12-16-2019

Novel Statistical Methods for Censored Medical Cost and Breast Cancer Data

Guanhao Wei

NOVEL STATISTICAL METHODS FOR CENSORED MEDICAL COST AND BREAST CANCER DATA

by

GUANHAO WEI

Under the Direction of Gengsheng Qin, PhD

ABSTRACT

Recent studies show that appropriate statistical analysis of cost data may lead to more cost-effective medical treatments, resulting in substantial cost savings. Even though the mean value is publicly accepted as a summary of medical costs, however, due to heavy censoring and heavy skewness, mean will be affected much by missing or extremely large values. Therefore, quantiles of medical costs like the median cost are more reasonable summaries of the cost data. In the first part of this dissertation, we first propose to use empirical

likelihood (EL) methods based on influence function and jackknife techniques to construct confidence regions for regression parameters in median cost regression models with censored data. We further propose EL-based confidence intervals for the median cost with given covariates. Compared with existing normal approximation-based confidence intervals, our proposed intervals have better coverage accuracy.

In the real world, there is a large proportion of patients having zero costs. In the second part, we propose to use fiducial quantity and EL-based inference for the mean of zero-inflated censored medical costs applying the method of variance estimates recovery (MOVER). We also provide EL-based confidence intervals for the upper quantile censored medical costs with many zero observations. Simulation studies are conducted to compare the performance between proposed EL-based methods and the existing normal approximation-based methods in terms of coverage probability. The novel EL-based methods are observed to have better finite sample performances than existing methods, especially when the censoring proportion is high.

In the third part of this dissertation, we focus on evaluating breast cancer recurrence risk. For early-stage cancer tumor recurrence study, existing methods do not have an overall powerful survival prediction ability. Preliminary studies show that centrosome amplification has a strong latent correlation with tumor progression. As a result, we propose to construct a novel quantitative centrosome amplification score to stratify patients' cancer recurrence risk. We prove that patients with higher centrosome amplification score will have a significantly higher probability to experience cancer recurrence given all demographic conditions, which could provide a potent reference for the future developing trend of early-stage breast cancer.

INDEX WORDS: Censored medical cost, Confidence region, Empirical likelihood, Jackknife, Median regression, Zero cost, Fiducial quantity, MOVER, Centrosome amplification, Breast cancer

NOVEL STATISTICAL METHODS FOR CENSROED MEDICAL COST AND BREAST CANCER DATA

by

GUANHAO WEI

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

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2019

NOVEL STATISTICAL METHODS FOR CENSROED MEDICAL COST AND BREAST CANCER DATA

by

GUANHAO WEI

Committee Chair: Gengsheng Qin

Committee: Yi Jiang

Jing Zhang

Ruiyan Luo

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2019

DEDICATION

This dissertation is dedicated to all my family for their love, support, and encouragement during my studies. This work is especially dedicated to my parents Yong Wei and Ping Liu for giving me so much with their sacrifices.

ACKNOWLEDGEMENTS

This dissertation work would not have been possible without the support of many people. I want to express my gratitude to my advisor Dr. Gengsheng Qin. His outstanding guidance, patience, and mentorship throughout my PhD program. I am especially thankful for his support during my dissertation research. His insight into the various methodological approaches for my research and his intensive feedback on the several versions of my dissertation draft has been invaluable. I appreciate Dr. Qin's time and dedication in working with me to understand the existing studies on medical cost estimates and on the mathematical proofs associated with my work.

I also would like to express my gratitude to my committee members, Dr. Yi Jiang, Dr. Jing Zhang, and Dr. Ruiyan Luo, for asking heuristic questions regarding my research and for providing me with helpful feedback. I appreciate their time and commitment to my work.

For my research related to medical cost study, I am really grateful for GSU alumni, Dr. Jenny V. Jeyarajah, for her previous study and detailed explanation in medical cost analysis. Without her help, I will not be able to solve theoretical or programming problems so smoothly. Also, I want to thank Dr. Andrew Willan from University of Toronto for providing me with the Canadian Implantable Defibrillator Study data.

For my research related to cancer study, I would like to express my gratitude to professors Dr. Remus Osan and Dr. Ritu Aneja for supporting me to enroll in the precious GSU Molecular Basis of Disease fellowship program. Being an MBD member brings a lot of opportunities to get involved with most frontier clinical trial projects related to cancer study, which helped me practiced my biostatistics modeling and visualization skills a lot. What's more, my colleagues and friends from biology department, Dr. Karuna Mittal and Dr. Sergey Klimov, provided me much help with their great expertise when we worked cooperatively in cancer clinical and genetic data modeling areas.

During my PhD study period at Georgia State University, I want to thank my seniors

Dr. Jun Xia, Dr. Chenxue Li, and Dr. Bing Liu for kindly providing me with their experience of growing to be a better fit a role as a teacher and researcher. I would love to appreciate my study mates, Yan Hai, Xinjie Hu, and Tianchi Zhang for their encouragement and cooperation, which keeps me release my stress while under hard working.

Finally, I would express my great love to my family members for their love, support, patience, and encouragement throughout my studies. My parents Yong Wei and Ping Liu provide me enough financial support and mental comfort, which helped me keep going further and insist on doing research. Last but not least, I want to thank my wife Huaxu Yang for her endless love, encouragement, and selfless contribution, which is tremendous support throughout my entire graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- BT - Bang and Tsiatis's Simple Weighted Estimator for Mean Cost
- CAS - Centrosome Amplification Score
- CIDS - Canadian Implantable Defibrillator Study
- CIs - Confidence Intervals
- DCIS - Ductal Carcinoma In Situ
- EF - Efficient Estimator for Median Cost
- EL - Empirical Likelihood
- HK - Hasan and Krishnamoorthy's Fiducial Quantity
- IFEL/ELI - Influence Function based Empirical Likelihood
- JEL - Jackknife Empirical Likelihood
- LZT - Li, Zhou and Tian's Fiducial Quantity
- MOVER - Method Of Variance Estimates Recovery
- NA - Normal Approximation
- SW - Simple Weighted Estimator for Median Cost
- ZT - Zhao and Tian's Efficient Estimator for Mean Cost

PART 1

INTRODUCTION

1.1 Medical Cost with Censored Observation

The National Health Expenditure Accounts (NHEA) reported that U.S. healthcare spending grew 3.9% in 2017, reaching $3.5 trillion or $10,739 per person. As a share of the nation's Gross Domestic Product, healthcare spending accounts for 17.9% and is projected to reach 25% by 2037 (Chen *et al.*,2016[9]). As a result, healthcare costs are a major social and economic concern. For example, in 2013, Steven Brill wrote an article appearing in *Time* magazine about 'How outrageous pricing and egregious profits are destroying our health care.'(Topol, 2015[59]). Due to a well established Medicare system, insurance companies, private employers, federal or state governments, and public charity associations will pay 80-90% of total health care costs. So few patients will care about what they have been charged. Consequently, missed prevention, unnecessary services, inefficiently delivered services, excessive high price, high administrative fees, and fraud claims waste a huge amount of American health care spending. Given these healthcare cost realities and the structure of the US healthcare system, we should be more aware of the importance of medical cost data analysis.

One of the goals in such analysis is to identify factors that affect medical costs and examine how these factors influence medical costs. Such identification can lead to more effective policymaking by health care providers and medical insurance companies. Therefore, a regression for mean or quantile medical cost on policy specific covariates is useful.

A number of statistical methods have been proposed for such cost regression models. Bang and Tsiatis (2000)[1], Bang and Zhao (2012)[3], Johnson (2015)[30], Lin (2000)[36] and Willan *et al.* (2005)[64] proposed linear models for estimating censored total costs.

However, since medical cost data are right skewed, the mean cost is highly sensitive to

outliers. As a result, regression on mean medical costs may not be so informative. To get a more comprehensive picture, it is more reasonable to estimate the median or other quantiles of medical costs with covariates.

Existing methods for median cost regression models with censored data, proposed by Bang and Tsiatis (2002)[2], focus on finding consistent and asymptotically normal estimators for the median regression parameters. Even though it is feasible to derive confidence intervals for the median cost of a patient over a certain period based on the asymptotic normality of the regression estimators, the normal approximation-based intervals can have poor coverage accuracy if the cost data is highly skewed and heavily censored. We note that the normal approximation-based confidence regions for the regression coefficients can have lower coverage probabilities than the desired nominal level when the cost data are moderately or heavily censored. Furthermore, it is analytically complicated and hard to estimate variance accurately. To overcome those drawbacks, we propose to use Empirical Likelihood-based inference for the median medical costs with covariates.

1.2 Zero Inflated Cost Data

As noted above, excessive use of medical and health care expenditures masks various cost inefficiencies in the health care system. Therefore, an assessment of the cost is well understood and therefore important for developing a treatment plan with appropriate cost considerations[28].

However, the statistical analysis of cost data is complicated due to following real-world data characteristics: (1) a large proportion of patients have zero costs because these patients did not take any treatments; (2) nonzero medical costs are highly skewed to the right with unknown distributions; (3) the percentage of censored observations is usually high.

