned}$$

Note that $\dot{l}_\Lambda$ maps $L_2^0(F)$ to $L_2^0(P)$, where P is the joint probability measure of (U, V, Z) . Let $\dot{l}_\Lambda^T : L_2^0(F) \rightarrow L_2^0(P)$ be the ad joint operator of $\dot{l}_\Lambda$, so that $\langle g^*, \dot{l}_\Lambda(h) \rangle_P = \langle \dot{l}_\Lambda^T(g^*), h \rangle_F$ for any $h \in L_2^0(F)$ and $g^* \in L_2^0(P)$, where $\langle \cdot, \cdot \rangle_P$ and $\langle \cdot, \cdot \rangle_F$ are the inner product in $L_2^0(P)$ and $L_2^0(F)$, respectively. To ensure an efficient score for θ , we need to find h^* such that $\langle \dot{l}_\theta - \dot{l}_\Lambda(h^*), \dot{l}_\Lambda(h) \rangle_P = \langle \dot{l}_\Lambda^T(\dot{l}_\theta - \dot{l}_\Lambda(h^*)), h \rangle_F = 0$, equivalently, to solve the following normal equations:

$$\dot{l}_\Lambda^T \dot{l}_\Lambda(h_\beta^*) = \dot{l}_\Lambda^T \dot{l}_\beta \quad (4.3.6)$$

$$\dot{l}_\Lambda^T \dot{l}_\Lambda(h_b^*) = \dot{l}_\Lambda^T \dot{l}_b \quad (4.3.7)$$

$$\dot{l}_\Lambda^T \dot{l}_\Lambda(h_\sigma^*) = \dot{l}_\Lambda^T \dot{l}_\sigma \quad (4.3.8)$$

We summarize the main results in the theorem below. The solution details are deferred to Appendix B.

Theorem 4.3.2. *For $\beta \in R^p$, $b \in R^{p+1}$, $\sigma \in R$, under Conditions 1-6, the equation (4.3.6) - (4.3.8) has a unique solution h_β^* , h_b^* , h_σ^* , respectively. The efficient score for θ is*

$$l_\theta^* = \dot{l}_\theta - \dot{l}_\Lambda(h^*),$$

where $h^* = (h_\beta^{*T}, h_b^{*T}, h_\sigma^*)^T$ with dimension $(2d+2)$, and each component has a bounded derivative. And the information bound for θ is $I(\theta) = E[l_\theta^* l_\theta^{*T}]$. Under Conditions 1-6, $I(\theta)$ is a positive definite matrix with finite entries.

4.3.3 Asymptotic properties for the estimators proposed

The following theorems demonstrate that the estimators developed in Section 4.2 are consistent, asymptotically normal and semiparametrically effi-

cient. Convergence rates of estimators of parameters and the nonparametric component are also obtained.

Theorem 4.3.3. *(Consistency) Under Conditions 1-6, we have $\hat{\theta}_n \rightarrow \theta_0$ almost surely (a.s.), and $\hat{\Lambda}_n \rightarrow \Lambda_0$ a.s. at all mass points of U and V .*

Theorem 4.3.3 affirms the consistency of MLEs of regression parameters in Huang (1996); Huang and Wellner (1995) and Ma (2010) for independent interval-censored data, meanwhile, establishes the consistency of the variance component estimator $\hat{\sigma}_n$. The proof is omitted here for the sake of space since it follows along the same lines as the proof of Theorem 5.14 in Van der Vaart (2000) and by using the fact that any two parameter vectors $\theta^* = \theta_0$ if θ^* and θ_0 satisfy equation (4.3.1).

Theorem 4.3.4. *(Rate of convergence) Under Conditions 1-6, $D((\hat{\theta}_n, \hat{\Lambda}_n), (\theta_0, \Lambda_0)) = O_p(n^{-1/3})$, where $D((\theta_1, \Lambda_1), (\theta_2, \Lambda_2)) = \|\theta_1 - \theta_2\| + \{\int[(\Lambda_1(U) - \Lambda_2(U))^2 + (\Lambda_1(V) - \Lambda_2(V))^2]d\Psi(U, V)\}^{1/2}$, $\|\cdot\|$ is Euclidean norm and $\Psi(U, V)$ is the joint distribution of (U, V) .*

This result shows that the rate of convergence of the estimated nonparametric component $\hat{\Lambda}_n$ is $n^{1/3}$, coincided with the results in Huang (1996) and Huang and Wellner (1995). The detailed proof of this theorem is contained in Appendix C.

Theorem 4.3.5. *(Asymptotic normality and efficiency) Assume Conditions 1-6 hold,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0)^{-1}),$$

where $I(\theta_0) = E[l_{\theta_0}^* l_{\theta_0}^{*T}]$ with $l_{\theta_0}^*$ being the efficient score as defined in Theorem 4.3.2 evaluated at θ_0 . Furthermore, $\hat{\theta}_n$ is asymptotically efficient.

The proof of Theorem 4.3.5 is mainly based on verifying the two assumptions given in Theorem 6.1 of Huang (1996). See Appendix D for a sketch of the proof. In Theorem 4.3.5, $\hat{\theta}_n$ is shown to be asymptotically efficient in the sense that its asymptotic covariance matrix attains the semiparametric efficiency bound developed by Bickel et al. (1993). Note that results in Theorems 4.3.4-4.3.5 are more general than those for a semiparametric two-part model provided by Ma (2010) since our results cover Ma's findings as special cases. In particular, we do not need to assume $E[l(\theta, \Lambda|X) - l(\theta_0, \Lambda_0|X)] \leq -k_1(||\theta - \theta_0|| + ||\Lambda - \Lambda_0||_2)$ with a fixed constant $k_1 > 0$ as assumption (A5) imposed in Ma (2010), which may be unrealistic in practice. Furthermore, from the asymptotic normality of $\hat{\sigma}_n$ given in Theorem 4.3.5, a Wald test statistic can be defined as $\hat{\sigma}_n/SE(\hat{\sigma}_n)$ and applied for testing whether the random effects are statistically significant in the model, i.e., $H_0 : \sigma = 0$ versus $H_1 : \sigma \neq 0$.

4.4 Numerical analysis

To assess the finite sample performance of the proposed method, we conduct intensive simulations and a real data analysis of relapse times from a smoking cessation study.

4.4.1 Simulation study

We carry out simulations under the mixture cure transformation model with random effects (4.1.1), where $\pi(Z) = [1 + \exp(-b_1 - b_2 Z)]^{-1}$ and $S(t|Z; \omega) = \exp(-G\{\Lambda(t) \exp(\beta Z + \omega)\})$. Four different transformations are considered from popular transformation families: the Box-Cox transformation $G(t) = \{(1 + t)^\rho - 1\}/\rho$ with $\rho = 1, 0.5$, and the logarithmic

transformation $G(t) = \log(1 + rt)/r$ with $r = 0.5, 1$. Let function $\Lambda(T)$ be $0.5T$ in the Box-Cox transformation family and T in the logarithmic transformation family. A single covariate Z is simulated from the *Bernoulli* (0.5) distribution and random effects ω are from $N(0, \sigma^2)$. The right censoring times are generated in a similar manner as Chapter 2, where times are censored by an exponential distribution with mean 10. For those interval/left censored cases, we first generated a time sequence on $[0, 5]$ with the length of each time interval from *uniform* $[0.05, 0.1]$, then extract two consecutive time points that event time T_{ij} locates in between to form an observed interval for subject j in cluster i (Ma, 2010). We set true values of $(\beta, b_1, b_2, \sigma)$ as $(0.5, 1, 0.5, 0.5)$ which yield the rates of right-censored and cured subjects at 34%, 38%, 36% and 32% on average under four types of transformed models, respectively. We consider two different numbers of clusters $n=100$ or 200 with 2 observations per cluster. The number of replications in all the simulations is 1000.

The estimates of parameters $(\beta, b_1, b_2, \sigma^2)$ are evaluated in terms of the bias, estimated standard error (SE), sample standard error (SSE), root mean square error (RMSE) and the coverage probability (CP) of the 95% confidence interval based on the normal approximations to parameter estimates. In particular, SEs of estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are obtained using the inverse of the observed information matrix given in Appendix A. Due to the complicated forms of asymptotic variances, SEs of $(\hat{b}_1, \hat{b}_2)$ are computed by the weighted bootstrap approach (Ma, 2010) with $\exp(1)$ distributed weights to estimate standard errors of $(\hat{b}_1, \hat{b}_2)$ based on 500 bootstrap samples.

Given the known transformations, simulation results summarized in Tables 4.1-4.2 show that the biases of all parameter estimates are close to

0, indicating that the parameter estimates are unbiased. The estimated standard errors of $(\hat{\beta}, \hat{b}_1, \hat{b}_2)$ are very close to their empirical counterparts, while the estimated SEs of $\hat{\sigma}^2$ are slightly greater than the corresponding empirical ones. The SEs, SSEs, and RMSEs of all parameters are almost compatible and diminished simultaneously when the number of clusters increases. The values of CP are close to the nominal value 95% and further evident to our proposed asymptotic properties.

When the transformation is unknown, to complete our proposed method, an extra step of model selection is required. We estimate parameters involved in the Box-Cox transformation or the logarithmic transformation based on the simulated data under the same scenarios as in Tables 4.1-4.2. Results are reported in Tables 4.3-4.4, which show that the performance of the estimators proposed is satisfactory and similar to that revealed by Tables 4.1-4.2.

4.4.2 Application to the smoking cessation study

We revisit the interval-censored relapse time data from a smoking cessation study to illustrate the proposed method in this chapter. Since the relapse times of subjects residing in the same geographical region might be correlated, we are dealing with clustered interval-censored data with a cured subgroup. In Chapter 2, we have applied a semiparametric mixture cure model adopting the PH model with random effects for the conditional survival component. We shall make a comparison among several transformed models and evaluate the suitability of the model selected. In this analysis, we aim to investigate the effect of risk factors on the risk of smoking resumption based on this data set. The risk factors in this study include sex (Female=1, Male=0), intervention type SI/UC (Special Intervention=1,

Table 4.1: *Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with known transformations $G(t) = [(1+t)^\rho - 1]/\rho$.*

n	Parameter	True	Bias	SE	SSE	RMSE	CP
$\rho = 1$							
100	β	0.5	0.0048	0.1626	0.1604	0.1604	0.9560
	b_1	1.0	-0.0199	0.1950	0.2033	0.2043	0.9430
	b_2	0.5	0.0782	0.3052	0.3208	0.3301	0.9360
	σ^2	0.5^2	-0.0027	0.0172	0.0066	0.0071	0.9710
200	β	0.5	0.0052	0.1147	0.1149	0.1150	0.9470
	b_1	1.0	-0.0037	0.1487	0.1382	0.1383	0.9550
	b_2	0.5	0.0729	0.2315	0.2144	0.2264	0.9580
	σ^2	0.5^2	-0.0018	0.0122	0.0032	0.0037	0.9620
$\rho = 0.5$							
100	β	0.5	-0.0140	0.2084	0.2001	0.2005	0.9600
	b_1	1.0	-0.1207	0.2211	0.1987	0.2325	0.9350
	b_2	0.5	0.0763	0.3398	0.2996	0.3092	0.9730
	σ^2	0.5^2	-0.0062	0.0170	0.0087	0.0107	0.9700
200	β	0.5	-0.0073	0.1478	0.1370	0.1372	0.9550
	b_1	1.0	-0.1078	0.1562	0.1294	0.1685	0.9380
	b_2	0.5	0.0742	0.2400	0.2113	0.2240	0.9570
	σ^2	0.5^2	-0.0043	0.0121	0.0035	0.0056	0.9630

Bias, the mean of bias; SE, the estimated standard error; SSE, the sample standard error; RMSE, the root mean square error; and CP, the coverage probability of 95% confidence interval based on the normal approximations to parameter estimates.

