8-2024

Estimation of Additive Cure Model with Applications

Modou Lamin Sanyang

Estimation of Additive Cure Model with Applications

by

Modou lamin Sanyang

Under the Direction of Committee Li-Hsiang Lin, Ph.D.

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

Masters of Science

in the College of Arts and Sciences

Georgia State University

2024

ABSTRACT

Survival analysis plays a crucial role in medical research for understanding the time until an event of interest occurs, such as disease recurrence or death. An important branch of survival analysis models is cure models, assuming that a proportion of subjects will never experience the event of interest. The value of the proportion is called the cured rate and is usually associated with many covariates with complex effect relationships. Studying cure models under such non-linear covariate effects remains an active research area. This thesis aims to investigate advancements in additive cure models, focusing on their ability to capture additive complex relationships between covariates and survival outcomes with a cured fraction through non-linear modeling techniques, such as basic splines. Additive cure models offer a robust framework for analyzing survival data when a subset of individuals is cured and does not experience the event. The thesis will involve simulation studies to assess the accuracy of parameter estimation and model fit in various scenarios, and the application of additive cure models to real-world datasets from medical research studies. The findings will enhance the understanding and application of additive cure models in analyzing survival data with non-linear covariate effects, with implications for clinical decision-making and prognostic modeling. The insights gained from this research have implications for various fields, including epidemiology, clinical research, and public health, providing valuable tools for analyzing survival data and enhancing decision-making processes.

INDEX WORDS: Nonparametric estimation, Cure fraction, Survival analysis, Additive covariates, Censored Data, Robust estimators, Covariates Effects.

Estimation of Additive Cure Models with Applications

by

Modou lamin Sanyang

Committee Chair: Li-Hsiang Lin

Committee: Gengsheng Qin

Yichuan Zhao

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2024

DEDICATION

This thesis is dedicated to the remarkable individuals who have shaped my academic journey. To Professor Lin, my esteemed supervisor, your guidance and expertise have been the cornerstone of this research endeavor. Your commitment to excellence has inspired me to push the boundaries of my understanding and strive for academic rigor. I am deeply grateful for the opportunities you have provided and the knowledge you have imparted.

To my parents, whose unwavering support and belief in my abilities have been the driving force behind my pursuit of knowledge. Your sacrifices and encouragement have been the bedrock of my academic achievements. I also extend my heartfelt thanks to my beloved grandfather, with whom I share this academic voyage. His wisdom and constant support have created an enriching environment for my studies. To my friends and the professors in the Department of Mathematics and Statistics, thank you for the camaraderie, insights, and shared passion for learning. This thesis is a testament to the collective impact of these individuals on my academic and personal growth.

ACKNOWLEDGMENTS

Completing this thesis has been a transformative experience, and I am indebted to several individuals who have played pivotal roles in this academic endeavor. First and foremost, I extend my sincere appreciation to Professor Lin, my supervisor, for his unwavering support, insightful guidance, and dedication to pushing the boundaries of knowledge. His mentorship has been instrumental in shaping the trajectory of my research, and I am truly grateful for the opportunity to learn under his expertise.

I would also like to express my deepest gratitude to my parents for their enduring encouragement, belief in my potential, and the sacrifices they made to facilitate my education. Additionally, living with my grandfather has been a source of inspiration and a reminder of the importance of resilience and perseverance. To my friends, whose camaraderie provided much-needed balance and laughter during challenging times, and to the professors and colleagues in the Department of Mathematics and Statistics, thank you for fostering an intellectually stimulating environment. Each of these individuals has left an indelible mark on my academic journey, and I am thankful for their contributions to the successful completion of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Background

In the realm of survival analysis, it is not uncommon to encounter situations in which a portion of the subjects studied never experience an event of interest. This scenario frequently arises in medical research, particularly when investigating the survival time of patients undergoing a specific treatment for a particular illness. When a patient successfully recovers from the disease, the time until death from that specific ailment will never be observed, leading to a phenomenon known as "cure" in the context of survival analysis [1]. Let T represent a random variable representing the time until a clearly defined event, henceforth termed "death", the key characteristic of survival model for describing the cure phenomenon is that the associated survival probability will go to a nonzero value when t for T = t goes to infinity; that is,

$$\lim_{t\to\infty} S(t) = \lim_{t\to\infty} P(T > t) = p > 0 \tag{1.1}$$

where $S(.)$ is the population survival function for the event time T. The value p is interpreted as the cure rates.

During the past two decades, numerous research papers have investigated the extension of survival models to incorporate the concept of a cure fraction. These specialized models, aptly named "cure models," naturally account for scenarios where a portion of the population does not experience an event of interest, leading to a distinctive focus in the literature. The existing body of literature encompasses a diverse array of statistical models to extend the

model (1.2) with covariates to understand the association between cure rates and covariates, ranging from parametric, semiparametric and fully nonparametric models.

When the cure phenomenon is observed from the dataset, it is also commonly to observe the right censoring together. That is, for individuals who are not cured may either be censored or experienced the target event. It can be difficult to distinguish censored data; therefore, to determine the cure fraction, it becomes evident that certain assumptions must be imposed on the model to identify and estimate this fraction accurately. An common method for assessing this assumption in practical terms is to examine whether the Kaplan and Meier curve demonstrates a sufficiently extended plateau, encompassing numerous censored observations, and the censored data occur after the last failure time points are treated as censored data. In this thesis, we will explore a new semiparametric cure model, with a nonparametric model for assessing the cure rate, under right censoring.

1.2 Literature Review on Cured Models

A cure model in survival analysis addresses scenarios in which a subset of individuals in a study population is immune or cure of experiencing the event of interest (e.g., disease recurrence or death). There are two main types of cure models: the mixture cure model and the promotion time cure model. The mixture cure model assumes a two-component mixture distribution, where one component represents individuals who are susceptible to the event (uncured) and the other component represents those who are not susceptible (cured)[1, 11, 12]. Such model enables the cure rate and the survival function to depend on distinct sets

of covariates. The promotion time cure model, which integrates the cured fraction implicitly into the survival function, suitable for situations such as long-term vaccine efficacy studies [14]. This framework allows for the estimation of both the cure fraction and the survival distribution for the uncured individuals.