Topics about inferences on zero-inflated data were investigated by a lot of scholars in the past few years. Zhou *et al.*(2000)[71] proposed methods to construct confidence intervals for the mean of diagnostic test charge data containing a lot of zeros by using normal approximation based approaches. Tian(2005)[58] and Hasan and K. Krishnamoorthy(2018)[24]

proposed methods to make inferences for the mean value of zero-inflated lognormal data. In general, their ideas are based on generalized pivotal quantities by decomposing data into zero and non-zero parts. Chen *et al.*(2003)[8] and Chen *et al.*(2010)[7] also proposed empirical-likelihood or pseudo-likelihood based inferences for the mean of a population containing many zero values under various conditions. However, this analysis uses only complete data without considering censored observations.

For nonzero medical costs, currently, a number of methods have been proposed in the literature with different types of assumptions. However, as Young[65] and others have pointed out, the main challenge in analyzing censored cost data is that even if the time to death and the time to censoring are independent, the total cost at censoring will vary with the total cost at the time of death. Hence, standard survival analysis techniques, which assume independent censoring and treat censored costs as censored survival times, cannot be directly used for the analysis of censored cost data[28].

Without assuming that the censoring time is discrete, Bang and Tsiatis(2000)[1] proposed a simple weighted estimator that uses the final cost for uncensored patients only. Later, a complicated estimator, proposed by Zhao and Tian(2001)[66], creatively used additional information based on the censored and uncensored patients' cost history[27]. Zhao and Tian's estimator is thus more efficient than Bang and Tsiatis' simple weighted estimator. Lin *et al.*(2003)[37] proposed a nonparametric estimator for mean medical cost considering discrete censoring time, and their approaches utilized patients' final total cost and cost history.

Moreover, it is a well-known fact that the cost data is generally highly skewed, and the mean medical cost is proved to be highly influenced by extremely large costs. Theoretically, Sherwood *et al.*[53] showed that when marginal effects vary across the conditional distribution, focusing on the marginal effects at the mean value may substantially distort information of interest at the tails. And also, a weak relationship between a risk factor and the mean medical cost does not preclude a stronger relationship at the upper or lower quantiles of the conditional distribution. Therefore, mean medical cost alone cannot offer complete information for cost analysis[28].

A complementary approach is the use of quantile costs, which have long been used for characterizing economic data such as housing prices, mortgages, and incomes[33]. Quantile-based methods are applicable to medical research because it provides more specific and comprehensive distributional information than a method based on the mean[57]. For example, for patients with chronic diseases, the longer they live, the more money they will spend on medication. In this scenario, any increasing or decreasing trend in quantile costs, especially upper quantile costs, which indicates upper cost limits, can affect patients' economic decisions based on their treatments and possible lifetime. Healthcare providers may also convincingly predict potential upper bound of overall cost change in the future to provide better healthcare plans given an available budget. Hence, estimating the quantiles of medical costs and comparing these costs among different groups is important to cost data analysis[32] in addition to mean cost alone.

In 2012, Zhao *et al.*[68] proposed non-parametric estimators for median costs with censored data. Their normal approximation-based confidence intervals for the median medical cost can have poor finite sample performance when the data are severely skewed. And also, those existing methods don't consider observations with zero cost, which is not ignorable.

In this part of the dissertation, we focus on interval estimation for the mean and upper quantiles of zero-inflated medical cost with censored data. We construct fiducial confidence intervals for zero cost proportion and EL-based confidence intervals for the mean of non-zero censored medical costs. Then we use the method of variance estimates recovery (MOVER) approach to making inferences for the overall mean of zero-inflated costs. In addition, we construct EL-based confidence intervals for the upper quantiles of medical costs while assuming zero cost proportion is a binomial proportion to fully use all the data information.

1.3 Empirical Likelihood Methods

Empirical likelihood (EL), introduced by Owen (1988, 1990, 2001)[47, 46, 48], is a powerful pure nonparametric method. EL methods allow us to employ likelihood methods without having to pick a parametric family for the data; the EL methods are not restricted

to symmetric confidence interval/region for the parameters, but instead, its shapes are determined by data; and the EL methods allow for confidence interval or confidence region construction without a variance estimator. The advantages of EL-based methods over normal approximation-based methods have been well-recognized (Hall and La Scala, 1990)[22].

According to the characteristics of medical cost data, which is skewed and censored, researchers believe that EL-based methods are especially suitable for making inferences for medical costs. Zhou *et al.*[70] proposed a plug-in empirical likelihood method for constructing a confidence region for a vector of regression parameters belonging to a censored cost regression model and a confidence interval for the expected total cost of a patient with given covariates. Jeyarajah and Qin[27] proposed influence function-based empirical likelihood method for mean medical cost with censored data. They demonstrated that the EL-based methods outperform the existing methods when analyzing highly skewed and heavily censored cost data. For those reasons, we want to make further progress on medical cost analysis utilizing the advantages of empirical likelihood.

1.4 DCIS Cancer

Ductal carcinoma in situ, presence of abnormal cells inside a milk duct in the breast, is considered a pre-invasive or earliest form of breast cancer. About twenty percent of screen-detected breast cancers can be classified as ductal carcinoma in situ. DCIS is not life-threatening, but having DCIS can increase the risk of developing invasive breast cancer later on. ([55, 39]). According to Page *et al.* (1982)[49], around 20 to 53% of women with untreated DCIS will progress to invasive breast cancer within a period of greater than or around 10 years. Due to the reason that we don't have a robust method to control latent DCIS progression into invasive cancer, breast-conserving surgery combined with radiation or surgery is primary treatment plans([16]). However, for such patients with lumpectomy or breast conservation surgery treatment, there are still around thirty percent of those experienced local recurrences ([4, 17]). And after initial treatment, if breast cancer does come back, there are as high as 50% of chance that DCIS will become invasive. That's why the

investigation on DCIS recurrence prediction is an important topic for early-stage cancer study. The major research interest in this paper is to develop prognostic biomarkers that can stratify DCIS patients based on their recurrence risk.

1.5 Centrosome Amplification

The centrosome is a small organelle normally localized at the periphery of the nucleus ([13, 18]). It is considered as the major microtubule-organizing center of animal cells([14]). And cases with an increased number of centrosomes at mitosis be greater than 2, the critical value for normal cells is publicly accepted as centrosome amplification numerically. Also, recent studies show that an unusually large volume of centrosome will also be recognized as another important feature of amplification([44]), which can be considered as structural amplification. It is well known that amplified centrosomes are the cause of abnormal tumor progression heterogeneity([19, 40]), which is considered as one vital characteristic in a growing list of human cancers and is a potential future hallmark of cancer cells ([6]). Additionally, centrosome amplification occurs within pre-invasive cancer including DCIS, suggesting that centrosome amplification is an early event in the formation of tumor cells([13]). Naturally, it is meaningful to investigate the association between centrosome amplification and DCIS tumor progression status.

A preliminary study by simply using the percentage of centrosomes that have amplification show that the extent of aberrations may differ in DCIS cases with or without recurrence. However, the prognostic value of centrosome amplification has remained unexplored for clinical application, as there is no methodology available for the rigorous quantization of centrosome amplification phenotype. Also, it is unclear whether a reliable prognostic value of centrosome amplification can be used in numerical or structural centrosome amplification. Additionally, for the features of centrosome amplification such as frequency (percentage of cells showing amplified centrosomes) and severity (how abnormal the number or volume of centrosomes), we want to be much more clear about whether one or the combination of the two features is more informative in recurrence risk prediction.

As a result, in this part of the dissertation, we proposed a novel methodology to quantify both numerical and structural centrosomal aberrations in clinical tissue samples. Our analytical procedure creatively generates a summary of full centrosome amplification information to predict the risk of local recurrence after a lumpectomy. We have developed an algorithm that generalizes the frequency and severity of numerical and structural centrosome amplification in clinical samples and computes a centrosome amplification score for each sample. Our discovery results show that centrosome amplification score (CAS) is a promising measurement that may improve treatment recommendations and allow identification of patients at high or low risk of recurrence. CAS demonstrates the highest effect on recurrence-free survival probability compared to commonly used clinicopathological variables such as grade, age, and comedo necrosis.

1.6 Cox Proportional-Hazards Model

In the third part of the dissertation, we need to illustrate the association between patients' recurrence free survival risk and one or more clinical predictors. A commonly used model is the Cox proportional hazard model[11]. For our breast cancer topic, the model examines how related factors of interest influence the hazard risk of cancer recurrence.

The general form of Cox model is

$$h(t, \mathbf{X}) = h_0(t) exp\{\sum_{i=1}^{p} \beta_i X_i\}$$

where t represents the survival time, $h_0(t)$ is the baseline hazard function, $h(t, \mathbf{X})$ is the hazard function determined by p dimensional vector $\mathbf{X} = (X_1, .., X_p)$.

Then the estimated hazard ratio between two sets of predictors will be defined as $\hat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = exp\{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\}$. A hazard ratio greater than 1 indicates that predictors are positively associated with the event probability, and vice versa. In our application problem, for example, if the hazard ratio is 2, then the rate of recurrence event in treatment group is twice the rate in the reference group.