Table 4.2: *Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with known transformations $G(t) = \log(1 + rt)/r$.*

n	Parameter	True	Bias	SE	SSE	RMSE	CP
$r = 0.5$							
100	β	0.5	0.0183	0.2218	0.2168	0.2176	0.9570
	b_1	1.0	0.0357	0.2003	0.2060	0.2091	0.9420
	b_2	0.5	0.0786	0.3113	0.3249	0.3343	0.9340
	σ^2	0.5^2	-0.0082	0.0168	0.0101	0.0130	0.9350
200	β	0.5	0.0053	0.1576	0.1520	0.1521	0.9500
	b_1	1.0	0.0606	0.1415	0.1492	0.1610	0.9380
	b_2	0.5	0.0563	0.2204	0.2283	0.2351	0.9450
	σ^2	0.5^2	-0.0064	0.0120	0.0060	0.0088	0.9480
$r = 1$							
100	β	0.5	-0.0096	0.2720	0.2721	0.2723	0.9520
	b_1	1.0	-0.1005	0.1921	0.1907	0.2155	0.9170
	b_2	0.5	0.0478	0.2953	0.3061	0.3098	0.9430
	σ^2	0.5^2	-0.0083	0.0168	0.0086	0.0120	0.9480
200	β	0.5	0.0016	0.1916	0.1790	0.1790	0.9480
	b_1	1.0	-0.0792	0.1359	0.1388	0.1598	0.9290
	b_2	0.5	0.0451	0.2101	0.2116	0.2164	0.9510
	σ^2	0.5^2	-0.0065	0.0120	0.0050	0.0082	0.9540

Table 4.3: *Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with estimated transformations $G(t) = [(1+t)^\rho - 1]/\rho$.*

n	Parameter	True	Bias	SE	SSE	RMSE	CP
$\rho = 1$							
100							$(\hat{\rho} = 0.820)$
	β	0.5	0.0220	0.1748	0.1753	0.1766	0.9430
	b_1	1.0	-0.0122	0.1954	0.2034	0.2035	0.9420
	b_2	0.5	0.0945	0.3056	0.3174	0.3312	0.9370
	σ^2	0.5^2	-0.0039	0.0171	0.0070	0.0081	0.9710
200							$(\hat{\rho} = 0.828)$
	β	0.5	0.0202	0.1243	0.1274	0.1290	0.9420
	b_1	1.0	0.0155	0.1378	0.1422	0.1431	0.9370
	b_2	0.5	0.0650	0.2159	0.2169	0.2264	0.9410
	σ^2	0.5^2	-0.0027	0.0122	0.0032	0.0041	0.9670
$\rho = 0.5$							
100							$(\hat{\rho} = 0.470)$
	β	0.5	0.0141	0.2109	0.2098	0.2103	0.9470
	b_1	1.0	-0.0952	0.1892	0.1974	0.2192	0.9250
	b_2	0.5	0.0848	0.2933	0.3106	0.3220	0.9350
	σ^2	0.5^2	-0.0057	0.0170	0.0071	0.0091	0.9680
200							$(\hat{\rho} = 0.476)$
	β	0.5	-0.0037	0.1494	0.1408	0.1408	0.9620
	b_1	1.0	-0.0888	0.1340	0.1366	0.1630	0.9300
	b_2	0.5	0.0790	0.2078	0.2163	0.2303	0.9370
	σ^2	0.5^2	-0.0047	0.0121	0.0040	0.0061	0.9620

Table 4.4: *Summary statistics of simulation studies in mixture cure model for clustered interval-censored survival data with estimated transformations $G(t) = \log(1 + rt)/r$.*

n	Parameter	True	Bias	SE	SSE	RMSE	CP
$r = 0.5$							
$(\hat{r} = 0.479)$							
100	β	0.5	0.0140	0.2202	0.2184	0.2188	0.9520
	b_1	1.0	0.0438	0.2000	0.2039	0.2085	0.9400
	b_2	0.5	0.0685	0.3117	0.3324	0.3394	0.9310
	σ^2	0.5^2	-0.0076	0.0169	0.0092	0.0120	0.9470
$(\hat{r} = 0.487)$							
200	β	0.5	0.0097	0.1561	0.1546	0.1549	0.9580
	b_1	1.0	0.0519	0.1414	0.1461	0.1550	0.9370
	b_2	0.5	0.0537	0.2204	0.2251	0.2314	0.9340
	σ^2	0.5^2	-0.0066	0.0120	0.0062	0.0091	0.9590
$r = 1$							
$(\hat{r} = 0.814)$							
100	β	0.5	-0.0128	0.2562	0.2638	0.2641	0.9430
	b_1	1.0	-0.1039	0.1924	0.2025	0.2276	0.9130
	b_2	0.5	0.0470	0.2953	0.3214	0.3249	0.9270
	σ^2	0.5^2	-0.0084	0.0168	0.0095	0.0127	0.9410
$(\hat{r} = 0.819)$							
200	β	0.5	-0.0112	0.1808	0.1740	0.1748	0.9590
	b_1	1.0	-0.1026	0.1358	0.1377	0.1717	0.9240
	b_2	0.5	0.0473	0.2089	0.2277	0.2325	0.9380
	σ^2	0.5^2	-0.0065	0.0120	0.0055	0.0085	0.9650

Usual Care=0), number of cigarettes smoked per day (Cigarettes) and duration (years) of smoking habit (Duration). All these factors are used in both the logistic regression and transformation survival model components. For the choice of transformation $G(\cdot)$, we consider all transformations from the Box-Cox family or the logarithm transformation families. To apply the proposed method, we need to determine the transformation parameter ρ or r . Using an ad hoc model selection method as those presented in Zeng et al. (2008) and Mao and Wang (2010), we calculate the Akaike's information criterion (AIC) $-2\log L + 2p$, where p represents the number of unknown parameters, for the transformation models over a grid of equally spaced points of (ρ, r) , and choose the model having the smallest AIC, equivalently, the largest likelihood function. The log likelihood function over difference values of transformation parameters is plotted in Figure 4.1, showing that the best choice of ρ in the Box-Cox transformation is 1.08. We use the closest integer $\rho = 1$, corresponding to the PH model. On the other hand, the value estimated for r in the logarithm transformation is approximately zero, which agrees with the choice of the well-known PH model for the conditional survival function. It is also shown in Figure 4.1 that the PH cure model gives the highest log-likelihood value among models with different choices of ρ and r under the transformations, and thus is preferable to the proportional odds cure model ($r = 1$) for this data set.

Table 4.5 reports the fitting result of the selected model with $\rho = 1$ in comparison with those of transformation models with $\rho = 0.5$ in the Box-Cox family and $r = 0.5, 1$ in the logarithm transformation family. It is evident from the estimates coefficient $\hat{b}_{SI/CU} = -0.67$ in the logistic regression component with $p = 0.047$ that subjects with special intervention have higher cure probability, i.e., more likely to quit smoking, than those

with usual care. However, the intervention type is nonsignificant in the survival model component, which coincides with the findings of Chapter 2 and Banerjee and Carlin (2004). The number of cigarettes smoked per day and the duration of smoking habit have significant effects in both the logistic regression and survival model components. Subjects with more cigarettes smoked per day are more likely to fail in their quit attempts. For those "uncured" subjects, long-term smokers have higher risk to resume smoking. Additionally, the variance component σ^2 in the survival model component is estimated to be 0.112 with a standard error of 0.011. Based on the asymptotic normality of estimated variance $\hat{\sigma}^2$ given by Theorem 4.3.5, the Wald statistic $\hat{\sigma}^2/SE(\hat{\sigma}^2) \approx 10.2$, indicating strong significance of the random cluster effects in the PH model for the conditional survival function. This is evident that there are nonnegligible correlation among relapse times in the same region.

To further illustrate the effect of factor SI/UC on the risk of smoking relapse, we examine the estimated survival curves for subjects resumed smoking and the overall study population under the transformation model selected. The survival probabilities, exhibiting in Figures 4.2 and 4.3, are estimated with the number of cigarettes smoked per day fixed at its median of 25 cigarettes and the duration of smoking habits at its median of 31 years, while factors sex and SI/UC vary for comparison purpose. Figure 4.2 demonstrates that different intervention types and gender do not make much difference in the survival probability estimated for smokers failed to quit smoking. It is evident from Figure 4.3 that in the overall study population smokers with the special intervention have obviously greater survival rate to relapse than those with the usual care. These results confirm the findings revealed in Table 4.5.

4.5 Concluding remarks

In this chapter, we provide a general class of semiparametric mixture cure transformation models for the analysis of clustered interval-censored survival data and develop the maximum likelihood estimation integrating the effect of latent random effects on the conditional survival function. Deriving asymptotic properties of the resulting estimators is a challenging task due to the complex data structure. There is a lack of the theoretical justification for estimators in mixture cure models with correlated interval-censored data. Our work fills this gap by establishing the consistency, asymptotic normality and convergence rate of the estimators using the modern empirical process techniques. We also have studied the model identifiability. In addition, given the asymptotic normality of the proposed estimator of the variance component, a hypothesis test concerning the significance of cluster effects can be consequently performed using the Wald test.

When the cure rates of subjects within the same cluster are considered to be correlated, it may be of interest to introduced another set of random effects v in the logistic regression to account for the within-cluster correlation. This requires a more complicated estimation procedure and can elicit additional problems, such as whether random effects ω and v are from the same family of distributions, and whether they are correlated. Future research along this direction is desirable. Asymptotic properties we induced for the estimators in this chapter are based on a known transformation. Further investigation on the property of the estimators under a transformation with unknown parameters also deserves a future research.

Figure 4.1: Maximum log-likelihood of the transformation mixture cure model over values of parameters ρ and r in the Box-Cox transformation and the logarithm transformation, respectively.

Table 4.5: *Parameter estimates for smoking relapse time data in mixture cure models with two types of transformations.*

Parameter	$G(t) = [(1+t)^\rho - 1]/\rho$			
	$\hat{\beta}_{(SE)}$	$\hat{b}_{(SE)}$	$\hat{\beta}_{(SE)}$	$\hat{b}_{(SE)}$
	$\rho = 1.5$		$\rho = 1.0$	
$\log L$	-94.540		-93.866	
Intercept	0.597 _(0.742)		0.708 _(0.742)	
Sex	0.505 _(0.350)	0.193 _(0.309)	0.580 _(0.404)	0.190 _(0.309)
SI/UC	0.213 _(0.376)	-0.653 _(0.340)	0.246 _(0.431)	-0.670* _(0.339)
Cigarettes	-0.046* _(0.016)	0.039* _(0.014)	-0.044* _(0.018)	0.036* _(0.014)
Duration	0.054* _(0.018)	-0.062* _(0.021)	0.058* _(0.020)	-0.063* _(0.021)
σ^2	0.116* _(0.011)		0.112* _(0.011)	

Parameter	$G(t) = \log(1+rt)/r$			
	$\hat{\beta}_{(SE)}$	$\hat{b}_{(SE)}$	$\hat{\beta}_{(SE)}$	$\hat{b}_{(SE)}$
	$r = 0.5$		$r = 1.0$	
$\log Lk$	-98.222		-104.522	
Intercept	0.904 _(0.744)		1.092 _(0.746)	
Sex	0.628 _(0.549)	0.216 _(0.308)	0.709 _(0.669)	0.226 _(0.307)
SI/UC	0.359 _(0.563)	-0.721* _(0.340)	0.431 _(0.672)	-0.764* _(0.340)
Cigarettes	-0.054* _(0.025)	0.039* _(0.014)	-0.056* _(0.030)	0.039* _(0.014)
Duration	0.076* _(0.028)	-0.068* _(0.021)	0.084* _(0.035)	-0.072* _(0.021)
σ^2	0.105* _(0.010)		0.102* _(0.010)	

* p -value < 0.05.

NOTE: The standard error estimates are given in parentheses. $\log L$, the value of log-likelihood.

Figure 4.2: *Estimated survival probabilities of subjects resumed smoking with the special intervention (solid) and usual care (dotted) by the proportional hazards cure model.*

Figure 4.3: *Estimated survival probabilities for the overall population of smokers with the special intervention (solid) and usual care (dotted) by the proportional hazards cure model.*

4.6 Appendices

Appendix A: Observed information matrix

Using a Newton-Raphson algorithm to maximize the complete data log-likelihood in the M-step in Section 4.2, we need to compute the following observed Fisher information matrix.

$$\begin{bmatrix} -\frac{\partial^2 l_1(\beta, \Lambda)}{\partial \beta \partial \beta^T} & -\frac{\partial^2 l_1(\beta, \Lambda)}{\partial \beta \partial \Lambda^T} & & \\ -\frac{\partial^2 l_1(\beta, \Lambda)}{\partial \Lambda \partial \beta^T} & -\frac{\partial^2 l_1(\beta, \Lambda)}{\partial \Lambda \partial \Lambda^T} & & \\ & & -\frac{\partial^2 l_2(b)}{\partial b \partial b^T} & \\ & & & -\frac{\partial^2 l_3(\sigma)}{\partial \sigma^2} \end{bmatrix}$$

In this appendix, we only derive the first and second derivatives of $l_1(\beta, \gamma)$ with respect to β and $\gamma = (\gamma_1, \dots, \gamma_Q)$ when replacing Λ by $\Lambda(t) = \sum_{q=1}^Q e^{\gamma_q} 1_{[t_q \leq t]}$ in $l_1(\beta, \Lambda)$. Derivatives of $l_2(b)$ with respect to b can be easily obtained because of the logistic model framework. So can the derivatives of $l_3(\sigma)$ with respect to σ .