Choosing between these models depends on the characteristics of the data and the need to distinguish between susceptible and cured individuals in the population under study. Since our focus is on a non-parametric model for the cure rate, we focus on reviewing the literature on non-parametric cure models in survival analysis. [7] explore nonparametric estimation and testing within a cure model framework. Their approach focuses on estimating the survival function using non-parametric methods, such as Kaplan-Meier estimation, to accommodate censoring in datasets where a subset of individuals may be indefinitely free from the event of interest. They propose testing procedures to assess the presence of a cured fraction in the population, providing statistical tools to determine if a proportion of individuals can be considered cured beyond a specified time point. This research contributes methodologically to survival analysis by offering robust techniques for identifying and analyzing cured individuals in medical and epidemiological studies, thereby informing clinical decisions and prognostic evaluations. [3] introduced a non-parametric estimator for cure rates from right-censored survival data, aiming to overcome limitations associated with parametric assumptions about survival distributions. Their method involves two key steps: first, estimating the nonparametric survival function $S(t)$, which describes the probability of surviving beyond time t, using techniques like Kaplan-Meier estimation to account for censoring. Second, they esti-

mate the cure fraction p, representing the proportion of patients considered cured beyond a specified time point where the risk of recurrence is assumed to be negligible. This nonparametric approach allows for flexible estimation of the cure rate λ, defined as $\lambda = p \cdot \lambda_0$, where λ_0 is the baseline hazard rate. [17] presented a versatile nonparametric estimator for evaluating cure rates, incorporating smoothness assumptions in the underlying hazard function. Their method utilizes kernel smoothing techniques and bootstrap resampling to accommodate diverse data distributions and censoring mechanisms effectively. [4] Investigated the nonparametric comparison of survival functions using interval-censored data with varying censoring rates. Their study addresses the methodological challenge of comparing survival distributions when event times are only known within intervals and censoring rates differ across groups or time periods. They proposed a robust methodology that employs nonparametric techniques to estimate and compare survival functions under these conditions, contributing to statistical methods tailored for interval-censored data analysis. These studies collectively highlight the importance of nonparametric methods in providing robust and reliable estimates of cure rates in survival analysis, but these works do not incorporate covariates additively, which is different from the approach in this paper.

We are interested in using the cure model in the promotion time because it has a nice biological interpretation [14]. The survival function of the promotion time cure model can be expressed as

$$S(t) \quad = \quad \exp(-\theta(X)F(t)), \tag{1.2}$$

where $S(.)$ is the population survival function for the event time T, X is the covariates and $F(.)$ is an unknown baseline cummulative distribution function. Previous Studies on PT cure models usually assume a nonparametric form (right-continuous function with jumps) for $F(\cdot)$ while they estimate $\theta(\cdot)$ through a parametric form [15, 9]. These methods cannot capture nonlinear regression effects for the cure rates. Then, several studies extend the parametric $\theta(\cdot)$ to nonparametric form, including nonparametric splines [2], neural network [13], and support vector machine [10]. However, more exploration are needed about the model identifiability and the estimation procedure may be inefficient when there are many inputs. To fix some of these issues, a method is proposed by [8], which does not suffer the identifiability problem but the model can be inefficient when there are many inputs. Along this research direction, solving the non-identifiable and many input issues motives us to incorporate additive model in the PT cure model, which is the main focus of the literatue.

1.3 Statement of Problem

The central focus of this thesis is to explore how additive models can be incorporated for the PT cure model to effectively estimate cure fractions. We suggest a non-parametric modeling approach for the cure rate, signifying our assumption that cure rate does not adhere to the constraints of a specific family of conditional probability functions. As for the sur-

vival function, we refrain from imposing any specific assumptions. This choice allows for a versatile and diverse range of models for the survival function, providing flexibility in model selection. Specifically, the research aims to investigate the impact of additional factors on the estimation of cure fractions within the framework of nonparametric survival analysis. Using additive models, this study seeks to assess the separate effects of each covariate on the survival function while maintaining the flexibility of nonparametric estimation. Through careful consideration of the potential interactions between covariates, alongside the utilization of techniques such as spline functions or smoothing methods to account for non-linear relationships.

The rest of the thesis is organized as follows; chapter 2 discuses the likelihood under (1.2) with an additive model and proposes an algorithm for estimation. Some estimation properties will also be discussed in this Chapter. In Chapter 3, we employ simulated data to evaluate the practical performance of the proposed estimators and extend our analysis to real-world datasets.

CHAPTER 2

Methodology

In this section, we discuss a new non-parametric form of cure model by assuming a non-parametric form for the cure rate, parametric form for the baseline and additivity for the covariates. Nonparametric methods provide a flexible framework for analyzing survival data without making strong distributional assumptions. In this study, our focus lies on estimating $\theta(\cdot)$ within (1.2) using a non-linear additivity approach. We express $\theta(X)$ as:

$$\theta(X) = \exp(f_1(X_1) + f_2(X_2) + f_3(X_3) + \ldots + f_{p-1}(X_{p-1}) + f_p(X_p)) = \exp(m(X)) \quad (2.1)$$

Here, X represents continuous covariates, and $f_i(X_i)$ for each i from 1 to p denotes an unknown smooth function. We assume a parametric family for $\exp(\lambda)$ within $F(\cdot;\gamma)$, where λ is unknown. Nonparametric methods are particularly well-suited for analyzing censored survival data, offering the flexibility to estimate survival probabilities and hazard rates without imposing stringent assumptions on the underlying data distribution. We'll focus more on estimating $\theta(\cdot)$ as it's the key contributor to the cure rate, rather than $F(\cdot)$. By estimating the coefficients of $\theta(\cdot)$ using this method, we can assess the separate contributions of covariates to the survival function while maintaining the flexibility of nonparametric estimation. To develop the estimation procedure, we first show the proposed model is identifiable.

2.1 Identifiability Properties

In the landscape of cure model, the identifiability of cure models is a significant concern. [5] explores conditions that can make the two types of cure models (mixture cure models and

PT cure models) become identifiable or non-identifiable. For PT cure models, one concrete result from Theorem 6 of their paper is that if $F(\cdot)$ follows a parametric distribution, then the PT cure model (1.2) are model identifiable. This motivates us to consider a parametric assumption on $F(\cdot)$ to make $\theta(\cdot)$ and $F(\cdot)$ identifiable. To further make each component function $f_j(\cdot)$ in (2.1) identifiable, we follow **Tibshirani** to assume $E(f_j(X_j)) = 0$ for $j = 1, \cdots, p$. The result is summarized below:

Theorem [Model Identifiability] If $E(f_j(X_j)) = 0$ for $j = 1, \cdots, p$ in model (2.1) and $F(\cdot)$ is assumed to follow a parametric distribution in model (1.2). Then, $\{f_j(X_j)\}_{j=1}^p$ and the unknown parameter in $F(\cdot)$ are identifiable and hence the cure rate $\exp(-\theta(X))$ from the PT cure model is identifiable [6].