In this dissertation, we also use Cox model to predict survival probabilities. The relationship between survival and hazards functions can be presented as

$$\hat{S}(t, \mathbf{X}) = \hat{S}_0(t)^{exp\{\sum_{i=1}^{p} \beta_i X_i\}}$$

where $\hat{S}_0(t) = exp(-\hat{\Psi}_0(t))$, and $\hat{\Psi}_0(t)$ is the Breslow estimator for the cumulative baseline hazard $\Psi_0(t) = \int_0^t h_0(s)ds$.

1.7 Brief Summary

This dissertation is organized as follows.

In Part 2, we present the notations and normal approximation-based confidence regions for median medical cost regression with censored data. Then we propose an influence function-based EL method and a jackknife EL method for median medical cost regression model with censored data. Next, we present simulation studies to compare finite sample performance of the proposed EL confidence regions with that of the normal approximation-based confidence regions. A real data example is used to illustrate the proposed methods. Proof of the theorem for influence function-based EL method will be given at the end of Part 2.

In Part 3, we first introduce the fiducial quantity based confidence intervals for the parameter of zero cost proportion. Next, we present EL-based confidence intervals for the mean of nonzero medical costs. Then, we apply the MOVER approach to construct confidence intervals of overall zero-inflated mean costs with censored data. We also propose EL-based methods for the upper quantiles of medical costs with censored and zero costs. The performance of those inferential methods was compared in simulation studies and a real example.

In Part 4, we first describe the existing research that is related to DCIS recurrence-free survival. Next, we describe how to prepare clinical sample records and introduce notations and formulas in the model we use. After that, we apply univariate and multivariate statistical

analyses to diagnose the ability of CAS and other related demographic variables that could affect patients' recurrence-free survival differences. Finally, we evaluate the robustness of random small sample performance that CAS can do to make survival status prediction.

The final conclusion will be presented in Part 5.

PART 2

EMPIRICAL LIKELIHOOD BASED INFERENCES FOR MEDIAN MEDICAL COST REGRESSION MODELS WITH CENSORED DATA

The first part of this dissertation mainly focuses on how to construct confidence intervals for median medical cost with censored observations with given covariates using empirical likelihood-based approaches. It is organized as follows, in Section 2.1, we describe the notations and existing estimation equations. In Section 2.2, we introduce the empirical likelihood methods based on influence functions and jackknife techniques that are used to construct confidence regions for regression coefficients. In Section 2.3, we perform simulation studies and compare the coverage probabilities we achieved from different methods in different scenarios. In Section 2.4, we introduce a numerical approach used to build confidence intervals for the median medical costs with given covariates. In Section 2.5, we demonstrate the application of our method to a real-world dataset. In Section 2.6, we present the conclusion and discussion for future work.

2.1 Existing Methods

2.1.1 Notations

We first define the survival time of a patient is T, accordingly, the overall survival function for a given observation is $S(u) = P(T \geq u)$. And we can use C to denote the censoring time of a patient with survival function $K(u) = P(C > u)$. Then, we have the assumption that patients' survival time and the total medical cost does not depend on censoring time provided at random. Especially when the censoring occurs due to termination of a study, which is referred to as administrative censoring, such independence usually exists. Censoring can also occur when a patient leaves from a study, lost to follow up or administrative reasons such as a patient's follow-up time is less than the survival time (Bang and Zhao, 2012[3]).

Due to the existence of censoring, the medical costs records for all the patients are not observed completely. The observed survival time of a patient is denoted by $X_i = min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$ will be used as an indicator of censoring status. If censoring happens, it is difficult to estimate the median medical cost for the whole survival time without any restriction. Hence, it is more practical to consider the cost incurred from the beginning to a maximum fixed L units of time for a given case, where a reasonable amount of complete data T_i's are available over the time period $[0, L]$. Let $M_i(u)$ denote the total medical cost for a patient cumulatively from study time 0 to time u, and let $\mathbf{Z}$ be a vector of covariates. For simplicity, we define $M_i \equiv M_i(X_i)$ as observed total cost for each patient at the end of the lifetime or study period.

The observed cost data with covariates for n patients are $\big\{(X_i, \Delta_i, \mathbf{Z_i}, M_i(u)) : 0 \leq u \leq X_i, i = 1, \cdots, n\big\}$.

In this paper, we consider the following median cost regression model:

$$M = \beta' \mathbf{Z} + \epsilon, \tag{2.1}$$

where $\beta = (\beta_0, \beta_1, .., \beta_p)'$ is a $(p+1)$-dimensional vector of regression coefficients, ϵ has a distribution that is absolutely continuous, and $P(\epsilon < 0|\mathbf{Z}) = 1/2$. Our goal is to make inferences for β and estimate the median medical cost given a covariate $\mathbf{Z}$ based on the observed cost data. Model (2.1) can be generalized to a quantile cost regression model with $P(\epsilon < 0|\mathbf{Z}) = q$ for any $q \in (0, 1)$.

2.1.2 Simple Estimation for the Median Cost Regression Model with Censored Data

Based on the inverse probability weighting method, the following estimating equation can be used to estimate β in the median cost regression model:

$$Q_n(\beta) = \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(T_i)} \mathbf{Z_i} \{I(M_i < \beta' \mathbf{Z_i}) - 1/2\} \approx 0, \tag{2.2}$$

where $\hat{K}(u)$ is the Kaplan-Meier estimator for the survival function of the censoring time C.

The solution $\hat{\beta}$ to (2.2) is a simple estimator for β. This estimator was shown to be consistent and asymptotically normal (Bang and Tsiatis, 2002), i.e.,

$$n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Lambda), \tag{2.3}$$

where $\Lambda = A^{-1}\Gamma(A^{-1})'$, $A = -E\{\mathbf{Z}\mathbf{Z}'f(0|\mathbf{Z})\}$ with $f(0|\mathbf{Z})$ being the conditional density of ϵ at $\epsilon = 0$ given $\mathbf{Z}$, and Γ is a $(p+1) \times (p+1)$ unknown matrix.

Since the asymptotic variance matrix Λ of $\hat{\beta}$ is still unknown, we have to estimate it in order to make inference for β. Let $\mathbf{B_i}(\beta) = \{I(M_i < \beta'\mathbf{Z_i}) - 1/2\}\mathbf{Z_i}$, $N_i^c(u) = I(X_i \leq u, \Delta_i = 0)$, and $G(\mathbf{W}, u) = E[\mathbf{W}I(T \geq u)]/S(u)$ for any random vector $\mathbf{W}$. Let $\hat{f}_n(\cdot|\mathbf{Z})$ be a consistent density estimator of ϵ given $\mathbf{Z}$ (e.g., a conditional kernel density estimator). Then A can be consistently estimated by

$$\hat{A} = -\frac{1}{n}\sum_i \mathbf{Z_i}\mathbf{Z_i}'\hat{f}_n(0|\mathbf{Z_i}),$$

and Γ can be consistently estimated by

$$\begin{aligned}\hat{\Gamma} &= \frac{1}{4n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(T_i)}\mathbf{Z_i}\mathbf{Z_i}' - [\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{K}(T_i)}\mathbf{B_i}(\hat{\beta})]^{\otimes 2} \\ &\quad + \frac{1}{n}\int_0^L \frac{dN^c(u)}{\hat{K}^2(u)}[\hat{G}(\mathbf{B}(\hat{\beta})^{\otimes 2}, u) - \hat{G}^{\otimes 2}(\mathbf{B}(\hat{\beta}), u)],\end{aligned} \tag{2.4}$$

where $\mathbf{B}(\beta) = \{I(M < \beta'\mathbf{Z}) - 1/2\}\mathbf{Z}$, $\hat{G}(\mathbf{W}, u) = n^{-1}\hat{S}^{-1}(u)\sum_{i=1}^n \Delta_i\mathbf{W_i}I(T_i \geq u)/\hat{K}(T_i)$ with $\mathbf{W_i}$'s being copies of $\mathbf{W}$, and $\hat{S}(u)$ is the Kaplan-Meier estimator for the survival function $S(u)$.

Therefore, Λ can be consistently estimated by $\hat{\Lambda} = \hat{A}^{-1}\hat{\Gamma}(\hat{A}^{-1})'$. A normal approximation-based confidence region for β can be constructed using (2.3) and $\hat{\Lambda}$.