For simplicity, we write $\eta_{ij} = \exp(\beta^T Z_{ij} + \omega_i)$, $S_{U_{ij}} = S(L_{ij}|Z_{ij}; \omega_i)$, $S_{R_{ij}} = S(R_{ij}|Z_{ij}; \omega_i)$, and $G_{L_{ij}} = G\{\Lambda(L_{ij})\eta_{ij}\}$, $G_{R_{ij}} = G\{\Lambda(R_{ij})\eta_{ij}\}$. $l_1(\beta, \gamma)$ can be rewritten as

$$l_1(\beta, \gamma) = \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \hat{E}_\omega[\log(S_{L_{ij}} - S_{R_{ij}})] - (1 - \delta_{ij}) \hat{E}_Y[Y_{ij}] \hat{E}_\omega[G_{L_{ij}}].$$

Accordingly, the first and second derivatives of $l_1(\beta, \gamma)$ with respect to β and γ are given as follows:

$$\begin{aligned} \frac{\partial l_1(\beta, \gamma)}{\partial \beta} &= \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G'_{R_{ij}} \Lambda(R_{ij}) - S_{L_{ij}} G'_{L_{ij}} \Lambda(L_{ij})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij} \right] Z_{ij} \\ &\quad - (1 - \delta_{ij}) \hat{E}_Y[Y_{ij}] \Lambda(L_{ij}) \hat{E}_\omega[G'_{L_{ij}} \eta_{ij}] Z_{ij}, \end{aligned}$$

$$\begin{aligned}
& \frac{\partial l_1(\beta, \gamma)}{\partial \gamma_q} \\
&= e^{\gamma_q} \sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}}(G'_{R_{ij}} - G'_{L_{ij}})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij} \right] - \hat{E}_Y[Y_{ij}] \hat{E}_\omega[G'_{L_{ij}} \eta_{ij}] \right\} 1_{[L_{ij} \geq t_q]} \\
& \quad + \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G'_{R_{ij}}}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij} \right] 1_{[L_{ij} < t_q \leq R_{ij}]},
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2 l_1(\beta, \gamma)}{\partial \beta \partial \beta^T} \\
&= \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G'_{R_{ij}} \Lambda(R_{ij}) - S_{L_{ij}} G'_{L_{ij}} \Lambda(L_{ij})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij} \right. \\
& \quad + \frac{S_{L_{ij}} (G'_{L_{ij}} \Lambda(L_{ij}))^2 - S_{R_{ij}} (G'_{R_{ij}} \Lambda(R_{ij}))^2}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \\
& \quad + \frac{S_{R_{ij}} G''_{R_{ij}} \Lambda(R_{ij})^2 - S_{L_{ij}} G''_{L_{ij}} \Lambda(L_{ij})^2}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \\
& \quad \left. - \frac{(S_{R_{ij}} G'_{R_{ij}} \Lambda(R_{ij}) - S_{L_{ij}} G'_{L_{ij}} \Lambda(L_{ij}))^2}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right] Z_{ij}^2 \\
& \quad - (1 - \delta_{ij}) \hat{E}_Y[Y_{ij}] \{ \Lambda(L_{ij}) \hat{E}_\omega[G'_{L_{ij}} \eta_{ij}] + \Lambda(L_{ij})^2 \hat{E}_\omega[G''_{L_{ij}} \eta_{ij}^2] \} Z_{ij}^2,
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2 l_1(\beta, \gamma)}{\partial \beta \partial \gamma_q} = \frac{\partial^2 l_1(\beta, \gamma)}{\partial \gamma_q \partial \beta^T} \\
&= e^{\gamma_q} \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G'_{R_{ij}}}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij} - \frac{S_{L_{ij}} S_{R_{ij}} G'_{R_{ij}} (G'_{R_{ij}} \Lambda(R_{ij}) - G'_{L_{ij}} \Lambda(L_{ij}))}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right. \\
& \quad \left. + \frac{S_{R_{ij}} G''_{R_{ij}} \Lambda(R_{ij})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \right] Z_{ij} 1_{[L_{ij} < t_q \leq R_{ij}]} \\
& \quad + \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} (G'_{R_{ij}} - G'_{L_{ij}})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij} \right. \\
& \quad \left. - \frac{S_{L_{ij}} S_{R_{ij}} (G'_{R_{ij}} - G'_{L_{ij}}) (G'_{R_{ij}} \Lambda(R_{ij}) - G'_{L_{ij}} \Lambda(L_{ij}))}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right. \\
& \quad \left. + \frac{S_{R_{ij}} (G''_{R_{ij}} \Lambda(R_{ij}) - G''_{L_{ij}} \Lambda(L_{ij}))}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \right] Z_{ij} 1_{[L_{ij} \geq t_q]} \\
& \quad - \hat{E}_Y[Y_{ij}] \hat{E}_\omega[G'_{L_{ij}} \eta_{ij} + G''_{L_{ij}} \Lambda(L_{ij}) \eta_{ij}^2] Z_{ij} 1_{[L_{ij} \geq t_q]}.
\end{aligned}$$

and $\frac{\partial^2 l_1(\beta, \gamma)}{\partial \gamma \partial \gamma^T}$ is a $Q \times Q$ symmetric matrix whose diagonal entries are

$$\begin{aligned} & \frac{\partial^2 l_1(\beta, \gamma)}{\partial \gamma_q \partial \gamma_q} \\ &= \frac{\partial l_2}{\partial \gamma_q} + e^{2\gamma_q} \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G''_{R_{ij}}}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 - \frac{S_{L_{ij}} S_{R_{ij}} G'^2_{R_{ij}}}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right] 1_{[L_{ij} < t_q \leq R_{ij}]} \\ & \quad + \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} (G''_{R_{ij}} - G''_{L_{ij}})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \right. \\ & \quad \quad \left. - \frac{S_{L_{ij}} S_{R_{ij}} (G'_{R_{ij}} - G'_{L_{ij}})^2}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right] 1_{[L_{ij} \geq t_q]} \\ & \quad - \hat{E}_Y[Y_{ij}] \hat{E}_\omega[G''_{L_{ij}} \eta_{ij}^2] 1_{[L_{ij} \geq t_q]}, \end{aligned}$$

and off-diagonal entries are

$$\begin{aligned} & \frac{\partial^2 l_1(\beta, \gamma)}{\partial \gamma_q \partial \gamma_k} \\ &= e^{\gamma_q + \gamma_k} \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G''_{R_{ij}}}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 - \frac{S_{L_{ij}} S_{R_{ij}} G'^2_{R_{ij}}}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right] 1_{[L_{ij} < t_k < t_q \leq R_{ij}]} \\ & \quad + \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} G''_{R_{ij}}}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \right. \\ & \quad \quad \left. - \frac{S_{L_{ij}} S_{R_{ij}} G'_{R_{ij}} (G'_{R_{ij}} - G'_{L_{ij}})}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right] 1_{[t_k \leq L_{ij} < t_q \leq R_{ij}]} \\ & \quad + \delta_{ij} \hat{E}_\omega \left[\frac{S_{R_{ij}} (G''_{R_{ij}} - G''_{L_{ij}})}{S_{L_{ij}} - S_{R_{ij}}} \eta_{ij}^2 \right. \\ & \quad \quad \left. - \frac{S_{L_{ij}} S_{R_{ij}} (G'_{R_{ij}} - G'_{L_{ij}})^2}{(S_{L_{ij}} - S_{R_{ij}})^2} \eta_{ij}^2 \right] 1_{[L_{ij} \geq t_q]} \\ & \quad - \hat{E}_Y[Y_{ij}] \hat{E}_\omega[G''_{L_{ij}} \eta_{ij}^2] 1_{[L_{ij} \geq t_q]}, \end{aligned}$$

for $k < q, q = 1, \dots, Q$.

Appendix B: Proof of Theorem 4.3.2

To prove Theorem 4.3.2, we need to specify h in the score operator of Λ , $\dot{l}_\Lambda(h)$. In other words, we find out h^* such that equations (4.3.6)-(4.3.8) holds. Using the result of Groeneboom and Wellner (1992) on Page 8-9, we know that

$$\begin{aligned} \dot{l}_\Lambda^T \dot{l}_\Lambda(h)(t) &= E[\dot{l}_\Lambda(h)|T=t] = E_Z E[\dot{l}_\Lambda(h)|T=t, Z=Z], \\ \dot{l}_\Lambda^T \dot{l}_\theta(t) &= E[\dot{l}_\theta|T=t] = E_Z E[\dot{l}_\theta|T=t, Z=Z]. \end{aligned}$$

Then, to solve equations (4.3.6)-(4.3.8) is to find $h^* = (h_\beta^{*T}, h_b^{*T}, h_\sigma^*)^T$ satisfying the following equations:

$$E_Z E[\dot{l}_\Lambda(h_\beta^{*T})|T=t, Z=Z] = E_Z E[\dot{l}_\beta|T=t, Z=Z], \quad (\text{B.1})$$

$$E_Z E[\dot{l}_\Lambda(h_b^{*T})|T=t, Z=Z] = E_Z E[\dot{l}_b|T=t, Z=Z], \quad (\text{B.2})$$

$$E_Z E[\dot{l}_\Lambda(h_\sigma^*)|T=t, Z=Z] = E_Z E[\dot{l}_\sigma|T=t, Z=Z]. \quad (\text{B.3})$$

Since $Z_1, \dots, Z_N$ are identically from the same cluster, they have the

same expectation. Then we can define the following functions.

$$\begin{aligned} & \mathbf{A}_1(U, V|Z; \omega) \\ &= \frac{N[1 - S(U|Z; \omega)]^{N-1} S(U|Z; \omega) G' \{ \Lambda(U) e^{\beta^T Z + \omega} \} e^{\omega}}{\int [1 - S(U|Z; \omega)]^N \phi(\omega; \sigma) d\omega}, \end{aligned}$$

$$\begin{aligned} & \mathbf{A}_2(U, V|Z; \omega) \\ &= \frac{N[S(U|Z; \omega) - S(V|Z; \omega)]^{N-1} S(U|Z; \omega) G' \{ \Lambda(U) e^{\beta^T Z + \omega} \} e^{\omega}}{\int [S(U|Z; \omega) - S(V|Z; \omega)]^N \phi(\omega; \sigma) d\omega}, \end{aligned}$$

$$\begin{aligned} & \mathbf{A}_3(U, V|Z; \omega) \\ &= \frac{N[S(U|Z; \omega) - S(V|Z; \omega)]^{N-1} S(V|Z; \omega) G' \{ \Lambda(V) e^{\beta^T Z + \omega} \} e^{\omega}}{\int [S(U|Z; \omega) - S(V|Z; \omega)]^N \phi(\omega; \sigma) d\omega}, \end{aligned}$$

$$\begin{aligned} & \mathbf{A}_4(U, V|Z; \omega) \\ &= \frac{N[1 - \pi(Z) + \pi(Z) S(V|Z; \omega)]^{N-1} \pi(Z) S(V|Z; \omega) G' \{ \Lambda(V) e^{\beta^T Z + \omega} \} e^{\omega}}{\int [1 - \pi(Z) + \pi(Z) S(V|Z; \omega)]^N \phi(\omega; \sigma) d\omega}, \end{aligned}$$

$$\begin{aligned} & \mathbf{A}_5(U, V|Z; \omega) \\ &= \frac{N[1 - \pi(Z) + \pi(Z) S(V|Z; \omega)]^{N-1} \pi(Z) S(V|Z; \omega)}{\int [1 - \pi(Z) + \pi(Z) S(V|Z; \omega)]^N \phi(\omega; \sigma) d\omega}. \end{aligned}$$

By Conditions 3-4, $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4$ and $\mathbf{A}_5$ are positive functions of (U, V) given Z and ω . We calculate each component in equations (B.1)-(B.3) as

follows. Corresponding to Λ ,

$$\begin{aligned}
& E[\dot{l}_\Lambda(h)|T = t, Z = Z] \\
&= N \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} h(U) e^{\beta^T Z} \int_{\omega} \mathbf{A}_1 \psi(U, V|Z; \omega) \phi(\omega; \sigma) d\omega dV dU \\
&\quad - \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} e^{\beta^T Z} \int_{\omega} (\mathbf{A}_2 h(U) - \mathbf{A}_3 h(V)) 1_{[V-U \geq \varsigma]} \psi(U, V|Z; \omega) \\
&\quad \quad \cdot \phi(\omega; \sigma) d\omega dV dU \\
&\quad - \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t e^{\beta^T Z} \int_{\omega} \mathbf{A}_4 h(V) \psi(U, V|Z; \omega) \phi(\omega; \sigma) d\omega dV dU.
\end{aligned}$$