2.2 Estimation Method and the Likelihood Function

For each component function, we consider basis expansion; that is,

$$f_q(X_q) = \sum_j \alpha_{qj} B_{qj}(X_1) \text{ for } q = 1, \cdots, p,$$

where $B_{qj}(\cdot)$ is the basis spline (cite?).

Our goal is to develop an iterative algorithm to find estimators of the basis expansions $\{\alpha_{qj}\}$ for the nonparametric cured rate, and λ from the parametric baseline function.

2.2.1 Data Generation

The variable we're interested in is T, a non-negative random variable representing the time until a specific event occurs. We assume that we have independent and identically distributed (i.i.d.) data (Y_i, Δ_i, X_i), $i = 1, \ldots, n$, having the same distribution as (Y, Δ, X), where $Y = \min(T, C)$, $\Delta = I(T \leq C)$, X is a vector of covariates, the event time T follows the cure model given in 1.2. We assume T is subject to random right censoring. Instead of directly observing T, we observe $Y = \min(T, C)$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ represents the indicator function and C is the random censoring time. When there is a cure fraction, the survival function $S(t) = P(T > t)$ for T ensures that as t approaches infinity, $\lim_{t\to\infty} S(t) > 0$. This signifies the proportion of cured subjects, known as the cure rate. Because of right censoring, T is never observed when it equals infinity. When $\Delta = 1$ (uncensored observation), we know the individual is susceptible (uncured). In contrast, when $\Delta = 0$ (censored observation), the individual could belong to either sub-population, and we lack certainty about their status.

2.2.2 Likelihood Function

The likelihood function, given $(X_1, X_2, \ldots, X_n)^\top$, can be expressed as:

$$L = \prod_{i=1}^{n} \left\{ [f_T(Y_i; \theta(X_i), \gamma)]^{\Delta_i} [S_T(Y_i; \theta(X_i), \gamma)]^{1-\Delta_i} \right\}^{I(Y_i<\infty)} [S_T(\infty; \theta(X_i), \gamma)]^{I(Y_i=\infty)} \tag{2.2}$$

For a detailed derivation of L under Equation 1.2, refer to [18] and [16]. The corresponding log-likelihood function is:

$$L = \sum_{i=1}^{n} \{\Delta_i[\log(\theta(X_i)) + \log f(Y_i;\gamma)] - \theta(X_i)F(Y_i;\gamma)\} \tag{2.3}$$

The derivation of 2.3 is provided in Appendix A. Please note that $F(\infty) = 1$ for observations where $Y_i = \infty$, indicating cure. Then, the optimizer of the likelihood function (2.3) are found by using optimization function in R.

CHAPTER 3

Simulation Study

3.1 Simulation

To assess the performance of the proposed estimation approach, we conducted two simulation examples. The simulation procedure encompasses several critical steps: generating covariates, computing true cure rates, simulating survival times, and fitting the model using nonparametric methods. The sample size for the simulations is set at $n = 200$, and the process is iterated 1000 times to ensure robustness and reliability of the results.

The data generation scheme for an improper $S_T(\cdot|x)$ is outlined as follows: Two covariates, X_1 and X_2, are generated independently for each observation. Specifically, X_1 is drawn from a uniform distribution over the interval $[1, 3]$, and X_2 is drawn from a normal distribution with mean 0 and standard deviation 1. These covariates are selected to encapsulate a variety of distributional characteristics, which can influence the cure rate in different manners. The true cure rate is calculated using the functions f_1 and f_2 combined linearly to form $m(\cdot)$. The values are then exponentiated and the true cure probability cure.p $= \exp(-\theta(X_i))$ is computed. Subsequently, independent uniform values are generated from $U(0, 1)$ to determine whether an observation is cured. For each individual, we first determine whether the individual is cured. This is done by comparing the individual's random uniform number U_i on $(0, 1)$ with the complement of their cure probability $1 -$ cure.p. If U_i exceeds this threshold, i.e., $U_i > 1 -$ cure.p, the individual is classified as cured. For cured individuals, the observed time Y_i is set to infinity (∞), indicating that no event will occur, and the event

indicator Δ_i is set to 2, representing being cured. For individuals who are not cured, we proceed to simulate their survival and censoring times. The survival time T_i is generated using an inverse cumulative distribution function (CDF) approach $T_i = F_T^{-1}(U_i|X_i)$ (The detailed is summarized in Appendix B), which takes the random number U_i, a rate parameter λ_i, and an individual-specific parameter. The censoring time C_i is generated from an exponential distribution with a rate parameter of 10. We then compare the survival time T_i to the censoring time C_i to determine which event occurs first. If T_i is less than C_i; $(T_i < C_i)$, the event occurs before censoring, so the observed time Y_i is set to the survival time T_i, and the event indicator Δ_i is set to 1, indicating that the event occurred. If T_i is greater than or equal to C_i; $(T_i \geq C_i)$, the individual is censored, so the observed time Y_i is set to the censoring time C_i, and the event indicator Δ_i is set to 0, indicating censoring. The fitting of the model uses basis splines to capture the effects of covariates X_1 and X_2.

Table 3.1: Simulation Settings and Some Summary Statistics for the simulation datasets

	Example 1	Example 2
Sample size	200	200
Simulation Time	1000	1000
% Cure	1%	10.6%
% Censored	10.6%	21.6%
% Censored but not Cure	8.4%	11%

3.2 Detailed Algorithm

The detailed algorithm for conducting the simulation is summarized in this section. The major step in our algorithm is to find the estimators of the unknown coefficients by optimizing

likelihood function (2.3). The algorithm is provided below with using two covariates $p = 2$ for illustrations:

- **Step 1: Generate Covariates**
  - Generate n samples for covariate X_1 from a uniform distribution $U(1, 3)$.
  - Generate n samples for covariate X_2 from a normal distribution $N(0, 1)$.
- **Step 2: Basis Expansion for Covariates**
  - Define a set of basis functions $B_{1j}(X_1)$ for X_1 and $B_{2k}(X_2)$ for X_2 using B-splines.
  - Expand $f_1(X_1)$ and $f_2(X_2)$ as linear combinations of the basis functions:

$$f_1(X_1) = \sum_j \alpha_j B_{1j}(X_1)$$

$$f_2(X_2) = \sum_k \beta_k B_{2k}(X_2)$$

- **Step 3: Compute True Cure Rate**
  - Compute the linear predictor $m(X_i) = f_1(X_1) + f_2(X_2)$ for each individual i.
  - Compute the true cure rate as cure.p $= \exp(-\theta(X_i))$, where $\theta(X_i) = \exp(m(X_i))$.
- **Step 4: Determine Cure Status and Generate Survival Times**
  - For each individual i:
    - Generate a random uniform variable U_i from $U(0, 1)$.
    - If $U_i > 1 -$ cure.p:

- Set survival time $Y_i = \infty$ (indicating cured).
- Set censoring indicator $\Delta_i = 2$.