2.1.3 Improved Estimation for the Median Cost Regression Model with Censored Data

Since the inverse probability weight-based estimating equation (2.2) uses the cost history of uncensored patients, it is not efficient when many patients are censored. To improve the efficiency of the estimation method, Bang and Tsiatis (2002)[2] also proposed an improved estimation function for β which utilizes the cost histories of available data:

$$Q_n(e,\beta) = Q_n(\beta) + \gamma^{opt} \sum_{i=1}^{n} \int_0^L \frac{dN_i^c(u)}{\hat{K}(u)} \{e\{M_i^H(u)\} - \hat{G}^*(e\{M^H(u)\},u)\}, \tag{2.5}$$

where, for each u, $e\{M_i^H(u)\}$ is a vector of the cost history $e_j\{M_i^H(u)\}$ up to time u, and $\hat{G}^*(e\{M^H(u)\},u)$ is a vector of $\hat{G}^*(e_j\{M^H(u)\},u)$, and the J-dimensional vector γ^{opt} can be obtained by minimizing the variance of $(y_i - \gamma_1\omega_{i1} - \cdots - \gamma_J\omega_{iJ})$ for the ith individual with

$$\begin{aligned} y_i &= \int_0^L dM_i^c(u) K^{-1}(u) \{\mathbf{B_i}(\beta) - G(\mathbf{B}(\beta),u)\}, \\ \omega_{ij} &= \int_0^L dM_i^c(u) K^{-1}(u) \times \{e_j\{M_i^H(u)\} - G(e_j\{M^H(u)\},u)\}, \end{aligned}$$

where $M_i^c(u) = N_i^c(u) - \int_0^u \lambda^c(t) Y_i(t) dt$, $Y_i(t) = I(X_i \geq t)$, and $\lambda^c(t)$ is the hazard function of the censoring distribution. For simplicity, let $e_j\{M_i^H(u)\} = M_{ij}(u)$ with $M_{ij}(u)$ being the cost incurred in the subinterval $[t_{j-1}, min(t_j,u)]$, and

$$\hat{G}^*(e\{M^H(u)\},u) = \sum_{i=1}^{n} e\{M_i^H(u)\} Y_i(u)/Y(u),$$

where $Y(u) = \sum_{i=1}^{n} Y_i(u) = \sum_{i=1}^{n} I(X_i \geq u)$.

The solution $\hat{\beta}^{imp}$ to $Q_n(e,\beta) \approx 0$ is an improved estimator for β. This estimator is also consistent and asymptotically normal, i.e.,

$$n^{1/2}(\hat{\beta}^{imp} - \beta) \xrightarrow{d} N(0, \Lambda_I). \tag{2.6}$$

The asymptotic variance matrix Λ_I of $\hat{\beta}^{imp}$ is still unknown and not explicitly given. But it can be consistently estimated by $\hat{\Lambda}_I = \hat{A}^{-1}[\hat{\Gamma} - \hat{\Gamma}_1](\hat{A}^{-1})'$ where $\hat{\Gamma}_1 = \hat{cov}(y_i, W_i)\hat{var}(W_i)^{-1}\hat{cov}(y_i, W_i)'$. Since $\hat{\Gamma}_1$ is a positively definite matrix, $\hat{\beta}^{imp}$ is asymptotically a more efficient estimator for β than the simple estimator $\hat{\beta}$. Similarly, a normal approximation-based confidence region for β can be constructed using (2.6) and $\hat{\Lambda}_I$. However, estimating Λ_I is much more complicated than estimating the asymptotic variance of $\hat{\beta}$.

2.2 Empirical Likelihood Methods

2.2.1 Empirical Likelihood Method Based on Influence Function

The existing inferential approaches for β can provide normal approximation-based confidence regions for β. However, as shown in section 2, these approaches require complex conditional density estimation for the error distribution and computationally burdensome calculations for the variance estimates of the simple estimator $\hat{\beta}$ and the improved estimator $\hat{\beta}^{imp}$. This provides the motivation for developing new EL-based confidence regions for the parameters in median cost regression models with censored data in this section.

According to the results from Bang and Tsiatis (2000)[1], the estimating function $Q_n(\beta)$ can be expressed in terms of a martingale process stochastic integral:

$$n^{-1/2}Q_n(\beta) \;=\; n^{-1/2}\sum_{i=1}^{n}\zeta_i(\beta) + o_p(1),$$

where $\zeta_i(\beta) = \mathbf{B_i}(\beta) - \int_0^L \frac{dM_i^c(u)}{K(u)}\{\mathbf{B_i}(\beta) - G(\mathbf{B}(\beta), u)\}$ is the i-th influence function of $Q_n(\beta)$, and

$$n^{-1/2}Q_n(e, \beta) \;=\; n^{-1/2}\sum_{i=1}^{n}D_i(e, \beta) + o_p(1),$$

where $D_i(e,\beta) = \zeta_i(\beta) + \gamma^{opt} \int_0^L \frac{dN_i^c(u)}{K(u)} \{e\{M_i^H(u)\} - G(e\{M^H(u)\}, u)\}$ is the i-th influence function of $Q_n(e,\beta)$.

Based on these influence functions, we propose the following empirical likelihood for β:

$$L_{IF}(\beta) = sup\{\prod_{i=1}^{n} p_i : p_1 \geq 0, ..., p_n \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{g}_i(\beta) = 0\}$$

where $\hat{g}_i(\beta) = \hat{\zeta}_i(\beta)$ or $\hat{D}_i(e,\beta)$, and

$$\begin{aligned} \hat{\zeta}_i(\beta) &= \mathbf{B_i}(\beta) - \int_0^L \frac{dM_i^c(u)}{\hat{K}(u)} \Big\{ \mathbf{B_i}(\beta) - \frac{\sum_{i=1}^n \Delta_i \mathbf{B_i}(\beta) I(T_i \geq u)/\hat{K}(T_i)}{n\hat{S}(u)} \Big\}, \\ \hat{D}_i(e,\beta) &= \hat{\zeta}_i(\beta) + \hat{\gamma}^{opt} \int_0^L \frac{dN_i^c(u)}{\hat{K}(u)} \Big\{ e\big\{M_i^H(u)\big\} - \hat{G}^*(e\big\{M^H(u)\big\}, u) \Big\}. \end{aligned}$$

Using the Lagrange multipliers, we can easily get $p_i = (1/n)\{1 + \lambda' \hat{g}_i(\beta)\}^{-1}$, $i = 1, ..., n$, where $\lambda = (\lambda_0, \lambda_1, ..., \lambda_p)'$ is the solution of the equation

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\hat{g}_i(\beta)}{1 + \lambda' \hat{g}_i(\beta)} = 0. \tag{2.7}$$

The Influence Function-based Empirical Log-likelihood (IFEL) ratio statistic for β can be defined as

$$l_{IF}(\beta) = 2 \sum_{i=1}^{n} log(1 + \lambda' \hat{g}_i(\beta)). \tag{2.8}$$

Theorem 2.2.1 *Let β_0 be the true value of β. Then $l_{IF}(\beta_0)$ converges in distribution to a χ^2_{p+1} random variable with $p+1$ degree of freedom as $n \to \infty$.*

Based on Theorem 2.2.1, at level $(1-\alpha)$, we can construct an IFEL-based confidence region for β as $\{\beta : l_{IF}(\beta) \leq \chi^2_{p+1,1-\alpha}\}$, where $\chi^2_{p+1,1-\alpha}$ is the $(1-\alpha)$-th quantile of χ^2_{p+1}.

2.2.2 Jackknife Empirical Likelihood Method

Another method, called jackknife empirical likelihood (JEL) (Jing, Yuan, and Zhou, 2009[29]), is shown to be very effective in confidence interval estimation. Here we define the JEL for β in the median cost regression model with censored data.

Let $T_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_i(\beta)$, $T_{n,-i}(\beta) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\hat{g}_{j,-i}(\beta)$, where $\hat{g}_{j,-i}(\beta)$ is j-th estimated influence function based on the $n-1$ observations from the original dataset after deleting the i-th observation. Then, the jackknife pseudo samples can be written as

$$\hat{J}_i(\beta) = nT_n(\beta) - (n-1)T_{n,-i}(\beta), i = 1, \cdots, n. \tag{2.9}$$

Applying Owen's EL to these jackknife pseudo samples, we get the following JEL for β:

$$L_J = sup\{\prod_{i=1}^{n} p_i : p_1 \geq 0, \cdots, p_n \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\hat{J}_i(\beta) = 0\}. \tag{2.10}$$

The Lagrange multiplier method provides the jackknife empirical log-likelihood ratio for β as follows:

$$l_J(\beta) = 2\sum_{i=1}^{n} log(1 + \lambda_J'\hat{J}_i(\beta)), \tag{LLJ}$$

where $\lambda_J = (\lambda_0, \lambda_1, ..., \lambda_p)'$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{J}_i(\beta)}{1 + \lambda_J'\hat{J}_i(\beta)} = 0. \tag{2.11}$$

Since the jackknife pseudo samples are asymptotically independent under mild conditions (Tukey 1958[62], Shi 1984[54]), it can be proved that the Wilks' theorem for $l_J(\beta)$ still holds, i.e., when β_0 is the true value of β, $l_J(\beta_0)$ converges in distribution to a χ^2_{p+1} random variable as $n \to \infty$. Hence, a $(1-\alpha)$ level JEL-based confidence region for β con be constructed as $\{\beta : l_J(\beta) \leq \chi^2_{p+1,1-\alpha}\}$.

2.3 Simulation Studies for Confidence Regions of the Regression Coefficients

In this section, we conduct simulation studies to compare our proposed IFEL and JEL based confidence regions for β with the existing normal approximation (NA) based confidence region in terms of coverage probabilities.

We assume that the subjects are independently and identically distributed. The total cost for each subject contains three cost components, and the total cost for the ith individual is generated by

$$M_i = M_i(0) + \sum_{j=1}^{L} b_{ij}[max\{min(T_i, j) - (j-1), 0\}] + d_i I(T \leq L),$$

where $M_i(0)$ is the initial diagnostic cost, b_{ij} is the cost that happened monthly over $(j-1)$th and jth months, and d_i is the terminal death cost, the total time period L is set to 10 and 20 months. Let each of the three cost components be related to a single covariate Z_i with following relationships:

$$M_i(0) = 50 + 50Z_i + e_i, \quad b_{ij} = 50 + 100Z_i + e'_{ij}, \quad d_i = 20Z_i + e''_i,$$

where we set $Z_i \sim U[0, 10]$, $\log e_i \sim N(\log 50, 0.245^2)$, $\log e'_{ij} \sim N(\log 10, 0.245^2)$, and $\log e''_{ij} \sim N(\log 10, 0.632^2)$.