Let

$$\begin{aligned}
\mathbf{B}_1(U, V) &= E_Z[e^{\beta^T Z} \int_{\omega} \mathbf{A}_1(U, V|Z; \omega) \psi(U, V|Z; \omega) \phi(\omega; \sigma) d\omega], \\
\mathbf{B}_2(U, V) &= E_Z[e^{\beta^T Z} \int_{\omega} \mathbf{A}_2(U, V|Z; \omega) \psi(U, V|Z; \omega) \phi(\omega; \sigma) d\omega], \\
\mathbf{B}_3(U, V) &= E_Z[e^{\beta^T Z} \int_{\omega} \mathbf{A}_3(U, V|Z; \omega) \psi(U, V|Z; \omega) \phi(\omega; \sigma) d\omega], \\
\mathbf{B}_4(U, V) &= E_Z[e^{\beta^T Z} \int_{\omega} \mathbf{A}_4(U, V|Z; \omega) \psi(U, V|Z; \omega) \phi(\omega; \sigma) d\omega],
\end{aligned} \tag{B.4}$$

then we obtain

$$\begin{aligned}
\mathbf{L}_h(t) &\equiv \dot{l}_\Lambda^T \dot{l}_\Lambda(h)(t) \\
&= E_Z E[\dot{l}_\Lambda(h)|T = t, Z = Z] \\
&= \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} \mathbf{B}_1(U, V) h(U) dV dU \\
&\quad - \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} (\mathbf{B}_2(U, V) h(U) - \mathbf{B}_3(U, V) h(V)) 1_{[V-U \geq \varsigma]} dV dU \\
&\quad - \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t \mathbf{B}_4(U, V) h(V) dV dU,
\end{aligned}$$

where $\mathbf{B}$'s are positive functions of (U, V) .

Similarly, corresponding to β , we have

$$\begin{aligned}
 & E[\dot{l}_\beta | T = t, Z = Z] \\
 = & N \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} \Lambda(U) Z e^{\beta^T Z} \int_{\omega} \mathbf{A}_1 \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega dV dU \\
 & - \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} Z e^{\beta^T Z} \int_{\omega} (\mathbf{A}_2 \Lambda(U) - \mathbf{A}_3 \Lambda(V)) 1_{[V-U \geq \varsigma]} \psi(U, V | Z; \omega) \\
 & \quad \cdot \phi(\omega; \sigma) d\omega dV dU \\
 & - \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t Z e^{\beta^T Z} \int_{\omega} \mathbf{A}_4 \Lambda(V) \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega dV dU.
 \end{aligned}$$

Let

$$\begin{aligned}
 \mathbf{C}_1(U, V) &= E_Z [Z e^{\beta^T Z} \int_{\omega} \mathbf{A}_1(U, V | Z; \omega) \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega], \\
 \mathbf{C}_2(U, V) &= E_Z [Z e^{\beta^T Z} \int_{\omega} \mathbf{A}_2(U, V | Z; \omega) \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega], \\
 \mathbf{C}_3(U, V) &= E_Z [Z e^{\beta^T Z} \int_{\omega} \mathbf{A}_3(U, V | Z; \omega) \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega], \\
 \mathbf{C}_4(U, V) &= E_Z [Z e^{\beta^T Z} \int_{\omega} \mathbf{A}_4(U, V | Z; \omega) \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega],
 \end{aligned}$$

then

$$\begin{aligned}
 \mathbf{R}_\beta(t) &\equiv \dot{l}_\Lambda^T \dot{l}_\beta(t) \\
 &= E_Z E[\dot{l}_\beta | T = t, Z = Z] \\
 &= \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} \mathbf{C}_1(U, V) \Lambda(U) dV dU \\
 &\quad - \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} (\mathbf{C}_2(U, V) \Lambda(U) - \mathbf{C}_3(U, V) \Lambda(V)) 1_{[V-U \geq \varsigma]} dV dU \\
 &\quad - \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t \mathbf{C}_1(U, V) \Lambda(V) dV dU.
 \end{aligned}$$

Next, we have the entries corresponding to b ,

$$\begin{aligned}
& E[\dot{l}_b | T = t, Z = Z] \\
&= -W\pi(Z) + \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} W \int_{\omega} \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega dV dU \\
&+ \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} W 1_{[V-U \geq \varsigma]} \int_{\omega} \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega dV dU \\
&+ \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t W \int_{\omega} \mathbf{A}_{5j} \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega dV dU.
\end{aligned}$$

Now let

$$\begin{aligned}
\mathbf{D}_1 &= E_Z[W\pi(Z)], \\
\mathbf{D}_2(U, V) &= E_Z[W \int_{\omega} \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega], \\
\mathbf{D}_3(U, V) &= E_Z[W \int_{\omega} \mathbf{A}_5(U, V | Z; \omega) \psi(U, V | Z; \omega) \phi(\omega; \sigma) d\omega],
\end{aligned}$$

then

$$\begin{aligned}
\mathbf{R}_b(t) &\equiv \dot{l}_\Lambda^T \dot{l}_b(t) \\
&= -\mathbf{D}_1 + \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} \mathbf{D}_2(U, V) dV dU \\
&+ \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} \mathbf{D}_2(U, V) 1_{[V-U \geq \varsigma]} dV dU + \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t \mathbf{D}_3(U, V) dV dU.
\end{aligned}$$

Last we calculate the entry associated with σ ,

$$\begin{aligned}
E[\dot{l}_\sigma | T = t, Z = Z] &= \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} \int_{\omega} \mathbf{I}_1 \psi(U, V | Z; \omega) d\omega dV dU \\
&+ \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} \int_{\omega} \mathbf{I}_2 1_{[V-U \geq \varsigma]} \psi(U, V | Z; \omega) d\omega dV dU \\
&+ \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t \int_{\omega} \mathbf{I}_3 \psi(U, V | Z; \omega) d\omega dV dU,
\end{aligned}$$

where

$$\mathbf{I}_1(U, V|Z; \omega) = \frac{[1 - S(U|Z; \omega)]^N \frac{\partial \phi(\omega; \sigma)}{\partial \sigma}}{\int [1 - S(U_k|Z_k; \omega)]^N \phi(\omega; \sigma) d\omega},$$

$$\mathbf{I}_2(U, V|Z; \omega) = \frac{[S(U|Z; \omega) - S(V|Z; \omega)]^N \frac{\partial \phi(\omega; \sigma)}{\partial \sigma}}{\int [S(U|Z; \omega) - S(V|Z; \omega)]^N \phi(\omega; \sigma) d\omega},$$

$$\mathbf{I}_3(U, V|Z; \omega) = \frac{[1 - \pi(Z) + \pi(Z)S(V|Z; \omega)]^N \frac{\partial \phi(\omega; \sigma)}{\partial \sigma}}{\int [1 - \pi(Z) + \pi(Z)S(V|Z; \omega)]^N \phi(\omega; \sigma) d\omega},$$

We define

$$\mathbf{J}_k(U, V) = E_Z \left[\int_{\omega} \mathbf{I}_k(U, V|Z; \omega) \psi(U, V|Z; \omega) d\omega \right] \quad \text{for } k = 1, 2, 3.$$

then

$$\begin{aligned} \mathbf{R}_{\sigma}(t) &\equiv \dot{l}_{\Lambda}^T \dot{l}_{\sigma}(t) \\ &= \int_{U=t}^{\tau_1} \int_{V=U+\varsigma}^{\tau_1} \mathbf{J}_1(U, V) dV dU + \int_{U=\tau_0}^t \int_{V=t}^{\tau_1} \mathbf{J}_2(U, V) 1_{[V-U \geq \varsigma]} dV dU \\ &\quad + \int_{U=\tau_0}^t \int_{V=U+\varsigma}^t \mathbf{J}_3(U, V) dV dU \end{aligned}$$

By Conditions 1-4, $\mathbf{r}_{\beta}$, $\mathbf{r}_b$, and $\mathbf{r}_{\sigma}$ have bound derivatives on $[\tau_0, \tau_1]$.

Thus, equations (B.1)-(B.3) reduce to

$$\mathbf{L}'_h(t) = \mathbf{R}'_{\beta}(t),$$

$$\mathbf{L}'_h(t) = \mathbf{R}'_b(t),$$

$$\mathbf{L}'_h(t) = \mathbf{R}'_{\sigma}(t),$$

where the derivative of $\mathbf{L}_h(t)$ is

$$\mathbf{L}'_h(t) = \frac{\partial}{\partial t} \mathbf{L}_h(t) = -\mu(t)h(t) + \int_{\tau_0}^{t-\varsigma} h(u) \mathbf{B}_2(u, t) du + \int_{t+\varsigma}^{\tau_1} h(v) \mathbf{B}_3(t, v) dv$$

with

$$\mu(t) = \int_{t+\varsigma}^{\tau_1} \mathbf{B}_1(t, v) dv + \int_{t+\varsigma}^{\tau_1} \mathbf{B}_2(t, v) dv + \int_{\tau_0}^{t-\varsigma} \mathbf{B}_3(u, t) du + \int_{\tau_0}^{t-\varsigma} \mathbf{B}_4(u, t) du,$$

Using Condition 2, $\inf_{\tau_0 \leq t \leq \tau_1} \mu(t) > 0$. The derivative of $\mathbf{R}_\beta(t)$ is

$$\begin{aligned} -\mathbf{r}_\beta(t) &\equiv \mathbf{R}'_\beta(t) = \frac{\partial}{\partial t} \mathbf{R}_\beta(t) \\ &= -\nu(t)\Lambda(t) + \int_{\tau_0}^{t-\varsigma} \Lambda(u)\mathbf{C}_2(u, t)du + \int_{t+\varsigma}^{\tau_1} \Lambda(v)\mathbf{C}_3(t, v)dv \end{aligned}$$

with

$$\nu(t) = \int_{t+\varsigma}^{\tau_1} \mathbf{C}_1(t, v)dv + \int_{t+\varsigma}^{\tau_1} \mathbf{C}_2(t, v)dv + \int_{\tau_0}^{t-\varsigma} \mathbf{C}_3(u, t)du + \int_{\tau_0}^{t-\varsigma} \mathbf{C}_4(u, t)du,$$

the derivative of $\mathbf{R}_b(t)$ is

$$-\mathbf{r}_b(t) \equiv \mathbf{R}'_b(t) = \frac{\partial}{\partial t} \mathbf{R}_b(t) = \int_{\tau_0}^{t-\varsigma} \mathbf{D}_2(u, t) - \mathbf{D}_3(u, t)du,$$

and the derivative of $\mathbf{R}_\sigma(t)$ is

$$\begin{aligned} -\mathbf{r}_\sigma(t) &\equiv \mathbf{R}'_\sigma(t) = \frac{\partial}{\partial t} \mathbf{R}_\sigma(t) \\ &= \int_{t+\varsigma}^{\tau_1} \mathbf{J}_2(t, v) - \mathbf{J}_1(t, v)dv + \int_{\tau_0}^{t-\varsigma} \mathbf{J}_3(u, t) - \mathbf{J}_2(u, t)du. \end{aligned}$$

Therefore, we obtain the following equations,

$$-\mu(t)h(t) + \int_{\tau_0}^{t-\varsigma} h(u)\mathbf{B}_2(u, t)du + \int_{t+\varsigma}^{\tau_1} h(v)\mathbf{B}_3(t, v)dv = -\mathbf{r}_\beta(t), \quad (\text{B.5})$$

$$-\mu(t)h(t) + \int_{\tau_0}^{t-\varsigma} h(u)\mathbf{B}_2(u, t)du + \int_{t+\varsigma}^{\tau_1} h(v)\mathbf{B}_3(t, v)dv = -\mathbf{r}_b(t), \quad (\text{B.6})$$

$$-\mu(t)h(t) + \int_{\tau_0}^{t-\varsigma} h(u)\mathbf{B}_2(u, t)du + \int_{t+\varsigma}^{\tau_1} h(v)\mathbf{B}_3(t, v)dv = -\mathbf{r}_\sigma(t). \quad (\text{B.7})$$

Let

$$K(t, x) = [\mathbf{B}_2(x, t)1_{[\tau_0 \leq x \leq t-\varsigma]} + \mathbf{B}_3(t, x)1_{[t+\varsigma \leq x \leq \tau_1]}] / \mu(t). \quad (\text{B.8})$$