* Otherwise:
  - Generate survival time T_i using the inverse cumulative distribution function $T_i = F_T^{-1}(U_i|X_i)$ (Please refer to APPENDIX B).
  - Generate censoring time C_i from an exponential distribution with a rate parameter of 10.
  - Compare T_i and C_i:
  - If $T_i < C_i$:
  - Set $Y_i = T_i$ and $\Delta_i = 1$.
  - If $T_i \geq C_i$:
  - Set $Y_i = C_i$ and $\Delta_i = 0$.

- **Step 5: Fit the Model Using Nonparametric Methods**
  - Express $\theta(X)$ as:

$$\theta(X) = \exp(f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)) = \exp(m(X))$$

  - Model the survival function $\log(S(t))$ as:

$$\log(S(t)) = -\theta(X)F(t)$$

- Construct the likelihood function:

$$L = \prod_{i=1}^{n} \left\{ [f_T(Y_i; \theta(X_i), \gamma)]^{\Delta_i} [S_T(Y_i; \theta(X_i), \gamma)]^{1-\Delta_i} \right\}^{I(Y_i<\infty)} [S_T(\infty; \theta(X_i), \gamma)]^{I(Y_i=\infty)}$$

- Derive the log-likelihood function:

$$\log L = \sum_{i=1}^{n} \{\Delta_i [\log(\theta(X_i)) + \log f(Y_i; \gamma)] - \theta(X_i) F(Y_i; \gamma)\}$$

- Continue until convergence (change in log-likelihood or coefficients is below a threshold).

- **Step 6: Simulation Study**

  - Set sample size $n = 200$.

  - Iterate the process 1000 times to ensure robustness.

  - For each iteration:

    * Generate covariates X_1 and X_2.

    * Compute true cure rates and determine cure status.

    * Simulate survival and censoring times.

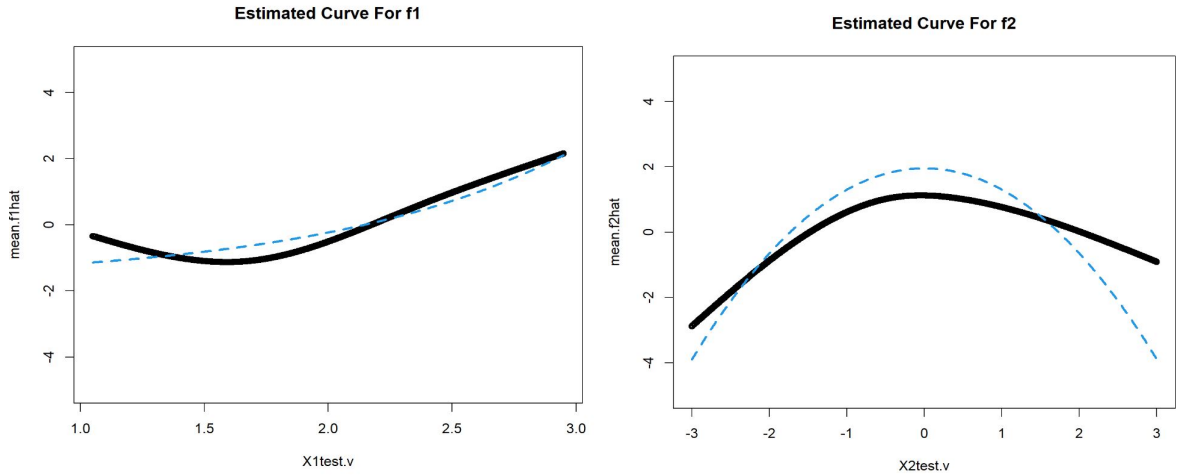
    * Fit the model using nonparametric methods and optimize the coefficients.

  - Analyze the results to assess the performance of the proposed estimation approach.

We will apply the simulation process to the examples in the following two sections.

3.3 Example 1: $\theta(X) = \exp\left(1.5 + 0.2\exp(X_1) - \frac{(1.3X_2^2)}{3}\right)$

The baseline $F(\cdot;\ \gamma)$ is taken as an exponential distribution with parameter $\gamma = -\log(10)$, and the censoring time C is approximately distributed as $U(0, 1)$. The resulting overall censoring and cure rates are 10.6% and 1.0%, respectively, which means that 8.4% are censored but not cured. This value reflects the subjects who were followed up until a certain point without experiencing the event but were not deemed cured.



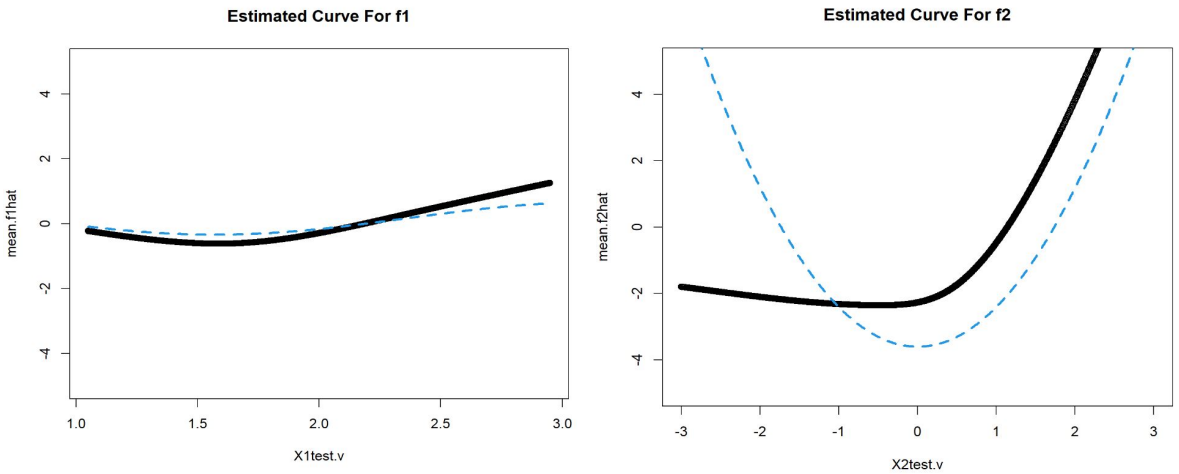
(a) (b)