The above generated M_i's satisfy $P(M_i < \alpha_0 + \beta_0 Z_i | Z_i) = 1/2$ with $\alpha_0 = 410$ and $\beta_0 = 570$. The survival time T_i's are generated from an exponential distribution with a median of 5 or 10 months.

The censoring time C_i's are generated from a uniform distribution on $[0, k]$ where different values of k, shown in Table 2.1 and Table 2.2, are used to generate censored survival times with light censoring proportion (25%-35%), moderate censoring proportion (45%-55%), and heavy censoring proportion (around 60%), respectively.

To investigate the effect of skewness of total costs on the performance of the proposed methods, we simulated total costs with skewness around 0.35, 1, or 2 such that $P(M_i <$

$\alpha_0 + \beta_0 Z_i | Z_i) = 1/2$ to make sure all scenarios with different skewness will have same median total cost. Two sample sizes $n = 400$, and 1000 are selected to generate cost data with different censoring proportions and skewness. 1000 simulations are run to calculate the coverage probabilities of the existing NA-based confidence region and the proposed IFEL and JEL based confidence regions for β at 95% confidence level in each simulation setting.

Tables 2.1-2.2 present the coverage probabilities of these (NA, IFEL, and JEL) regions at a 95% confidence level in each scenario of combinations of estimation procedures (Naive Estimation, Simple Estimation, Improved Estimation), skewness and censoring proportions. We use uncensored cases only in simulation of Naive Estimation procedures.

From these tables, we observe that IEFL and JEL based confidence regions for the regression parameters in the median cost regression model have higher coverage probabilities than the NA-based confidence regions in most cases. Especially, JEL method performs overall better in most of the scenarios.

The distributions of coverage probabilities of the NA, IFEL, and JEL based confidence regions with different censoring proportions are displayed in Figures 2.1-2.2 using box plots of coverage probabilities of these regions, where group labels are named as 'L' for light censoring, 'M' for moderate censoring, and 'H' for heavy censoring. For example, H.JEL stands for the boxplot of the coverage probabilities of the JEL confidence region based on 1000 simulation runs with heavy censoring. From Figures 2.1-2.2, we observe that when censoring is heavy, NA confidence regions have under coverage problems and could have a very low coverage probability, but IFEL and JEL regions have acceptable coverage probabilities that are closer to the nominal confidence level than the NA regions. Especially, when the sample size is small and the censoring proportion is high, coverage probabilities of NA regions will be much lower than the nominal level. As sample size increases, the performances of all the types of confidence regions improve, but still, coverage probabilities of IFEL and JEL based regions are overall higher than those of the NA-based regions.

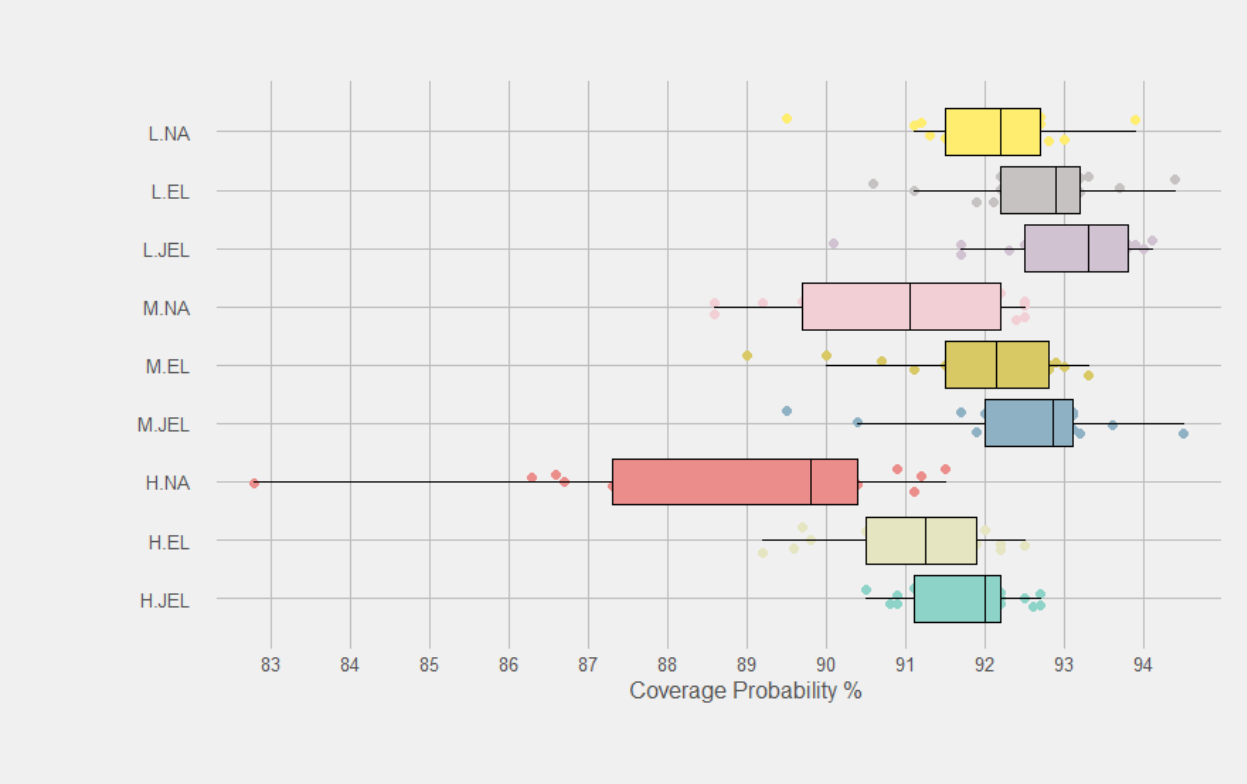

Figure 2.1 Boxplots of coverage probabilities under different simulation settings when sample size is 400

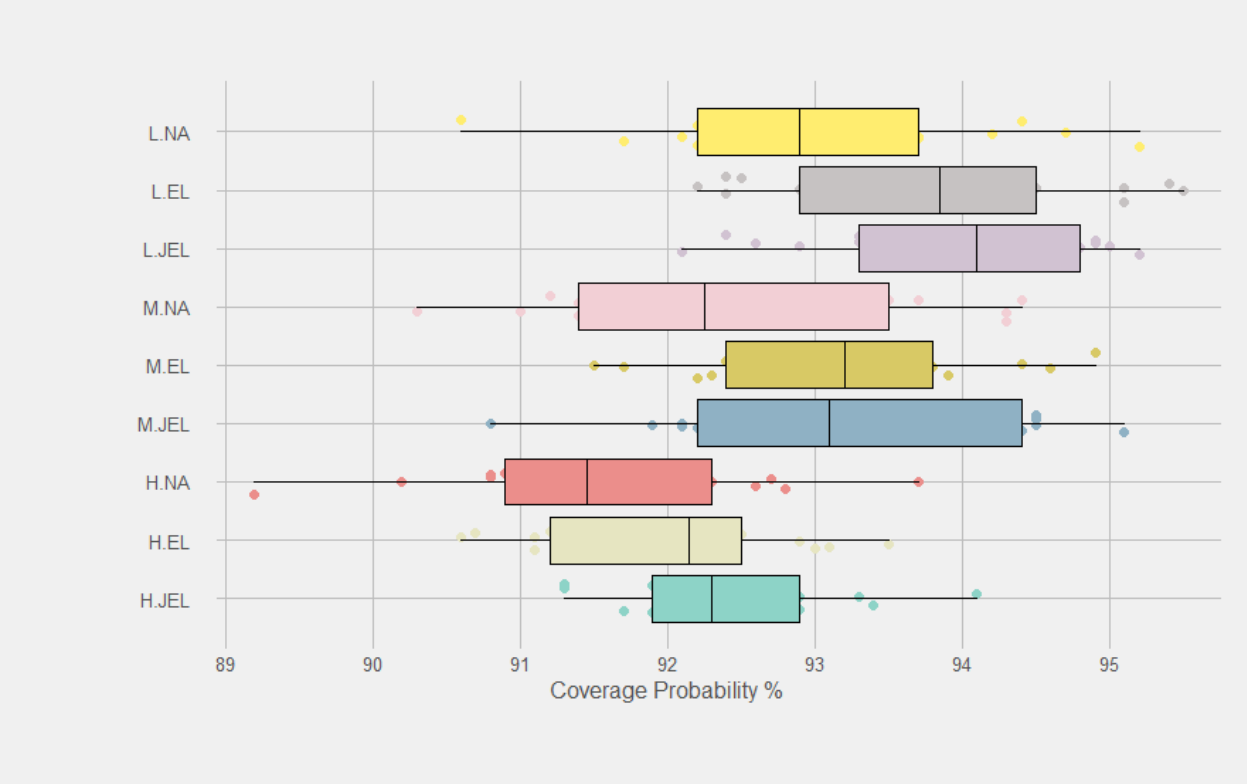

Figure 2.2 Boxplots of coverage probabilities under different simulation settings when sample size is 1000

Table 2.1 Coverage probabilities of various 95% level confidence regions for regression coefficients in median cost regression models. Median survival time =5 months. L= 10 months.