From equations (B.5)-(B.7), we obtain a group of Fredholm integral equations of the second kind,

$$h_\beta(t) - \int K(t, x)h_\beta(x)dx = \mathbf{r}_\beta(t)/\mu(t),$$

$$h_b(t) - \int K(t, x)h_b(x)dx = \mathbf{r}_b(t)/\mu(t),$$

$$h_\sigma(t) - \int K(t, x)h_\sigma(x)dx = \mathbf{r}_\sigma(t)/\mu(t).$$

By the definitions of $\mathbf{B}$'s in (B.4) and $K(t, x)$ in (B.8) and Condition 2, we have that $h(t) \equiv 0$ on $[\tau_0, \tau_1]$ if $h(t) - \int K(t, x)h(x)dx = 0$. Because $K(t, x)$ is a bounded kernel, by the classical results on Fredholm integral equations, there exists a resolvent $K^*(t, x)$, completely determined by K , such that the solution h^* can be given by

$$h_\beta^*(t) = \mathbf{r}_\beta(t)/\mu(t) + \int K^*(t, x)\mathbf{r}_\beta(x)/\mu(x)dx \quad (\text{B.9})$$

$$h_b^*(t) = \mathbf{r}_b(t)/\mu(t) + \int K^*(t, x)\mathbf{r}_b(x)/\mu(x)dx \quad (\text{B.10})$$

$$h_\sigma^*(t) = \mathbf{r}_\sigma(t)/\mu(t) + \int K^*(t, x)\mathbf{r}_\sigma(x)/\mu(x)dx \quad (\text{B.11})$$

Note that h_β^* , h_b^* and h_σ^* are differentiable as they are bounded on $[\tau_0, \tau_1]$ by (B.5)-(B.7), and continuous from (B.9)-(B.11). Furthermore, we show the derivatives of h_β^* , h_b^* and h_σ^* are bounded. Since $\mu(t)$ is bounded away from zero, the partial derivative of $\mathbf{B}_2(x, t)$ with respect to t is bounded on $[\tau_0, t - \varsigma]$, and the partial derivative of $\mathbf{B}_3(t, x)$ with respect to t is bounded on $[t + \varsigma, \tau_1]$, by Conditions 1-6 $\mathbf{r}_\beta$, $\mathbf{r}_b$ and $\mathbf{r}_\sigma$ have bounded derivatives on $[\tau_0, \tau_1]$. It follows that the derivatives of h_β^* , h_b^* and h_σ^* are bounded. Hence Theorem 4.3.2 is shown.

Appendix C: Proof of Theorem 4.3.4

We shall use the results of Theorem 3.4.1 and Lemma 3.4.2 in Van Der Vaart and Wellner (1996) to derive Theorem 4.3.4. First we introduce the following lemma under the proposed mixture model framework.

Recall that log-likelihood $l(\theta, \Lambda|X)$ is defined in 4.3.5 for N independent and identically distributed samples $X = (X_1, \dots, X_N)$ within a common

cluster. Define $L_2(P) = \{f : \int f^2 dP < \infty\}$, denote by the notation $\|\cdot\|_2$ the $L_2(P)$ -norm, i.e., $\|f\|_2 = (\int f^2 dP)^{1/2}$.

Lemma 4.6.1. *Define the uniformly bounded class of functions*

$\mathcal{L} = \{l(\theta, \Lambda|X) : (\theta, \Lambda) \in \Theta \times \Gamma^*\}$ *with*

$$\Gamma^* = \{\Lambda : \Lambda \text{ is an increasing function and bounded on } [\tau_0, \tau_1]$$

$$\text{with } \Lambda(0) = 0, \text{ and } \Lambda(V) - \Lambda(U) \geq \epsilon\}.$$

The $L_2(P)$ bracketing number of $\mathcal{L}$ is $N_{[]}(\epsilon, \mathcal{L}, L_2(P)) = O(\epsilon^{-2d-2}e^{1/\epsilon})$.

Proof. Conditions 2-3 imply that there exists an $\epsilon > 0$ satisfying $\Lambda_0(V) - \Lambda_0(U) \geq \Lambda_0(U + c) - \Lambda_0(U) > 2\epsilon$ in the support of (U, V) . From Theorem 4.3.3, we have $\hat{\Lambda}_n(V) - \hat{\Lambda}_n(U) \geq \epsilon$ for n large enough. Hence we can drop our attention to the likelihood associated with Λ in the class of functions Γ^* .

First, we consider the likelihood contributed by the survival model of uncured subjects. It is known that for the class of functions Γ^* , its ϵ -bracketing number is of order of $m_0 = N_{[]}(\epsilon, \Gamma^*, L_2(P)) = O(e^{1/\epsilon})$. This means there exists a set of functions, $\Lambda_1^{*Lw}, \Lambda_1^{*Up}, \dots, \Lambda_{m_0}^{*Lw}, \Lambda_{m_0}^{*Up}$ such that for each $\Lambda \in \Gamma^*$, $\Lambda_i^{*Lw} \leq \Lambda \leq \Lambda_i^{*Up}$ for some i and $\|\Lambda_i^{*Lw} - \Lambda_i^{*Up}\|_2 \leq \epsilon$. We can certainly use the following set of functions: $\{\Lambda_i^{*Lw} - \epsilon, \Lambda_i^{*Up} + \epsilon, i = 1, \dots, m_0\}$. Let $\Lambda_i^{Lw} = \Lambda_i^{*Lw} - \epsilon$, and $\Lambda_i^{Up} = \Lambda_i^{*Up} + \epsilon$. For any $\Lambda \in \Gamma^*$, we have $\Lambda_i^{Lw} + \epsilon = \Lambda_i^{*Lw} \leq \Lambda \leq \Lambda_i^{*Up} = \Lambda_i^{Up} - \epsilon$ and $\|\Lambda_i^{Lw} - \Lambda_i^{Up}\|_2 \leq 3\epsilon$. Since Γ^* is uniformly bounded away from zero, we choose a small ϵ such that all the bracketing functions stay away from zero.

Similarly to the proof of the Lemma 3.1 in Huang (1996), it is shown that we can choose m_1 points $\beta_1, \dots, \beta_{m_1}$ in $B(\beta_0, \eta) = \{\beta : \|\beta - \beta_0\| \leq \eta\}$, which is a ball with center β_0 and radius η defined as the set of all points

within Euclidean distance η from β_0 , such that for any $(\beta, \Lambda) \in B(\beta_0, \eta) \times \Gamma^*$,

$$e^{\beta_j^T Z} \Lambda_i^{Lw}(t) \leq e^{\beta^T Z} \Lambda(t) \leq e^{\beta_j^T Z} \Lambda_i^{Up}(t) \quad (C.1)$$

for $i = 1, \dots, m_0$ and some $1 \leq j \leq m_1$. And the minimum value of m_1 can be in the order of $O(\varepsilon^{-d})$.

Next, consider the likelihood contributed by the cured subgroup. For a fixed ω , define $S(t|\omega) = \exp(-e^{\beta^T Z + \omega} \Lambda(t))$, $S_{ij}^{Lw}(t|\omega) = \exp(-e^{\beta_j^T Z + \omega} \Lambda_i^{Lw}(t))$, and

$S_{ij}^{Up}(t|\omega) = \exp(-e^{\beta_j^T Z + \omega} \Lambda_i^{Up}(t))$. From (C.1), we have

$$S_{ij}^{Up}(t|\omega) \leq S(t|\omega) \leq S_{ij}^{Lw}(t|\omega).$$

We want to show that we can choose m_2 points $b_1, \dots, b_{m_2}$ in $B((\beta_0^T, b_0^T)^T, \eta)$ with the fixed m_1 points $\beta_1, \dots, \beta_{m_1}$ such that for any $(\beta, b, \Lambda) \in B((\beta_0^T, b_0^T)^T, \eta) \times \Gamma^*$,

$$\pi_k[1 - S_{ij}^{Lw}(t|\omega)] \leq \pi[1 - S(t|\omega)] \leq \pi_k[1 - S_{ij}^{Up}(t|\omega)], \quad (C.2)$$

for some $1 \leq k \leq m_2$, where $\pi_k = \pi(Z) = \frac{\exp(b_k^T W)}{1 + \exp(b_k^T W)}$ with $W = (1, Z^T)^T$. π is bounded on $(0, 1)$. Under Condition 1, we can find a constant $0 < C_1 < \infty$ such that $e^{\beta^T Z} \leq C_1$ for all Z and $\beta \in B(\beta_0, \eta)$. It follows that

$$\begin{aligned} & \pi[1 - S(t|\omega)] - \pi_k[1 - S_{ij}^{Lw}(t|\omega)] \\ &= \pi[1 - S(t|\omega)] - \pi[1 - S_{ij}^{Lw}(t|\omega)] + \pi[1 - S_{ij}^{Lw}(t|\omega)] - \pi_k[1 - S_{ij}^{Lw}(t|\omega)] \\ &= \pi[S_{ij}^{Lw}(t|\omega) - S(t|\omega)] - |\pi - \pi_k|[1 - S_{ij}^{Lw}(t|\omega)] \\ &\geq C_2 e^\omega [e^{\beta^T Z} \Lambda(t) - e^{\beta_j^T Z} \Lambda_i^{Lw}(t)] - C_3 e^\omega |b - b_k| \\ &\geq (C_2 \varepsilon - C_3 |b - b_k|) e^\omega, \end{aligned}$$

and

$$\pi_k[1 - S_{ij}^{Up}(t|\omega)] - \pi[1 - S(t|\omega)] \geq (C_2 \varepsilon - C_4 |b - b_k|) e^\omega.$$

where C_2, C_3, C_4 are positive and $1 - \exp(-ce^\omega) \approx ce^\omega$ for some constant c . If we choose $b_1, \dots, b_{m_2}$ with the known $\beta_1, \dots, \beta_{m_1}$ such that $\|b - b_k\| \leq \min\{(C_2 + 1)\varepsilon/C_3, (C_2 + 1)\varepsilon/C_4\}$ for any $(\beta, b, \Lambda) \in B((\beta_0^T, b_0^T)^T, \eta)$ and some $1 \leq k \leq m_2$, then (C.2) holds. It is known, see van de Geer (1993), that the minimum value of m_2 can be in the order of $O(\varepsilon^{-d-1})$.

Note that (C.2) implies that

$$1 - \pi_k + \pi_k S_{ij}^{Up}(t|\omega) \leq 1 - \pi + \pi S(t|\omega) \leq 1 - \pi_k + \pi_k S_{ij}^{Lw}(t|\omega),$$

and

$$\begin{aligned} \pi[S(U|\omega) - S(V|\omega)] &= [1 - \pi + \pi S(U|\omega)] - [1 - \pi + \pi S(V|\omega)] \\ &\leq [1 - \pi_k + \pi_k S_{ij}^{Lw}(U|\omega)] - [1 - \pi_k + \pi_k S_{ij}^{Up}(V|\omega)] \\ &= \pi_k [S_{ij}^{Lw}(U|\omega) - S_{ij}^{Up}(V|\omega)]. \end{aligned}$$

In addition,

$$\begin{aligned} \pi[S(U|\omega) - S(V|\omega)] &\geq [1 - \pi_k + \pi_k S_{ij}^{Up}(U|\omega)] - [1 - \pi_k + \pi_k S_{ij}^{Lw}(V|\omega)] \\ &= \pi_k \cdot \max\{S_{ij}^{Up}(U|\omega) - S_{ij}^{Lw}(V|\omega), S_{ij}^*(\omega)\}, \end{aligned}$$

where $S_{ij}^*(\omega) = \min\{S_0^*(\omega), S_{ij}^{Lw}(U|\omega) - S_{ij}^{Up}(V|\omega)\}$, which is positive for any ω , and $S_0^*(\omega) = \inf\{S(U|\omega) - S(V|\omega) : (\beta, \Lambda) \in \Theta_d \times \Gamma^*\}$ in the bounded support of Z .