Figure 3.1 Estimated Curve Example 1

The plots in Figure 3.1 show the estimated relationships between the covariates X_1 and X_2 on the x-axes and the mean estimated values of the functions for f_1 and f_2 on the y-axes, respectively. In the first plot, the black solid line represents the estimated curve for the mean estimate of f_1, showing a slight initial decline followed by a rise, indicating a non-linear relationship with covariate X_1. The close alignment with the blue dashed line suggests a strong similarity between the estimated model and the theoretical curve. In the

second plot, the solid black line for the mean estimate of f_2 follows a parabolic trajectory, peaks around a zero value of X_2, and then declines, with the blue dashed line similarly indicating a comparable trend. The proximity of the two lines in both plots highlights minor deviations, suggesting slight differences in model predictions but overall similar trends. These visualizations are effective for assessing the model fits and their capability to capture the underlying relationships between the predictor variables and the estimated functions, indicating the robustness and accuracy of the models used.

3.4 Example 2: $\theta(X) = \exp\left(0.5 + 0.5\cos(2X_1) - 1.2X_2^2\right)$



(a) (b)

Figure 3.2 Estimated Curve Example 2

Taking the same baseline, the overall censoring rate of 21.6% indicates that 21.6% of the subjects were censored, meaning their follow-up ended before the event of interest (e.g., relapse, failure) occurred. Censoring can result from subjects dropping out of the study or the study ending before the event is observed. The overall cure rate of 10.6% represents

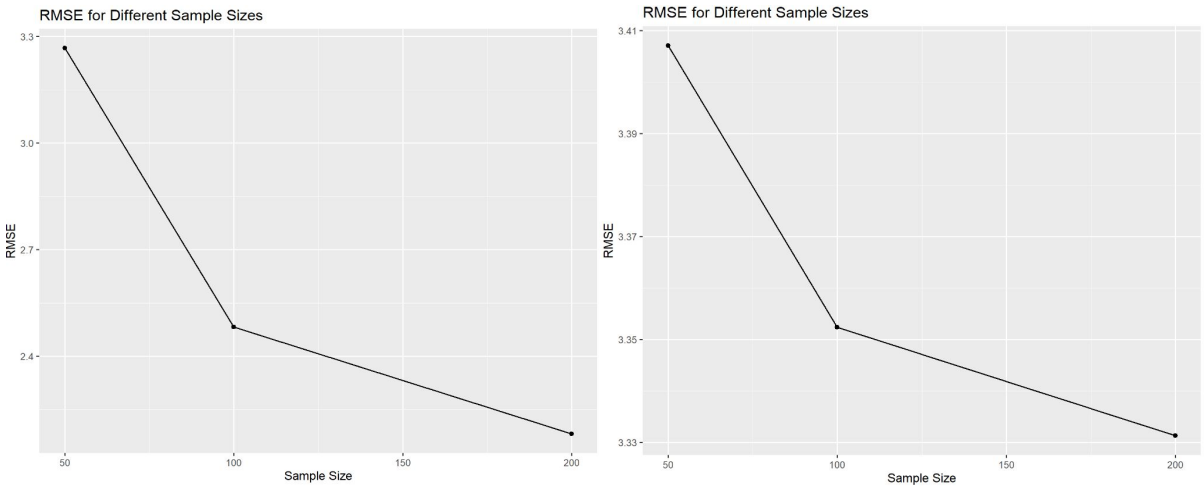
the proportion of subjects considered cured, implying that they were free from the event of interest over the study period. This low cure rate suggests that only a small fraction of subjects reached a state in which they were unlikely to experience the event in the future. Finally, the percentage of subjects censored but not cured was 11%, reflecting the subjects who were followed up to a certain point without experiencing the event but were not deemed cured.

The plots in Figure 3.1 illustrate the estimated relationships between the covariates X1 and X2 and the mean estimated values of the functions f1 and f2, respectively. In the first plot, the black solid line represents the estimated curve for f1, showing a slight initial decline followed by a rise, closely aligning with the blue dashed theoretical curve and indicating a non-linear relationship with X1. In the second plot, the estimated curve for f2 follows a parabolic trajectory, peaking around a zero value of X2 and then rising sharply, again closely matching the theoretical curve despite some deviations at the extremes. The strong alignment between the estimated and theoretical curves in both plots suggests that the models accurately capture the underlying relationships between the predictor variables and the estimated functions, demonstrating their robustness and effectiveness.

3.5 Check Robustness of the Proposed Method

In this section, we check the performance of the proposed method under different sample size settings for Examples 1 and 2. The sample size we choose is 50, 100, and 200 and we report the root mean square error (RMSE) values versue the 3 sample sizes in the Figures

3.3: RMSE for both examples consistently decreases as the sample size increases, indicating



(a) (b)

Figure 3.3 RMSE Curve for Examples 1 and 2

that the model predictions become more accurate with larger data sets. The most significant improvement in RMSE occurs between the smallest sample sizes (50 to 100), suggesting that a substantial portion of the model's robustness is achieved with the initial increase in sample size. Beyond this point, the rate of improvement slows, demonstrating diminishing returns. Both plots suggest a convergence trend as the sample size approaches 200, where further increases in sample size result in minimal gains in accuracy. This stabilization indicates that the model's performance is robust, as it does not fluctuate unpredictably with increasing sample sizes. The consistent decline in RMSE across all sample sizes and the absence of sudden fluctuations confirm the robustness of the model, ensuring reliable generalization and reducing the likelihood of overfitting. Based on these observations, we identify an optimal sample size range of 150 to 200, balancing the effort of data collection with the resulting

improvement in model performance.

CHAPTER 4

Real Data Analysis

In this chapter, we apply the proposed method to a real dataset about Melanoma.