Procedure	Skewness	Light Censoring (25%-35%,$k = 20$)			Moderate Censoring (40%-50%,$k = 12.5$)			Heavy Censoring (Around 60%,$k = 8$)		
		NA	IFEL	JEL	NA	IFEL	JEL	NA	IFEL	JEL
Sample Size n=400										
Naive	0.3347	0.917	0.922	0.925	0.904	0.918	0.930	0.828	0.897	0.917
Simple	0.3347	0.925	0.927	0.929	0.904	0.915	0.927	0.912	0.922	0.926
Improved	0.3347	0.928	0.933	0.941	0.917	0.920	0.931	0.909	0.925	0.927
Naive	0.9936	0.924	0.931	0.938	0.912	0.922	0.928	0.915	0.916	0.909
Simple	0.9936	0.927	0.937	0.933	0.916	0.924	0.927	0.901	0.905	0.916
Improved	0.9936	0.939	0.944	0.940	0.925	0.928	0.921	0.902	0.911	0.922
Naive	1.9926	0.913	0.921	0.917	0.892	0.907	0.929	0.866	0.896	0.913
Simple	1.9926	0.919	0.922	0.923	0.925	0.930	0.932	0.911	0.915	0.920
Improved	1.9926	0.922	0.923	0.929	0.922	0.928	0.931	0.899	0.917	0.921
Sample Size n=1000										
Naive	0.3702	0.922	0.938	0.944	0.927	0.931	0.930	0.922	0.924	0.927
Simple	0.3702	0.935	0.941	0.949	0.923	0.939	0.945	0.927	0.922	0.929
Improved	0.3702	0.937	0.951	0.950	0.944	0.949	0.951	0.937	0.935	0.941
Naive	1.0205	0.931	0.937	0.933	0.928	0.925	0.930	0.917	0.916	0.923
Simple	1.0205	0.932	0.942	0.943	0.929	0.934	0.937	0.911	0.915	0.920
Improved	1.0205	0.937	0.945	0.949	0.935	0.933	0.939	0.918	0.922	0.928
Naive	2.0102	0.917	0.924	0.935	0.914	0.917	0.922	0.909	0.911	0.913
Simple	2.0102	0.924	0.929	0.933	0.910	0.923	0.919	0.908	0.925	0.923
Improved	2.0102	0.927	0.931	0.926	0.914	0.933	0.929	0.912	0.921	0.923

2.4 Empirical Likelihood Based Confidence Intervals for the Median Cost with Given Covariates

Since IFEL and JEL methods can produce confidence regions with good coverage accuracy, our goal is to construct confidence intervals for the median medical cost based on the IFEL and JEL regions in this section.

Note that the median cost over $[0, L]$ given covariate $\mathbf{Z}$ is $M \approx \beta'\mathbf{Z}$. Since there is no closed form for the confidence intervals of the median cost when the empirical likelihood method is used, we suggest applying a numerical method to construct an EL-based confidence interval for the median cost with given covariates. Let $R_\alpha(\beta) = \{\beta : l(\beta) \leq \chi^2_{p+1,1-\alpha}\}$ be

Table 2.2 Coverage probabilities of various 95% level confidence regions for regression coefficients in median cost regression models. Median survival time =10 months. L= 20 months.

Procedure	Skewness	Light Censoring (25%-35%,$k = 45$)			Moderate Censoring (40%-50%,$k = 22$)			Heavy Censoring (Around 60%,$k = 15$)		
		NA	IFEL	JEL	NA	IFEL	JEL	NA	IFEL	JEL
Sample Size n=400										
Naive	0.2823	0.895	0.906	0.901	0.897	0.900	0.904	0.867	0.892	0.911
Simple	0.2823	0.912	0.911	0.917	0.925	0.927	0.936	0.903	0.919	0.920
Improved	0.2823	0.922	0.931	0.938	0.924	0.933	0.945	0.892	0.920	0.922
Naive	1.1312	0.927	0.932	0.935	0.886	0.890	0.895	0.873	0.898	0.908
Simple	1.1312	0.930	0.929	0.939	0.909	0.921	0.931	0.893	0.914	0.925
Improved	1.1312	0.925	0.929	0.934	0.920	0.929	0.930	0.904	0.922	0.927
Naive	1.9884	0.915	0.929	0.927	0.886	0.911	0.919	0.863	0.907	0.905
Simple	1.9884	0.911	0.919	0.933	0.897	0.923	0.920	0.885	0.911	0.909
Improved	1.9884	0.917	0.932	0.934	0.902	0.918	0.917	0.897	0.908	0.922
Sample Size n=1000										
Naive	0.3236	0.927	0.939	0.934	0.917	0.922	0.921	0.911	0.907	0.917
Simple	0.3236	0.925	0.932	0.941	0.920	0.926	0.945	0.917	0.922	0.929
Improved	0.3236	0.947	0.944	0.945	0.943	0.946	0.944	0.923	0.931	0.934
Naive	1.0467	0.944	0.951	0.941	0.912	0.925	0.928	0.902	0.911	0.919
Simple	1.0467	0.942	0.955	0.952	0.943	0.938	0.945	0.928	0.930	0.921
Improved	1.0467	0.952	0.954	0.948	0.937	0.944	0.943	0.926	0.929	0.933
Naive	1.9926	0.906	0.922	0.921	0.903	0.915	0.908	0.892	0.906	0.913
Simple	1.9926	0.922	0.924	0.929	0.915	0.924	0.921	0.911	0.915	0.920
Improved	1.9926	0.921	0.925	0.924	0.922	0.935	0.932	0.908	0.912	0.919

the confidence region for β where $l(\beta) = l_{IF}(\beta)$ or $l_J(\beta)$. Then we can write the confidence interval for the median cost over $[0, L]$ given covariate $\mathbf{Z}$ as (q_l, q_u):

$$q_l = min\{\beta'\mathbf{Z} : l(\beta) = c,\ 0 \le c \le \chi^2_{p+1,1-\alpha}\} \approx min\{\bigcup_{i=1}^{N}\{\beta'\mathbf{Z} : l(\beta) = c_i\}\},$$

$$q_u = max\{\beta'\mathbf{Z} : l(\beta) = c,\ 0 \le c \le \chi^2_{p+1,1-\alpha}\} \approx max\{\bigcup_{i=1}^{N}\{\beta'\mathbf{Z} : l(\beta) = c_i\}\},$$

where N is a large integer number (e.g., $N = 10000$), $\{c_1, ..., c_N\}$ is a random sample generated from the uniform distribution on $[0, \chi^2_{p+1,1-\alpha}]$.

To estimate (q_l, q_u), we need to solve the equation $l(\beta) = c_i$ for each c_i $(i = 1, \cdots, N)$.

After solving the N equations and obtaining the solution β's, we can calculate the estimated median costs $\beta'\mathbf{Z}$'s with given $\mathbf{Z}$.

Finally, an EL-based confidence interval for the median medical cost with given $\mathbf{Z}$ can be obtained by finding the minimum and maximum of the $\beta'\mathbf{Z}$'s.

2.5 Real Data Analysis

In this section, we illustrate the application of our proposed methods based on a Canadian implantable defibrillator study (CIDS) provided by Dr. Willan (Willan *et al.*, 2005[64]). The data we used is a trial study of patients at risk of cardiac arrest. A total of 659 patients with resuscitated ventricular defibrillation or sustained ventricular tachycardia or with unmonitored syncope were randomized between amiodarone and implantable cardioverter defibrillators in a 7 year study period started from October 1990. However, only the first 430 patients' historical costs are well kept in the records, while the costs for the remaining 229 cases are left as 0. As a result, in this application study, we only focus on 430 cases with known medical costs. The primary outcome measure was all-cause mortality. Figure 2.3 displays the distributions of total costs for patients with different ages at diagnosis. Clearly, the total costs are highly skewed. Figure 2.4 shows the survival curves of patients in different sex and treatment groups. We can see that, overall, people with defibrillator treatment have a little bit better survival than those with amiodarone treatment. However, female cases with defibrillator treatment have significantly lower risk when patients are older than 70.

Three variables are highly correlated to total costs: age, sex (Male vs Female) and treatment (Defibrillator vs Amiodarone). For this analysis, it is interesting to estimate the median cost of patients with a given age. Here the ages 57, 65 and 70 are selected, which are 25_{th}, 50_{th} and 75_{th} percentiles of patients' ages in the sample.

Based on our simulation studies, the simple estimation equation is used to estimate the parameters in median cost regression with given covariates.