Now, by defining

$$L(\omega) = [\pi(1 - S(U|\omega))]^{\delta^1} [\pi[S(U|\omega) - S(V|\omega)]]^{\delta^2} [1 - \pi + \pi S(V|\omega)]^{1-\delta^1-\delta^2},$$

$$\begin{aligned} L_{ijk}^{Up}(\omega) &= [\pi_k(1 - S_{ij}^{Up}(U|\omega))]^{\delta^1} [\pi_k[S_{ij}^{Lw}(U|\omega) - S_{ij}^{Up}(V|\omega)]]^{\delta^2} \\ &\quad [1 - \pi_k + \pi_k S_{ij}^{Lw}(V|\omega)]^{1-\delta^1-\delta^2}, \end{aligned}$$

$$\begin{aligned} L_{ijk}^{Lw}(\omega) &= [\pi_k(1 - S_{ij}^{Lw}(U|\omega))]^{\delta^1} [\pi_k \cdot \max\{S_{ij}^{Up}(U|\omega) - S_{ij}^{Lw}(V|\omega), S_{ij}^*(\omega)\}]^{\delta^2} \\ &\quad [1 - \pi_k + \pi_k S_{ij}^{Up}(V|\omega)]^{1-\delta^1-\delta^2}, \end{aligned}$$

we obtain $L_{ijk}^{Lw}(\omega) \leq L(\omega) \leq L_{ijk}^{Up}(\omega)$, which are functions of X for fixed ω .

Last, we show that we can choose m_3 points $\sigma_1, \dots, \sigma_{m_3}$ in $B(\theta_0, \eta)$ with the fixed m_1 points $\beta_1, \dots, \beta_{m_1}$ and m_2 points $b_1, \dots, b_{m_2}$ such that for any $(\theta, \Lambda) \in B(\theta_0, \eta) \times \Gamma^*$,

$$\int \prod_N L_{ijk}^{Lw}(\omega) \phi(\omega; \sigma_r) d\omega \leq \int \prod_N L(\omega) \phi(\omega; \sigma) d\omega \leq \int \prod_N L_{ijk}^{Up}(\omega) \phi(\omega; \sigma_r) d\omega \quad (C.3)$$

for i, j, k decided above and some $1 \leq r \leq m_3$.

By the definition of $L(\omega)$, $L_{ijk}^{Lw}(\omega)$, $L_{ijk}^{Up}(\omega)$ given ω , we have that

$$\begin{aligned} L(\omega) - L_{ijk}^{Lw}(\omega) &\geq \varepsilon e^\omega, \\ L_{ijk}^{Up}(\omega) - L(\omega) &\geq \varepsilon e^\omega. \end{aligned}$$

Then

$$\begin{aligned} &\int \prod_N L(\omega) \phi(\omega; \sigma) d\omega - \int \prod_N L_{ijk}^{Lw}(\omega) \phi(\omega; \sigma_r) d\omega \\ &= \int [\prod_N L(\omega) - \prod_N L_{ijk}^{Lw}(\omega)] d\phi(\omega; \sigma) d\omega - \int \prod_N L_{ijk}^{Lw}(\omega) [\phi(\omega; \sigma_r) - \phi(\omega; \sigma)] d\omega \\ &\geq \int \prod_N [L(\omega) - L_{ijk}^{Lw}(\omega)] \phi(\omega; \sigma) d\omega - \int \prod_N L_{ijk}^{Lw}(\omega) [\phi(\omega; \sigma_r) - \phi(\omega; \sigma)] d\omega \\ &\geq \varepsilon \inf_\sigma \int e^\omega \phi(\omega; \sigma) d\omega - |\sigma - \sigma_r| \sup_\sigma \int \left| \frac{\partial \phi(\omega; \sigma)}{\partial \sigma} \right| d\omega \\ &= C_5 \varepsilon - C_6 |\sigma - \sigma_r| \geq \varepsilon, \end{aligned}$$

and

$$\int \prod_N L_{ijk}^{Up}(\omega) \phi(\omega; \sigma_r) d\omega - \int \prod_N L(\omega) \phi(\omega; \sigma) d\omega \geq C_5 \varepsilon - C_6 |\sigma - \sigma_r| \geq \varepsilon,$$

where $C_5 = \inf_\sigma \int e^\omega \phi(\omega; \sigma) d\omega$ and $C_6 = \sup_\sigma \int \left| \frac{\partial \phi(\omega; \sigma)}{\partial \sigma} \right| d\omega$ are positive and finite by Condition 6. So if we choose $\sigma_1, \dots, \sigma_{m_3}$ with the known $\beta_1, \dots, \beta_{m_1}$ and $b_1, \dots, b_{m_2}$ such that for any $(\theta, \Lambda) \in B(\theta_0, \eta)$, $\|\sigma - \sigma_r\| \leq$

$(C_5 + 1)\varepsilon/C_6$ for some $1 \leq r \leq m_3$, then (C.3) holds. It is well known that the minimum value of m_3 can be in the order of $O(\varepsilon^{-1})$.

For each $(\theta, \Lambda) \in \Theta \times \Gamma^*$, we can therefore find $i, j, k, r; 1 \leq i \leq m_0, 1 \leq j \leq m_1, 1 \leq k \leq m_2, 1 \leq r \leq m_3$ such that

$$\begin{aligned} l_{ijk}^{Lw} &= \log \int \prod_N L_{ijk}^{Lw}(\omega) \phi(\omega; \sigma_r) d\omega \leq l = \log \int \prod_N L(\omega) \phi(\omega; \sigma) d\omega \\ &\leq l_{ijk}^{Up} = \log \int \prod_N L_{ijk}^{Up}(\omega) \phi(\omega; \sigma_r) d\omega \end{aligned}$$

By the mean value theorem for the logarithm function, $|l_{ijk}^{Up} - l_{ijk}^{Lw}| \leq C\varepsilon$, we have $\int |l_{ijk}^{Up} - l_{ijk}^{Lw}|^2 dP \leq C\varepsilon^2$ for some positive constant C . It follows that $N_{[]}(\varepsilon, \mathcal{L}, L_2(P))$ is of order $m_0 \cdot m_1 \cdot m_2 \cdot m_3 = O(\varepsilon^{-2d-2}e^{1/\varepsilon})$.

□

Proof of Theorem 4.3.4

By taking a Taylor expansion of $l(\theta, \Lambda|X)$ at true values (θ_0, Λ_0) , we can obtain $El(\theta, \Lambda|X) - El(\theta_0, \Lambda_0|X) = O(D^2((\theta, \Lambda), (\theta_0, \Lambda_0)))$. The Kullback - Leibler information indicates that the term $E(l(\theta, \Lambda|X) - l(\theta_0, \Lambda_0|X))$ is non-positive. That is,

$$El(\theta, \Lambda|X) - El(\theta_0, \Lambda_0|X) \leq -CD^2((\theta, \Lambda), (\theta_0, \Lambda_0))$$

for some constant $C > 0$. Therefore,

$$\sup_{\eta/2 \leq D((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \eta} El(\theta, \Lambda|X) - El(\theta_0, \Lambda_0|X) \leq -C\eta^2.$$

By Lemma 4.6.1, the bracketing integral for the class $\mathcal{L}$ is

$$\begin{aligned} J_{[]}(\eta, \mathcal{L}, L_2(P)) &= \int_0^\eta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{L}, L_2(P))} d\varepsilon \\ &= \int_0^\eta \sqrt{1 + 1/\varepsilon - (2d + 2) \log \varepsilon} d\varepsilon \\ &= O(1) \int_0^\eta \sqrt{1/\varepsilon} d\varepsilon = O(\eta^{1/2}). \end{aligned}$$

for η close to zero. Incorporating Lemma 3.4.2 of Van Der Vaart and Wellner (1996), we have

$$\begin{aligned} & E^* \sup_{D((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \eta} |\sqrt{n}(P_n - P)(l(\theta, \Lambda|X) - l(\theta_0, \Lambda_0|X))| \\ &= O(1)\eta^{1/2}(1 + \frac{\eta^{1/2}}{\eta^2\sqrt{n}}M^*). \end{aligned}$$

where $\|l\|_\infty \leq M^*$ for every $l \in \mathcal{L}$. Let $\varphi_n(\eta) = \eta^{1/2}(1 + \frac{\eta^{1/2}}{\eta^2\sqrt{n}}M^*)$, then $\varphi_n(\eta)/\eta$ is a decreasing function, and we can verify that $n^{2/3}\varphi_n(n^{-1/3}) = O(\sqrt{n})$ for n large. With Theorem 4.3.3, the conditions of Theorem 3.4.1 in Van Der Vaart and Wellner (1996) are satisfied. This implies $D((\hat{\theta}_n, \hat{\Lambda}_n), (\theta_0, \Lambda_0)) = O_p(n^{-1/3})$.

Appendix D: Proof of Theorem 4.3.5

Theorem 4.3.5 can be proved by verifying the assumptions given in Theorem 6.1 of Huang (1996) with the underlying model setting. To do this, we first need to show the following lemma first.

Lemma 4.6.2. *Define the class of functions $\mathcal{F} = \{l(\theta, \Lambda|X) : (\theta, \Lambda) \in \Theta \times \Gamma\}$, and for any $\eta > 0$, $\mathcal{F}(\eta) = \{l(\theta, \Lambda|X) : D((\theta, \Lambda), (\theta_0, \Lambda_0)) \leq \eta \text{ and } \Lambda \in \Gamma\}$ where $\Gamma = \{\Lambda : \Lambda \text{ is increasing and bounded on } [\tau_0, \tau_1] \text{ with } \Lambda(0) = 0\}$. The L_2 covering number of $\mathcal{F}$ is $\sup_Q N(\varepsilon, \mathcal{F}, L_2(Q)) \leq O(\varepsilon^{-2d-2}e^{1/\varepsilon})$. Or the entropy number is $\sup_Q \log N(\varepsilon, \mathcal{F}, L_2(Q)) \leq O(1/\varepsilon)$ for ε close to zero, where Q runs through all probability measures. This implies that $\mathcal{F}(\eta)$ is a Donsker class.*

Proof. It was shown by van de Geer (1993) that for the uniformly bounded class of functions Γ , its L_2 covering number is on the order of $m_0 = O(e^{1/\varepsilon})$. This implies that, there exists a set of functions $\Lambda_1, \dots, \Lambda_{m_0}$ such that

$\|\Lambda - \Lambda_i\|_2 \leq \varepsilon$ for some $1 \leq i \leq m_0$ and any $\Lambda \in \Gamma$. Moreover, there exist $\beta_1, \dots, \beta_{m_1}$, $b_1, \dots, b_{m_2}$, and $\sigma_1, \dots, \sigma_{m_3}$ such that $\|\beta - \beta_j\| \leq O(\varepsilon)$, $\|b - b_k\| \leq O(\varepsilon)$, $|\sigma - \sigma_r| \leq O(\varepsilon)$ for any $\theta = (\beta, b, \sigma) \in B(\theta_0, \eta)$ and $1 \leq j \leq m_1, 1 \leq k \leq m_2, 1 \leq r \leq m_3$, where $m_1 = O(\varepsilon^{-p})$; $m_2 = O(\varepsilon^{-p-1})$; $m_3 = O(\varepsilon^{-1})$. Define $l_{ijk}(X) = l(\beta_i, b_k, \sigma_r, \Lambda_i|X)$. Since $\Lambda \in \Gamma$, Λ is uniformly bounded away from zero and ∞ . Moreover, Z is bounded. For any $l(X) \in \mathcal{F}$ and $i = 1, \dots, m_0, j = 1, \dots, m_1, k = 1, \dots, m_2$ and $r = 1, \dots, m_3$, it follows by the mean value theorem that,

$$\begin{aligned}
& |l(X) - l_{ijk}(X)| \\
&= \left| \log \int \prod_N L(\beta, b, \Lambda|X) \phi(\omega; \sigma) d\omega - \log \int \prod_N L(\beta_j, b_k, \Lambda_i|X) \phi(\omega; \sigma_r) d\omega \right| \\
&\leq c_0 \left| \int \prod_N L(\beta, b, \Lambda|X) \phi(\omega; \sigma) d\omega - \int \prod_N L(\beta_j, b_k, \Lambda_i|X) \phi(\omega; \sigma_r) d\omega \right| \\
&= c_1(I + II)
\end{aligned}$$

with

$$\begin{aligned}
I &= \left| \int \prod_N L(\beta, b, \Lambda|X) [\phi(\omega; \sigma) - \phi(\omega; \sigma_r)] d\omega \right| \\
&\leq |\sigma - \sigma_r| \sup_{\sigma} \int \prod_N L(\beta, b, \Lambda|X) \frac{\partial \phi(\omega; \sigma)}{\partial \sigma} d\omega \\
&\leq O(|\sigma - \sigma_r|) = O(\varepsilon)
\end{aligned}$$

and

$$\begin{aligned}
II &= \left| \int \prod_N L(\beta, b, \Lambda|X) \phi(\omega; \sigma_r) d\omega - \int \prod_N L(\beta_j, b_k, \Lambda_i|X) \phi(\omega; \sigma_r) d\omega \right| \\
&\leq \int \left| \prod_N L(\beta, b, \Lambda|X) - \prod_N L(\beta_j, b_k, \Lambda_i|X) \right| \phi(\omega; \sigma_r) d\omega \\
&\leq \int \varphi_1(\omega)(\|\beta - \beta_j\|) + \varphi_2(\omega)(\|b - b_k\|) + \varphi_3(\omega)(\|\Lambda(U) - \Lambda_i(U)\|_2) \\
&\quad + \varphi_4(\omega)(\|\Lambda(V) - \Lambda_i(V)\|_2) \phi(\omega; \sigma_r) d\omega \\
&\leq \int \varphi_1(\omega)O(\varepsilon) + \varphi_2(\omega)O(\varepsilon) + \varphi_3(\omega)\varepsilon + \varphi_4(\omega)\varepsilon \phi(\omega; \sigma_r) d\omega \\
&= O(\varepsilon)
\end{aligned}$$

where c_0, c_1 , are positive constants, and $\int \varphi_a(\omega) \phi(\omega; \sigma_r) d\omega < \infty$ for $a = 1, 2, 3$ and $r = 1, \dots, m_3$. This implies that the covering number for the class $\mathcal{F}$ is on the order of $m_0 m_1 m_2 m_3 = \varepsilon^{-2d-2} e^{1/\varepsilon}$