4.1 Data Illustration

The "Survival from Malignant Melanoma" dataset is a critical resource for investigating survival outcomes in patients treated for malignant melanoma at the Department of Plastic Surgery, University Hospital of Odense, Denmark, from 1962 to 1977. This dataset comprises 205 observations and 7 variables, capturing detailed measurements from patients who underwent surgical tumor removal, including excision of approximately 2.5 cm of surrounding skin. Key prognostic variables such as tumor thickness and ulceration status are pivotal for assessing melanoma mortality risk. Patients were monitored until the end of 1977 to document their survival status. Summary statistics highlight significant findings: the dataset shows a median survival time of 2005 days (approximately 5.5 years) and a mean of 2153 days (almost 5.9 years). Patient ages range from 4 to 95 years, with a median of 54 years and a mean of 52.46 years, while tumor thickness varies widely, with a median of 1.94 mm and a mean of 2.92 mm, ranging from 0.1 to 17.42 mm.

The variable status indicates patient outcomes: 1 signifies deaths from melanoma, 2 indicates patients who remain alive, and 3 represents deaths from unrelated causes. Categorical data analysis shows a mean status of 1.79, with a median of 2, highlighting predominantly surviving patients. Moreover, the variable ulcer, denoting the presence (1) or absence (0) of ulceration, exhibits a mean prevalence of 0.439, indicating that ulceration was present in

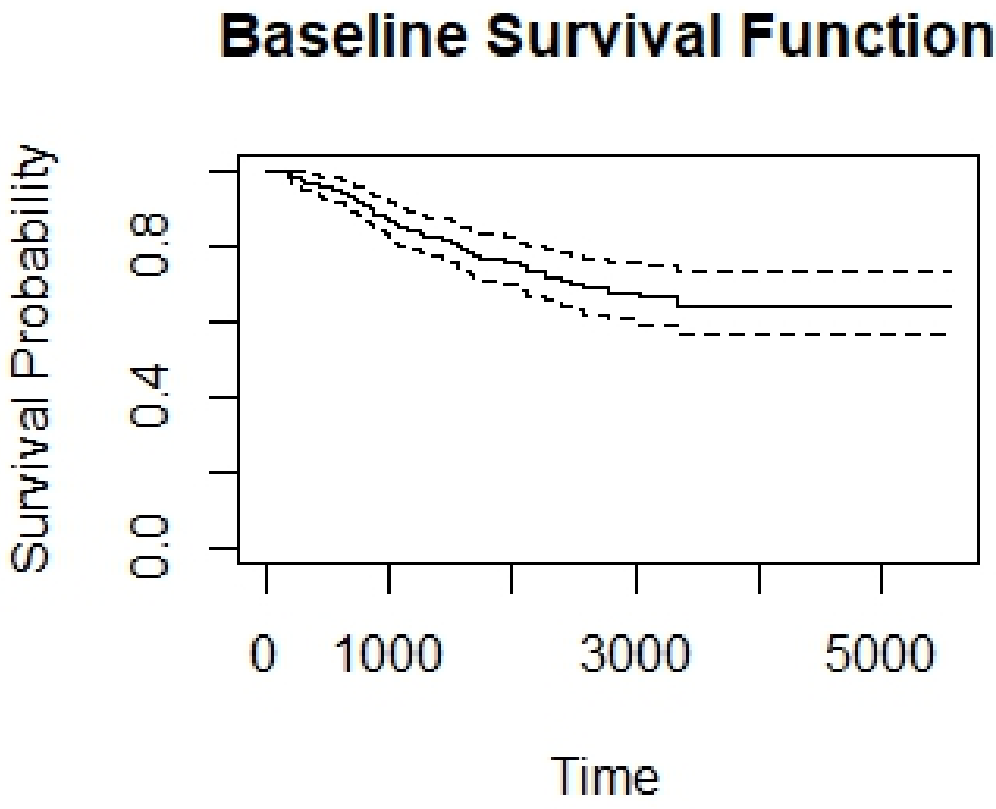
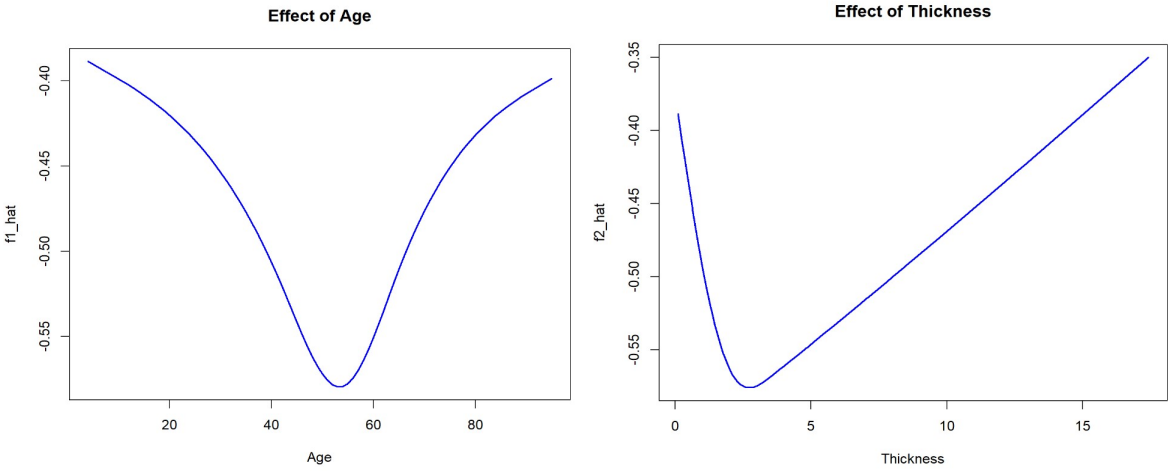