We choose $N = 10000$ to calculate confidence intervals for the median medical cost using the numerical method in section 5. The results are presented in Table 2.3. From Table 2.3,

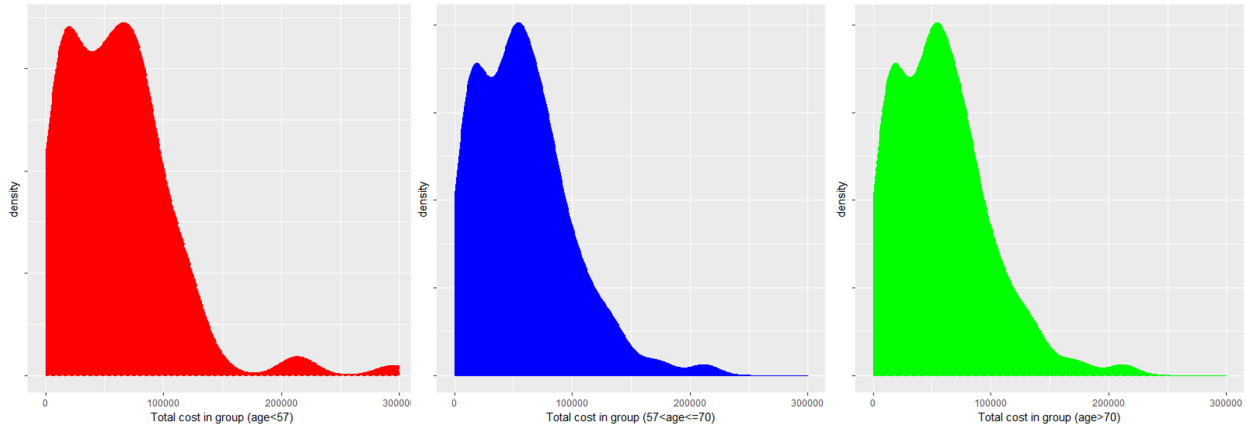

Figure 2.3 Distribution of total medical cost in different age groups

we observe that the median costs decrease when age increases, which is reasonable since the elderly have shorter survival times and their total medical costs until death could be less. The results also show that male patients cost more than female patients. What's more, defibrillator treatment leads to much more expenditure than amiodarone leads to. To those who are concerned with budget control, these results suggest that a large number of expenditures may not improve survival rates. From the table, we notice that, in most of the cases, the lower bounds of 95% IFEL and JEL confidence intervals are higher than those of the NA-based confidence intervals. Also, IFEL and JEL confidence intervals have a shorter length than the NA-based intervals for the median costs.

2.6 Discussion

In this part of the dissertation, we have developed empirical likelihood-based region-s/intervals for the parameters in median cost regression models when the costs of some patients are censored. The proposed EL-based methods have sound asymptotic properties (i.e., Wilks' theorem). Our simulation studies showed that the proposed IFEL and JEL confidence regions have coverage probabilities much closer to the nominal confidence level than the normal approximation-based confidence regions in most cases. Especially, when the

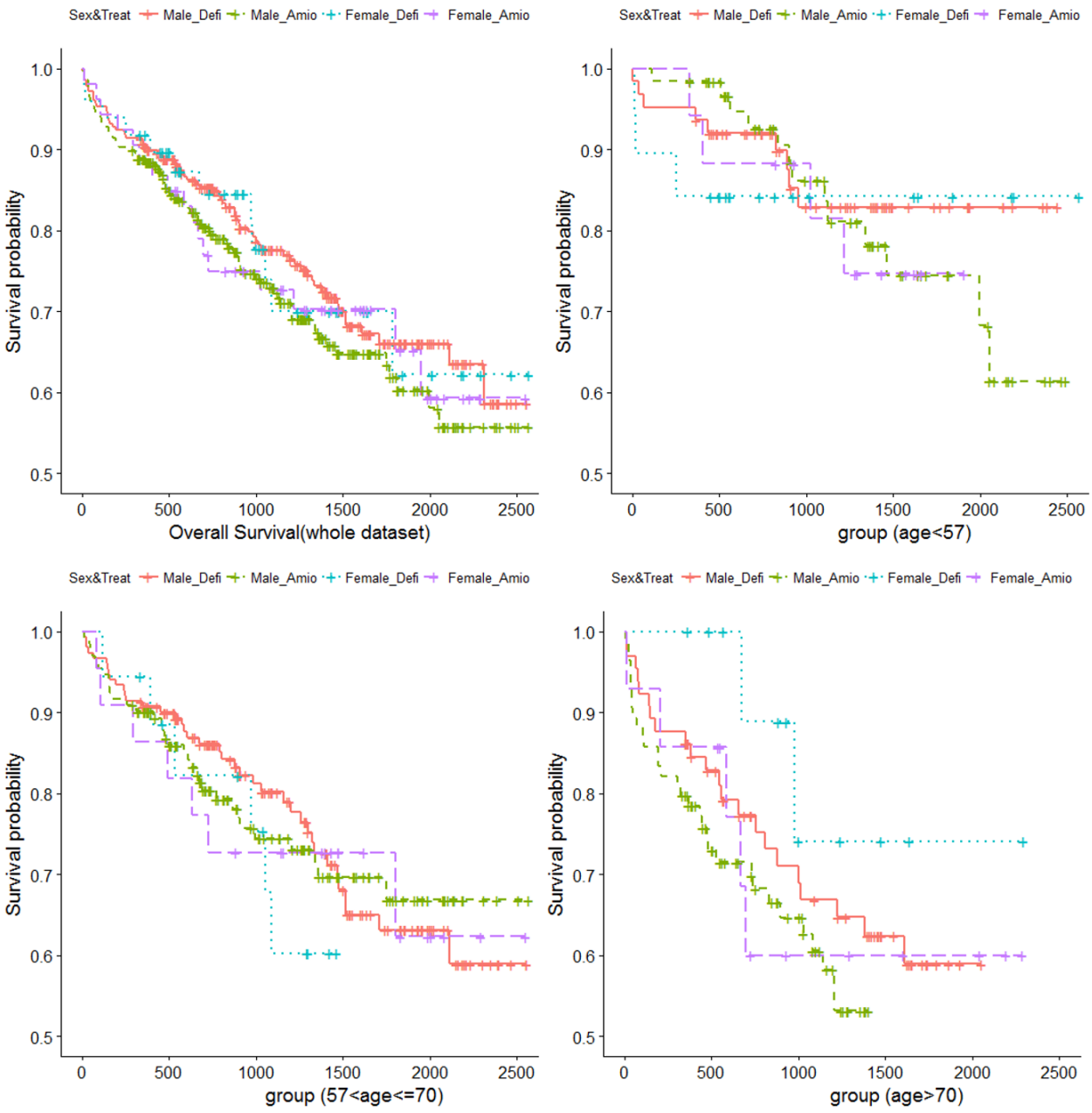

Figure 2.4 Kaplan Meier survival plot for patients under different treatment methods

Table 2.3 95% level confidence intervals for the median costs with given covariate age, sex, and treatment.

Age	Sex	Treatment	Median Cost	NA	IFEL	JEL
	M	Defibrillator	134618.6	[104418.1,145649.0]	[111358.2,146176.3]	[118714.4,142835.5]
57 (1st quartile)	M	Amiodarone	87029.7	[65069.3,103220.1]	[68165.5,95220.1]	[71179.0,94962.7]
	F	Defibrillator	114925.2	[95762.6,131104.2]	[105910.3,118549.7]	[97692.3,124177.2]
	F	Amiodarone	67336.3	[49552.3,91120.3]	[57658.4,94533.6]	[55924.7,96789.5]
	M	Defibrillator	118783.2	[102352.5,136244.3]	[109456.3,142356.5]	[108889.4,139214.5]
65 (Median)	M	Amiodarone	71194.3	[50113.2,93111.5]	[50226.5,94710.2]	[49133.2,92982.8]
	F	Defibrillator	99089.8	[81454.0,120319.6]	[88615.3, 125049.7]	[89154.1,119226.2]
	F	Amiodarone	51500.9	[26724.3,74302.6]	[30216.4,72350.6]	[29543.1,74169.5]
	M	Defibrillator	108886.0	[88512.5,126844.7]	[109225.6,135425.3]	[101231.4,130825.6]
70 (3rd quartile)	M	Amiodarone	61297.2	[42133.8,85219.5]	[43566.5,92619.3]	[42179.0,93129.6]
	F	Defibrillator	89192.7	[71249.1, 116928.1]	[72188.1,99861.2]	[74520.1,111425.2]
	F	Amiodarone	41603.8	[16793.2, 66532.7]	[20139.5,66533.1]	[20793.4, 61354.8]

sample size is small, censoring is heavy, and skewness is high, the IFEL and JEL approaches perform much better than the existing normal approximation-based methods that need complex variance estimation. The proposed confidence regions/intervals for the parameters in median cost regression models can be directly calculated by implementing the algorithm for computing the standard empirical likelihood interval/region without the variance estimation (Hall and La Scala, 1990[22]). Based on this study, we recommend the use of the proposed IFEL and JEL methods for inferences in median cost regression models with censored data.