□

Proof of Theorem 4.3.5

Denote by $l_n(\theta, \Lambda)$ the log-likelihood function of the samples $\{X_{ij}\}$, $i = 1, \dots, n$, $j = 1, \dots, n_i$,

$$l_n(\theta, \Lambda) = \frac{1}{n} \log L(\theta, \Lambda),$$

where $L(\theta, \Lambda)$ is given in (4.2.1). Remember that the estimators $\hat{\theta}_n$ and $\hat{\Lambda}_n$ are obtained by maximizing $l_n(\theta, \Lambda)$ over θ and Λ , treating Λ as an increasing step function, left continuous having right limits, with possible jumps only at the interval endpoints and $\Lambda(0) = 0$. Denote the jump points of $\hat{\Lambda}_n$ by $t_1^* < \dots < t_{Q^*-1}^*$, and $t_0^* = \tau_0$, $t_{Q^*}^* = \tau_1$. By Condition 2, there exists two constant μ_0 and μ_1 such that $\mu_0 < \hat{\Lambda}_n(t_1^*) < \dots < \hat{\Lambda}_n(t_{Q^*-1}^*) < \mu_1$, and $\mu_0 \leq \hat{\Lambda}_n(t)$ for $t_0^* \leq t < t_1^*$, $\hat{\Lambda}_n(t) \leq \mu_1$ for $t_{Q^*-1}^* < t \leq t_{Q^*}^*$.

Following Huang (1996), Theorem 4.3.5 is proved if the assumptions listed in Theorem 6.1 of Huang (1996) are satisfied.

In particular, Huang's Assumption (i) is that $S_{1n}(\hat{\theta}_n, \hat{\Lambda}_n) = o_p(n^{-1/2})$ and

$S_{2n}(\hat{\theta}_n, \hat{\Lambda}_n)[h^*] = o_p(n^{-1/2})$. Since θ_0 is an interior point of Θ , $\hat{\theta}_n$ converges to θ_0 almost surely, then $\hat{\theta}_n$ lies in Θ for n sufficiently large. It follows that

$$S_{1n}(\hat{\theta}_n, \hat{\Lambda}_n) \equiv \frac{\partial}{\partial \theta} l_n(\hat{\theta}_n, \hat{\Lambda}_n) = 0$$

with probability one for n sufficiently large. But for the nonparametric component Λ , its score function evaluated at $(\hat{\theta}, \hat{\Lambda}_n)$ along the least favourable direction will in general not be zero. However, it suffices to show that the score function

$$S_{2n}(\hat{\theta}_n, \hat{\Lambda}_n)[h^*] \equiv \frac{\partial}{\partial a} l_n(\hat{\theta}_n, \hat{\Lambda}_n + ah^*) \Big|_{a=0} = o_p(n^{-1/2}). \quad (\text{D.1})$$

where h^* is found in Theorem 4.3.2. Let $\delta_{ij}^3 = 1 - \delta_{ij}^1 - \delta_{ij}^2$ in the following. The partial derivative of $l_n(\theta, \Lambda)$ with respect to Λ for fixed θ along with any function h is

$$\begin{aligned} & \frac{\partial}{\partial a} l_n(\theta, \Lambda + ah) \Big|_{a=0} \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{n_i} e^{\beta Z_{ij}} \{ \delta_{ij}^1 A_{1ij} h(U_{ij}) - \delta_{ij}^2 [A_{2ij} h(U_{ij}) - A_{3ij} h(V_{ij})] - \delta_{ij}^3 A_{4ij} h(V_{ij}) \} \end{aligned}$$

with $A_{1ij}, A_{2ij}, A_{3ij}$ and A_{4ij} are followed by the formulas in Section 3.2 for samples $\{X_{ij}\}$, $i = 1, \dots, n$, $j = 1, \dots, n_i$. Let $\hat{A}_{1n}, \hat{A}_{2n}, \hat{A}_{3n}$ and $\hat{A}_{4n}$ be the empirical versions of A_1, A_2, A_3 and A_4 with $(\hat{\theta}_n, \hat{\Lambda}_n)$. Because $\hat{\Lambda}_n$ is constant on each interval $I_k = [t_{k-1}^*, t_k^*)$, $k = 2, \dots, Q^* - 1$, then $l_n(\theta_n, \Lambda)$ is also constant which is only varied by the values of $\Lambda(U)$ and $\Lambda(V)$. It means the partial derivative of $l_n(\theta_n, \Lambda)$ with respect to Λ is zero on I_k ,

$$\int_{[U \in I_k] \cup [V \in I_k]} \sum_{j=1}^{n_i} e^{\hat{\beta}_n^T Z_{ij}} [\delta_{ij}^1 \hat{A}_{1n} - \delta_{ij}^2 \hat{A}_{2n}] + e^{\hat{\beta}_n^T Z} [\delta_{ij}^2 \hat{A}_{3n} - \delta_{ij}^3 \hat{A}_{4n}] dP_n = 0,$$

where P_n is the empirical measure. Multiplying the value of $\Lambda(U) = \Lambda(V)$ for $U, V \in I_k$, we obtain

$$\begin{aligned} \int_{[U \in I_k] \cup [V \in I_k]} \sum_{j=1}^{n_i} e^{\hat{\beta}_n^T Z_{ij}} \hat{\Lambda}_n(U) [\delta_{ij}^1 \hat{A}_{1n} - \delta_{ij}^2 \hat{A}_{2n}] \\ + e^{\hat{\beta}_n^T Z} \hat{\Lambda}_n(V) [\delta_{ij}^2 \hat{A}_{3n} - \delta_{ij}^3 \hat{A}_{4n}] dP_n = 0. \end{aligned} \quad (D.2)$$

To prove (D.1), we first derive an important equation as follows. By Condition 3, Λ_0^{-1} is well defined. Denote $\xi = h^* \circ \Lambda_0^{-1}$, h^* composed with Λ_0^{-1} . Then $\xi \circ \hat{\Lambda}_n = h^*(\Lambda_0^{-1}(\hat{\Lambda}_n))$ is constant on the intervals $\{I_k\}_{k=2}^{Q^*-1}$. From (D.1), we further get

$$\begin{aligned} \int_{[t_1^* \leq U < t_{Q^*-1}^*] \cup [t_1^* < V < t_{Q^*-1}^*]} \sum_{j=1}^{n_i} e^{\hat{\beta}_n^T Z_{ij}} \xi(\hat{\Lambda}_n(U)) [\delta_{ij}^1 \hat{A}_{1n} - \delta_{ij}^2 \hat{A}_{2n}] \\ + e^{\hat{\beta}_n^T Z_{ij}} \xi(\hat{\Lambda}_n(V)) [\delta_{ij}^2 \hat{A}_{3n} - \delta_{ij}^3 \hat{A}_{4n}] dP_n = 0. \end{aligned} \quad (D.3)$$

Along the similar lines as the proof of Theorem 3.3 in Huang and Wellner (1995) and using Conditions 1-4, Theorems 4.3.2-4.3.4 and the proof procedure of Lemma 4.6.2, we can show that the integral on the left side of (D.3) over the region outside of $[t_1^*, t_{Q^*-1}^*)$ is of rate $o_p(n^{-1/2})$. Theorem 4.3.4 confirms that h^* is well approximated by $\xi \circ \hat{\Lambda}_n$. Therefore,

$$\begin{aligned} \int \sum_{j=1}^{n_i} e^{\hat{\beta}_n^T Z_{ij}} h^*(U) [\delta_{ij}^1 \hat{A}_{1n} - \delta_{ij}^2 \hat{A}_{2n}] + e^{\hat{\beta}_n^T Z_{ij}} h^*(V) [\delta_{ij}^2 \hat{A}_{3n} - \delta_{ij}^3 \hat{A}_{4n}] dP_n \\ = o_p(n^{-1/2}). \end{aligned}$$

Thus (D.1) is proved, Assumption (i) of Theorem 6.1 in Huang (1996) is verified.

To complete the proof, we need only to show that Huang's assumption (ii) holds as well, which consists of five parts. The first part regarding consistency and rate of convergence readily follows from Theorems 4.3.3-4.3.4.

The second part about positive information can be obtained from Theorem 4.3.2. The third part on stochastic equicontinuity can be established by using Theorem 3.2.5 of Van Der Vaart and Wellner (1996) and the results of Lemma 4.6.2. A Taylor expansion of $S_1(\theta, \Lambda)$, which is the limit version of $S_{1n}(\theta_n, \Lambda_n)$ at (θ_0, Λ_0) , guarantees the fourth part concerning smoothness of the model. By Conditions 1-3 and Theorem 4.3.2, the fifth part holds, i.e.,

$$\sqrt{n}P_n l_{\theta_0}^* \longrightarrow_d N(0, I(\theta_0)).$$

with $l_{\theta_0}^*$ and $I(\theta_0)$ defined in Theorem 4.3.2.

Therefore, the asymptotic normality and efficiency of $\hat{\theta}_n$ are obtained directly from Theorem 6.1 of Huang (1996), and the proof of Theorem 4.3.5 is finished.

Chapter 5

Conclusion

In this chapter, we summarises the major finding of the studies undertaken in previous chapters and outline potential research on such complex survival data under interval censoring for the future.

5.1 Major findings

Analysis of interval-censored data has been attracting much research interest as such data often arise in many fields, especially in clinical trials and public health studies due to periodic visits of cases. In this thesis, we are interested in case 2 interval-censored data with more complex data structure, including left-, interval-, and right-censored observations. Due to the effectiveness of medical treatments, a cured subgroup of subjects exists who are not susceptible to the event of interest. When event times are observed as repeated measurements or collected from several clusters due to geographic exposures, clustered interval-censored data are then encountered.

To analyze such complex survival data with a non-negligible cure fraction, we propose a mixture cure modelling procedure in Chapter 2 by in-

tegrating a logistic regression on the cure proportion and a proportional hazards model for the uncured subgroup. Random effects are incorporated into two components of mixture cure model to account for the intracluster dependence. Under the GLMM framework, we develop the REML estimation for the parameters, as well as an iterative algorithm for estimating the unknown baseline survival function. The estimation procedure is implemented via an EM algorithm. The performance of the proposed method is evaluated by a simulation study and a real data set from a smoking cessation study.

It is obvious that the implementation of mixture model is more complicated than that of a single model. The complexity will be increased greatly when mixture model is applied to the clustered interval-censored data. We thus set up a score test to assess the presence of cured subjects in clustered interval-censored survival data in Chapter 3. A simulation study is conducted to evaluate the sampling distribution and power behaviour of the score test and a bootstrap approach is further proposed to adjust the estimated significance levels of the test. We illustrate applications of the test using data sets from a smoking cessation study and a retrospective study of early breast cancer patients, respectively.

In order to provide a wide range in model selection and avoid prespecified model assumptions, we introduce a general class of semiparametric mixture cure transformation models with random effects for clustered interval-censored data in Chapter 4. Under appropriate conditions, we show that the proposed mixture cure transformation model with random effects is identifiable. Under the proposed model, we first obtain the maximum likelihood estimators for the parameters of interest using an EM algorithm and estimator for the nonparametric component applying a self-consistent iter-

ative algorithm. Then we prove the estimator for the nonparametric component is consistent with a cube-root- n convergence rate. Furthermore, the asymptotic normality of the estimators for regression parameter and variance component of random effects is eventually obtained. We conduct intensive simulations to assess the performance of the estimators proposed and revisit the smoking cessation data to illustrate the methodology mentioned.