Figure 4.1 Survival curve

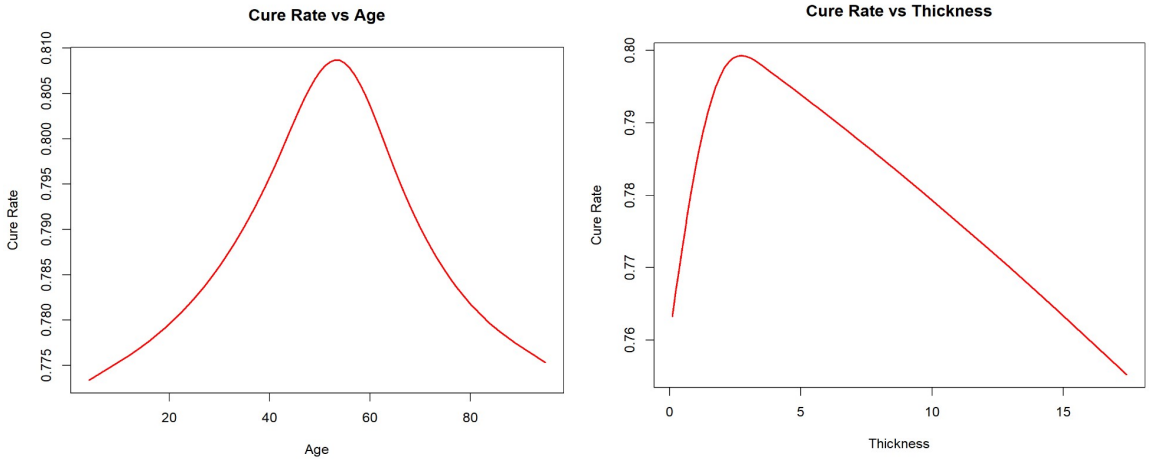
approximately 43.9% of cases. This dataset provides a robust platform for analyzing survival outcomes based on various clinical and demographic factors.

By analyzing this dataset, we aim to contribute further to the understanding of prognostic indicators and to refine predictive models for melanoma survival. This analysis will enhance the ability to tailor treatments and improve patient outcomes. The dataset's comprehensive nature allows for detailed exploration of survival outcomes and the factors influencing them, aiding in the development of more effective clinical strategies and interventions for managing



(a) (b)

Figure 4.2 Effect of covariate



(a) (b)

Figure 4.3 Marginal cure rate plot

malignant melanoma.

Figure 3.3 shows the survival probability over time, with a clear decrease indicating the occurrence of events (deaths) throughout the study period. The extended probability plateau indicates a period during the study in which no deaths events from melanoma are

observed. The survival probability remains constant during this plateau period, reflecting an increased number of censors. The step-like pattern and the widening confidence intervals indicate greater uncertainty in the survival estimates as fewer individuals remain in the study over time. This plot serves as a fundamental representation of the survival experience within the cohort, highlighting the importance of temporal dynamics in understanding patient outcomes.

The subsequent plots in Figure 3.4 explore the effects of age and tumor thickness on survival and cure rates, revealing intricate nonlinear relationships. The effect of age on melanoma deaths shows a U-shaped curve, with the highest risk in older and younger patients and a lower risk for average age. Similarly, the effect of thickness shows a complex pattern, with the less substantial risk around the mean tumor thickness and most influences as thickness increases or decreases. These non-linearities emphasize the necessity of considering the full range of covariate values to accurately capture their impact on survival and cure rates.

Lastly, the cure rate graphs in Figure 3.5 reveal critical insights into how age and tumor thickness influence treatment success. The highest cure rates are observed for individuals with average age and slightly below-average tumor thickness. The cure rate declines as age deviates from the mean in either direction and as tumor thickness increases, indicating that both younger and older individuals, as well as those with higher thickness values, experience poorer outcomes. These findings underscore the importance of personalized treatment strategies that account for age and tumor thickness to improve patient prognosis and optimize cure

rates.

CHAPTER 5

Conclusion and Future Works

The present study provides a comprehensive approach to modeling survival data using a mixture cure model framework. By leveraging sophisticated statistical techniques and non-parametric methods, our approach demonstrates substantial potential for accurate parameter estimation and modeling of survival outcomes, particularly in the presence of cure fractions. Here, we discuss the implications of our findings, methodological strengths, limitations, and potential future directions for research.

One of the key strengths of our approach is the incorporation of B-splines for basis expansion of covariates additively, allowing for flexible modeling of covariate effects. This is particularly advantageous when dealing with complex, non-linear relationships between covariates and survival outcomes. The simulation study, which iterated the process 1000 times with a sample size of 200, robustly validated the model's performance, ensuring reliability and generalizability of the results.

Our simulation results underscore the efficacy of the proposed model in accurately estimating the cure rate and survival distribution among uncured individuals.

The application of our model to the "Survival from Malignant Melanoma" dataset illustrates its practical utility in real-world clinical scenarios. The analysis of the data set revealed significant information on the prognostic factors influencing melanoma survival, highlighting the impact of tumor thickness, age, ulceration status, and other covariates on patient outcomes. These findings align with the existing literature, strengthening the validity

and relevance of the model in clinical decision making and public health interventions.

In summary, this study presents a robust and flexible framework for survival analysis using cure models. The integration of B-splines for covariate expansion and optimization for parameter estimation demonstrates significant promise in accurately modeling survival data. While limitations exist, the model's strengths and practical utility in clinical applications underscore its potential as a valuable tool in survival analysis. Future research endeavors aimed at addressing the identified limitations and exploring advanced methodologies could further enhance the model's efficacy and broaden its applicability in various fields of study.

Appendices

A Derivation of the Log-Likelihood Function

The likelihood function is given by:

$$L = \prod_{i=1}^{n} \left\{ [f_T(Y_i; \theta(X_i), \gamma)]^{\Delta_i} [S_T(Y_i; \theta(X_i), \gamma)]^{1-\Delta_i} \right\}^{I(Y_i<\infty)} [S_T(\infty; \theta(X_i), \gamma)]^{I(Y_i=\infty)} \tag{1}$$

Taking the logarithm of the likelihood function, we get the log-likelihood:

$$\log L = \sum_{i=1}^{n} I(Y_i < \infty) \left[\Delta_i \log f_T(Y_i; \theta(X_i), \gamma) + (1 - \Delta_i) \log S_T(Y_i; \theta(X_i), \gamma)\right] + I(Y_i = \infty) \log S_T(\infty; \theta(X_i), \gamma \tag{2}$$

Assuming no observations are censored beyond a finite time ($I(Y_i = \infty) = 0$), we simplify:

$$\log L = \sum_{i=1}^{n} \left[\Delta_i \log f_T(Y_i; \theta(X_i), \gamma) + (1 - \Delta_i) \log S_T(Y_i; \theta(X_i), \gamma)\right] \tag{3}$$