5.2 Future works

Based on current research, we briefly outline several future works on interval-censored survival data.

We employ the semiparametric proportional hazards function as the survival function component for uncured subjects in Chapter 2, an alternative approach would be to use the parametric accelerated failure time (AFT) model. Although a general class of semiparametric mixture cure transformation models introduced in Chapter 4 provides a wide range in model selection, the AFT models, which can be framed as linear models for the logarithm of the survival time, are excluded. Unlike proportional hazards models, the regression parameter estimates from AFT models are robust to omitted covariates. They are also less affected by the choice of probability distribution. In the literature, the work of Odell et al. (1992) has shown the ability of the parametric AFT model to handle interval-censored data. Furthermore, Lambert et al. (2004) introduced a mixture model for the hazard function with two components based on a fully parametric AFT model with shared random effects. The extension of the method proposed in Chapter 2 to a mixture cure model with the semiparametric AFT model employed

as the survival function component in the GLMM setting would thus be feasible. Such an extension would provide another potential way of dealing with interval-censored multivariate survival data with a cure fraction.

When the random effects ω and v are present in both the logistic regression and the survival function components in Chapter 2, the association between them should be kept in mind. In Chapter 2, we only consider the cases that the random effects ω and v are independent. The mixture cure model with correlated random effects has been studied by Lai and Yau (2008) for right-censored data, in which random effects followed a bivariate normal distribution. The cases of dependent random effects involved in the mixture cure model can also be studied in future under interval censoring.

Throughout Chapter 2 and Chapter 4, the covariates Z associated in the logistic regression for the portion of cured subjects are the same as those in the survival function component for uncured subjects. Along the lines of Ma (2010), the covariates in the logistic regression may have no or partial overlap with Z . It leads to a variable selection problem which is even more important and necessary for high dimensional data. The performance of proposed models will be improved when most irrelevant and redundant covariates to the response are removed.

Bibliography

- Banerjee, S. and Carlin, B. P. “Parametric spatial cure rate models for interval-censored time-to-relapse data.” *Biometrics*, 60(1):268–275 (2004).
- Bellamy, S. L., Li, Y., Ryan, L. M., Lipsitz, S., Canner, M. J., and Wright, R. “Analysis of clustered and interval censored data from a community-based study in asthma.” *Statistics in medicine*, 23(23):3607–3621 (2004).
- Bickel, P. J., Klaassen, C. A., and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore (1993).
- Cook, R. J., White, B. J. G., Yi, G. Y., Lee, K. A., and Warkentin, T. E. “Analysis of a nonsusceptible fraction with current status data.” *Statistics in medicine*, 27(14):2715–2730 (2008).
- Davidson, R. and MacKinnon, J. G. “The power of bootstrap and asymptotic tests.” *Journal of Econometrics*, 133(2):421–441 (2006).
- Farewell, V. T. “The use of mixture models for the analysis of survival data with long-term survivors.” *Biometrics*, 1041–1046 (1982).

- . “Mixture models in survival analysis: Are they worth the risk?” *Canadian Journal of Statistics*, 14(3):257–262 (1986).
- Finkelstein, D. M. “A proportional hazards model for interval-censored failure time data.” *Biometrics*, 42(4):845–854 (1986).
- Finkelstein, D. M. and Wolfe, R. A. “A semiparametric model for regression analysis of interval-censored failure time data.” *Biometrics*, 41(4):933–945 (1985).
- Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. “A note on permutation tests for variance components in multilevel generalized linear mixed models.” *Biometrics*, 63(3):942–946 (2007).
- Goggins, W. B. and Finkelstein, D. M. “A proportional hazards model for multivariate interval-censored failure time data.” *Biometrics*, 56(3):940–943 (2000).
- Goldman, A. I. “Survivorship analysis when cure is a possibility: a Monte Carlo study.” *Statistics in Medicine*, 3(2):153–163 (1984).
- Groeneboom, P. and Wellner, J. A. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser Basel (1992).
- Huang, J. “Efficient estimation for the proportional hazards model with interval censoring.” *The Annals of Statistics*, 24(2):540–568 (1996).
- Huang, J. and Wellner, J. A. “Efficient estimation for the proportional hazards model with” Case 2” interval censoring.” Technical report (1995).
- Jiang, J. and Lahiri, P. “Mixed model prediction and small area estimation.” *Test*, 15(1):1–96 (2006).

- Kim, D. K. “Regression analysis of interval-censored survival data with covariates using log-linear models.” *Biometrics*, 53:1274–1283 (1997).
- Kim, Y. J. and Jhun, M. “Cure rate model with interval censored data.” *Statistics in medicine*, 27(1):3–14 (2008).
- Kuk, A. Y. C. and Chen, C. H. “A mixture model combining logistic regression with proportional hazards regression.” *Biometrika*, 79(3):531–541 (1992).
- Lai, X. and Yau, K. K. W. “Long-term survivor model with bivariate random effects: Applications to bone marrow transplant and carcinoma study data.” *Statistics in medicine*, 27(27):5692–5708 (2008).
- Lam, K. F. and Xue, H. “A semiparametric regression cure model with current status data.” *Biometrika*, 92(3):573–586 (2005).
- Lambert, P., Collett, D., Kimber, A., and Johnson, R. “Parametric accelerated failure time models with random effects and an application to kidney transplant survival.” *Statistics in medicine*, 23(20):3177–3192 (2004).
- Li, J. and Ma, S. “Interval-censored data with repeated measurements and a cured subgroup.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(4):693–705 (2010).
- Li, Y., Ryan, L., Bellamy, S., and Satten, G. A. “Inference on clustered survival data using imputed frailties.” *Journal of Computational and Graphical Statistics*, 12(3):640–662 (2003).
- Lindsey, J. C. and Ryan, L. M. “Methods for interval-censored data.” *Statistics in Medicine*, 17(2):219–238 (1998).

- Liu, H. and Shen, Y. “A Semiparametric Regression Cure Model for Interval-Censored Data.” *Journal of the American Statistical Association*, 104(487):1168–1178 (2009).
- Lu, W. and Ying, Z. “On semiparametric transformation cure models.” *Biometrika*, 91(2):331–343 (2004).
- Ma, S. “Cure model with current status data.” *Statistica Sinica*, 19(1):233–249 (2009).
- . “Mixed case interval censored data with a cured subgroup.” *Statistica Sinica*, 20:1165–1181 (2010).
- Maller, R. A. and Zhou, S. “Testing for the presence of immune or cured individuals in censored survival data.” *Biometrics*, 51(4):1197–1205 (1995).
- Mao, M. and Wang, J. L. “Semiparametric Efficient Estimation for a Class of Generalized Proportional Odds Cure Models.” *Journal of the American Statistical Association*, 105(489):302–311 (2010).
- McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. *Generalized, Linear, and Mixed Models*. Wiley: New York, 2 edition (2008).
- McGilchrist, C. A. “REML estimation for survival models with frailty.” *Biometrics*, 49(1):221–225 (1993).
- Murray, R. P., Anthonisen, N. R., Connett, J. E., Wise, R. A., Lindgren, P. G., Greene, P. G., Nides, M. A., et al. “Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function.” *Journal of clinical epidemiology*, 51(12):1317–1326 (1998).

- Nelson, K. P., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J., Parzen, M., and Strawderman, R. “Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects.” *Journal of Computational and Graphical Statistics*, 15(1):39–57 (2006).
- Ng, S. K., McLachlan, G. J., Yau, K. K. W., and Lee, A. H. “Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment.” *Statistics in medicine*, 23(17):2729–2744 (2004).
- Odell, P. M., Anderson, K. M., and D’Agostino, R. B. “Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model.” *Biometrics*, 48(3):951–959 (1992).
- Pan, W. “A multiple imputation approach to cox regression with interval-censored data.” *Biometrics*, 56(1):199–203 (2000).
- Pan, W. and Chappell, R. “Estimation in the cox proportional hazards model with left-truncated and interval-censored data.” *Biometrics*, 58(1):64–70 (2002).
- Peng, Y. and Dear, K. B. G. “A nonparametric mixture model for cure rate estimation.” *Biometrics*, 56(1):237–243 (2000).
- Peng, Y., Dear, K. B. G., and Carriere, K. C. “Testing for the presence of cured patients: a simulation study.” *Statistics in medicine*, 20(12):1783–1796 (2001).
- Rabinowitz, D., Tsiatis, A., and Aragon, J. “Regression with interval-censored data.” *Biometrika*, 82(3):501–513 (1995).

- Shen, X. “Proportional odds regression and sieve maximum likelihood estimation.” *Biometrika*, 85(1):165–177 (1998).
- Sinha, S. K. “Bootstrap tests for variance components in generalized linear mixed models.” *Canadian Journal of Statistics*, 37(2):219–234 (2009).
- Sun, J. *The statistical analysis of interval-censored failure time data*. Springer Verlag (2006).
- Sy, J. P. and Taylor, J. M. G. “Estimation in a cox proportional hazards cure model.” *Biometrics*, 56(1):227–236 (2000).
- Taylor, D. J., Weaver, M. A., and Roddy, R. E. “Evaluating factors associated with STD infection in a study with interval-censored event times and an unknown proportion of participants not at risk for disease.” *Statistics in medicine*, 22(13):2191–2204 (2003).
- Turnbull, B. W. “The empirical distribution function with arbitrarily grouped, censored and truncated data.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 38:290–295 (1976).
- van de Geer, S. “Hellinger-consistency of certain nonparametric maximum likelihood estimators.” *The Annals of Statistics*, 21(1):14–44 (1993).
- Van den Broek, J. “A score test for zero inflation in a Poisson distribution.” *Biometrics*, 738–743 (1995).
- Van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, Cambridge (2000).
- Van Der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York (1996).

- Wong, M. C. M., Lam, K. F., and Lo, E. C. M. “Bayesian analysis of clustered interval-censored data.” *Journal of dental research*, 84(9):817–821 (2005).
- Xiang, L. M., Lee, A. H., Yau, K. K. W., and McLachlan, G. J. “A score test for zero-inflation in correlated count data.” *Statistics in medicine*, 25(10):1660–1671 (2006).
- Yang, Z., Hardin, J. W., and Addy, C. L. “Testing overdispersion in the zero-inflated Poisson model.” *Journal of Statistical Planning and Inference*, 139(9):3340–3353 (2009).
- Yau, K. K. W. and Ng, A. S. K. “Long-term survivor mixture model with random effects: application to a multi-centre clinical trial of carcinoma.” *Statistics in medicine*, 20(11):1591–1607 (2001).
- Yu, B. and Peng, Y. “Mixture cure models for multivariate survival data.” *Computational Statistics & Data Analysis*, 52(3):1524–1532 (2008).
- Zeng, D., Cai, J., and Shen, Y. “Semiparametric additive risks model for interval-censored data.” *Statistica Sinica*, 16(1):287–302 (2006a).
- Zeng, D. and Lin, D. Y. “Semiparametric transformation models with random effects for recurrent events.” *Journal of the American Statistical Association*, 102(477):167–180 (2007).
- Zeng, D., Lin, D. Y., and Lin, X. “Semiparametric transformation models with random effects for clustered failure time data.” *Statistica Sinica*, 18(1):355–377 (2008).
- Zeng, D., Yin, G., and Ibrahim, J. G. “Semiparametric transformation

- models for survival data with a cure fraction.” *Journal of the American Statistical Association*, 101(474):670–684 (2006b).
- Zhang, J. and Peng, Y. “A new estimation method for the semiparametric accelerated failure time mixture cure model.” *Statistics in medicine*, 26(16):3157–3171 (2007).
- Zhao, Y., Lee, A. H., Yau, K. K. W., Burke, V., and McLachlan, G. J. “A score test for assessing the cured proportion in the long-term survivor mixture model.” *Statistics in medicine*, 28(27):3454–3466 (2009).
- Zhou, S. and Maller, R. A. “The likelihood ratio test for the presence of immunes in a censored sample.” *Statistics*, 27:181–201 (1995).

List of Works

Below is the list of works done during my PhD studies in NTU.

1. Xiang, L.M., Ma, X., and Yau, K.K.W. "Mixture cure model with random effects for clustered interval-censored survival data." *Statistics in Medicine*, 30(9):995-1006 (2011).
2. Ma, X. and Xiang, L.M. "Testing for the presence of a cure fraction in clustered interval-censored survival data." *Australian and New Zealand Journal of Statistics* (Revised).
3. Ma, X. and Xiang, L.M. "Semiparametric mixture cure transformation models for clustered interval-censored survival data." Submitted.