We express f_T and S_T in terms of the cumulative distribution function (CDF) F_T and the hazard function λ_T:

$$f_T(t; \theta, \gamma) = \frac{d}{dt} F_T(t; \theta, \gamma) \tag{4}$$

$$S_T(t; \theta, \gamma) = 1 - F_T(t; \theta, \gamma) \tag{5}$$

$$\lambda_T(t; \theta, \gamma) = \frac{f_T(t; \theta, \gamma)}{S_T(t; \theta, \gamma)} \tag{6}$$

Substituting these relationships, the log-likelihood becomes:

$$\log L = \sum_{i=1}^{n} \left[\Delta_i \log(\lambda_T(Y_i; \theta(X_i), \gamma) S_T(Y_i; \theta(X_i), \gamma)) + (1 - \Delta_i) \log S_T(Y_i; \theta(X_i), \gamma)\right] \tag{7}$$

Since $\lambda_T(t)S_T(t) = f_T(t)$, we simplify:

$$\log L = \sum_{i=1}^{n} [\Delta_i \log \lambda_T(Y_i; \theta(X_i), \gamma) + \log S_T(Y_i; \theta(X_i), \gamma)] \tag{8}$$

Using $\lambda_T(t; \theta, \gamma) = \theta(X_i) f(Y_i; \gamma)$ and $S_T(t; \theta, \gamma) = e^{-\theta(X_i)F(Y_i;\gamma)}$:

$$\log L = \sum_{i=1}^{n} [\Delta_i \log(\theta(X_i) f(Y_i; \gamma)) - \theta(X_i) F(Y_i; \gamma)] \tag{9}$$

Breaking this down, we obtain:

$$\log L = \sum_{i=1}^{n} \{\Delta_i [\log(\theta(X_i)) + \log f(Y_i; \gamma)] - \theta(X_i) F(Y_i; \gamma)\} \tag{10}$$

B Generating Survival Time

The survival time T_i is generated using the inverse cumulative distribution function (CDF) approach:

$$T_i = F_T^{-1}(U_i|X_i) \tag{11}$$

where U_i is a uniform random variable in the interval (0, 1), and X_i represents the covariates for the i-th individual.

C Inverse Transformation Method

We can prove the inverse transform method. Let $Y = F(X)$. Since Y is a random variable, it has a CDF, which we can denote $G(y)$. By definition:

$$G(y) = P(Y \leq y)$$

Since $Y = F(X)$:

$$G(y) = P(F(X) \leq y)$$

Since X is a continuous random variable, its CDF is continuous. Therefore, we can apply the inverse, F^{-1}, to both sides of the inequality:

$$G(y) = P(F^{-1}(F(X)) \leq F^{-1}(y))$$

What is $F^{-1}(F(X))$? Simply, X:

$$G(y) = P(X \leq F^{-1}(y))$$

Notice that we have an expression of the form $P(X \leq x)$, where $x = F^{-1}(y)$. We know, by definition, $F(x) = P(X \leq x)$, so:

$$G(y) = F(F^{-1}(y)) = y$$

In summary, the CDF of Y is $G(y) = y$. If we take the derivative of the CDF to get the PDF, we see that $g(y) = 1$. Let's remember the PDF for a uniform random variable:

D Exponential Example

Consider the Exp(λ) distribution, which has the following CDF:

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

Let's set $F(X) = U$ and solve for X:

$$U = 1 - e^{-\lambda X}$$

$$U - 1 = -e^{-\lambda X}$$

$$\ln(U - 1) = -\lambda X$$

$$\frac{\ln(U - 1)}{-\lambda} = X$$

Also, we know that the expression $U - 1$ is itself uniform, so we can simplify:

$$X = \frac{\ln(U)}{-\lambda}$$

REFERENCES

[1] Joseph Berkson and Robert P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[2] Mu Chen and Peirong Du. A unified cure rate model based on the generalized odds-rate class. *Statistics in Medicine*, 37(24):3446–3457, 2018.

[3] James R. Cook and Victor DeGruttola. Flexible regression models for survival data: with application to HIV infection in the Swiss HIV Cohort Study. *Statistics in Medicine*, 18(11):1489–1501, 1999.

[4] Yuting Feng, Ruiyan Duan, and Jianguo Sun. Nonparametric comparison of survival functions based on interval-censored data with unequal censoring. *Statistics in Medicine*, 36(12):1895–1906, 2017.

[5] Leonid Hanin and Li-Shan Huang. Identifiability of cure models revisited. *Journal Name*, Volume:Pages, Year. Department of Mathematics, Idaho State University, 921 S. 8th Avenue, Stop 8085, Pocatello, ID 83209-8085, USA; Institute of Statistics, National Tsing-Hua University, 101, Section 2, Kuang-Fu Road, Hsin-Chu 30013, Taiwan.

[6] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

[7] E. M. Laska and M. J. Meisner. Nonparametric estimation and testing in a cure model. *Biometrics*, 48(4):1223–1234, December 1992.

[8] L.-H. Lin and L.-S. Huang. Promotion time cure model with local polynomial estima-

tion. *Statistics in Biosciences*, To Appear, 2024.

[9] Yanyuan Ma and Geng Yin. Cure rate model based on a generalized transformation class. *Journal of the American Statistical Association*, 103(482):637–645, 2008.

[10] Saurabh Pal and Emmanuel Aselisewine. Advanced survival models for analyzing long-term survivor data. *Journal of Statistical Research*, 57(2):123–145, 2023.

[11] L. Wang, P. Du, and H. Liang. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68(3):726–735, 2012.

[12] S. Wang, C. Wang, and J. Sun. An additive hazards cure model with informative interval censoring. *Lifetime Data Analysis*, 27(2):244–268, – 2021.

[13] Wenyu Xie and Weixin Yu. A flexible semiparametric cure rate model based on an em algorithm. *Computational Statistics & Data Analysis*, 143:106839, 2020.

[14] Andrei Y. Yakovlev and Alexey D. Tsodikov. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, New Jersey, 1996.

[15] Donglin Zeng and DY Lin. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 93(3):627–640, 2006.

[16] J. Zhang and M. Davidian. A semiparametric approach to cure rate models with flexible hazard functions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4):965–983, 2020.

[17] Jialiang Zhang and Jianguo Sun. Flexible nonparametric estimator for assessing cure rates. *Journal of the American Statistical Association*, 110(512):1189–1198, 2015.

[18] Y. Zhang and Y. Wu. Flexible cure rate models with nonparametric baseline hazard

functions. *Statistical Modelling*, 18(3):242–263, 2018.