2.7 Proof of Theorems

We first introduce some preliminary results and three Lemmas which are needed to prove Theorem 2.2.1. Let $X = min(T, C)$, $H(u) = P(X \leq u)$, and $b_H = sup\{u : H(u) < 1\}$. The following results are due to Zhou (1992)[69]:

$$\Psi_n \equiv \sup_{s \leq t} \frac{|K(s^-) - \hat{K}(s^-)|}{\hat{K}(s)} = o_p(1) \ for \ \forall t < b_H, \tag{2.12}$$

$$\Phi_n \equiv \sup_{s \leq max\{X_i\}} \frac{|K(s^-) - \hat{K}(s^-)|}{\hat{K}(s)} = O_p(1). \tag{2.13}$$

Lemma 1 (This Lemma is cited from Lemma 3.3. in He and Liang (2014)[25].

For $t < b_H$, let $\{h_n(t)\}$ be a random sequence such that $h_n(t) \to h(t)$ in distribution as $n \to \infty$, and $h(t) = o_p(1)$ as $t \to b_H$. As $n \to \infty$, if $V_n = O_p(1)$ and the random sequence $\{S_n\}$ can be written as $S_n = o_p(1) + V_n h_n(t)$ for any $t < b_H$, then $S_n = o_p(1)$.

Here we provide the proof of Theorem 2.2.1 when the influence function $\hat{g}_i(\beta) = \hat{\zeta}_i(\beta)$ based on the simple estimation method for β, and the theorem can be proved similarly when $\hat{g}_i(\beta) = \hat{D}_i(e, \beta)$.

Lemma 2. $n^{-1}\sum_{i=1}^{n} ||\hat{\zeta}_i(\beta_0) - \zeta_i(\beta_0)||^2 = o_p(1)$.

Proof of Lemma 2

From

$$\begin{aligned}
\zeta_i(\beta_0) &= \mathbf{B}_i(\beta_0) - \int_0^L \frac{dM_i^c(u)}{K(u)}\{\mathbf{B}_i(\beta_0) - G(\mathbf{B}(\beta_0), u)\}, \\
\hat{\zeta}_i(\beta_0) &= \mathbf{B}_i(\beta_0) - \int_0^L \frac{dM_i^c(u)}{\hat{K}(u)}\Big\{\mathbf{B}_i(\beta_0) - \hat{G}(\mathbf{B}(\beta_0), u)\Big\},
\end{aligned}$$

where $\hat{G}(\mathbf{B}(\beta_0), u) = \dfrac{\sum_{i=1}^n \Delta_i \mathbf{B}_i(\beta_0) I(T_i \geq u)/\hat{K}(T_i)}{n\hat{S}(u)}$, we get that

$$\hat{\zeta}_i(\beta_0) - \zeta_i(\beta_0) = A_i + B_i,$$

where

$$A_i = \int_0^L \Big(\frac{\mathbf{B_i}(\beta_0)}{K(u)} - \frac{\mathbf{B_i}(\beta_0)}{\hat{K}(u)}\Big) dM_i^c(u), \tag{2.14}$$

$$B_i = \int_0^L \Big(\frac{\hat{G}(\mathbf{B}(\beta_0), u)}{\hat{K}(u)} - \frac{G(\mathbf{B}(\beta_0), u)}{K(u)}\Big) dM_i^c(u). \tag{2.15}$$

The lemma will be proved by showing that the sample means of $||A_i||^2$ and $||B_i||^2$ tend to zero in probability. The proof will be presented in Part 1 and Part 2 below.

Part 1: $n^{-1}\sum_{i=1}^n ||A_i||^2 = o_p(1)$.

If $L < b_H$, the proof is easy. Let's take $L = b_H$ and split the first integral on $[0, b_H]$ in the

following equation into two integrals on intervals $[0,t]$ and $(t,b_H]$ with $t<b_H$.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}||A_i||^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\int_0^L \left\|\frac{\mathbf{B_i}(\beta_0)}{K(u)}-\frac{\mathbf{B_i}(\beta_0)}{\hat{K}(u)}\right\|^2 dM_i^c(u)\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\|\frac{\mathbf{B_i}(\beta_0)}{K(u)}-\frac{\mathbf{B_i}(\beta_0)}{\hat{K}(u)}\right\|^2 dM_i^c(u)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\int_t^L \left\|\frac{\mathbf{B_i}(\beta_0)}{K(u)}-\frac{\mathbf{B_i}(\beta_0)}{\hat{K}(u)}\right\|^2 dM_i^c(u)\\
&\leq \Psi_n^2\frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{||\mathbf{B_i}(\beta_0)||^2}{K^2(u)}dM_i^c(u)+\Phi_n^2\frac{1}{n}\sum_{i=1}^{n}\int_t^L \frac{||\mathbf{B_i}(\beta_0)||^2}{K^2(u)}dM_i^c(u) \qquad (2.16)
\end{aligned}$$

From Lemma 1, it follows that $h(t)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\int_t^L \frac{||\mathbf{B_i}(\beta_0)||^2}{K^2(u)}dM_i^c(u)=o_p(1)$, as $t\to L$. Then using (2.12) and (2.13), we get that,

$$\frac{1}{n}\sum_{i=1}^{n}||A_i||^2 \leq o_p(1)O_p(1)+O_p(1)o_p(1)=o_p(1). \qquad (2.17)$$

Part 2: $n^{-1}\sum_{i=1}^{n}||B_i||^2=o_p(1)$.

Similar to the proof of Part 1, we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}||B_i||^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\int_0^L \left\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)}-\frac{G(\mathbf{B}(\beta_0),u)}{K(u)}\right\|^2 dM_i^c(u)\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)}-\frac{G(\mathbf{B}(\beta_0),u)}{K(u)}\right\|^2 dM_i^c(u)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\int_t^L \left\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)}-\frac{G(\mathbf{B}(\beta_0),u)}{K(u)}\right\|^2 dM_i^c(u)\\
&\equiv \frac{1}{n}\sum_{i=1}^{n}\Lambda_i+\frac{1}{n}\sum_{i=1}^{n}\Xi_i. \qquad (2.18)
\end{aligned}$$

From (2.12), we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\Lambda_i &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)} - \frac{G(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\le \frac{2}{n}\sum_{i=1}^{n}\int_0^t \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)} - \frac{\hat{G}(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\int_0^t \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{K(u)} - \frac{G(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\le \Psi_n^2\frac{2}{n}\sum_{i=1}^{n}\int_0^t \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\int_0^t \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u) - G(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&= o_p(1)O_p(1) + \frac{2}{n}\sum_{i=1}^{n}\Lambda_i^*, \qquad (2.19)
\end{aligned}$$

where $\Lambda_i^* = \displaystyle\int_0^t \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u) - G(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u)$. By the uniform consistency of Kaplan-Meier estimator, we have that $\hat{G}(\mathbf{B}(\beta_0),u) = G(\mathbf{B}(\beta_0),u) + o_p(1)$ uniformly on $[0,t]$. Hence, $\frac{1}{n}\sum_{i=1}^{n}\Lambda_i^* = o_p(1)$, and $\frac{1}{n}\sum_{i=1}^{n}\Lambda_i = o_p(1)O_p(1) + o_p(1) = o_p(1)$.
For the second term in (2.18), we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\Xi_i &= \frac{1}{n}\sum_{i=1}^{n}\int_t^L \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)} - \frac{G(\mathbf{B}(\beta_{\mathbf{0}}),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\le \frac{2}{n}\sum_{i=1}^{n}\int_t^L \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{\hat{K}(u)} - \frac{\hat{G}(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\int_t^L \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{K(u)} - \frac{G(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\le \Phi_n^2\frac{2}{n}\sum_{i=1}^{n}\int_t^L \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\int_t^L \Big\|\frac{\hat{G}(\mathbf{B}(\beta_0),u) - G(\mathbf{B}(\beta_0),u)}{K(u)}\Big\|^2 dM_i^c(u) \\
&= O_p(1)o_p(1) + \frac{2}{n}\sum_{i=1}^{n}\Xi_i^*, \qquad (2.20)
\end{aligned}$$

$$
\begin{aligned}
\Xi_i^* &= \int_t^L \Big\| \frac{\hat{G}(\mathbf{B}(\beta_0),u) - G(\mathbf{B}(\beta_0),u)}{K(u)} \Big\|^2 dM_i^c(u) \\
&= \int_t^L \frac{dM_i^c(u)}{K^2(u)} \Big\| \frac{1}{n\hat{S}(u)} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} - \frac{E(\mathbf{B_i}(\beta_0) I(T_i \geq u))}{S(u)} \Big\|^2 \\
&= \int_t^L \frac{dM_i^c(u)}{K^2(u)} \Big\| \Big(\frac{1}{\hat{S}(u)} - \frac{1}{S(u)} \Big) \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} \\
&\qquad + \frac{1}{S(u)} \Big(\frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} - E(\mathbf{B_i}(\beta_0) I(T_i \geq u)) \Big) \Big\|^2 \\
&\leq 2 \int_t^L \frac{dM_i^c(u)}{K^2(u)} \Big(\frac{1}{\hat{S}(u)} - \frac{1}{S(u)} \Big)^2 \Big\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} \Big\|^2 \\
&\qquad + 2 \int_t^L \frac{dM_i^c(u)}{K^2(u)} \Big\| \frac{1}{S(u)} \Big(\frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} - E(\mathbf{B_i}(\beta_0) I(T_i \geq u)) \Big) \Big\|^2 \\
&\equiv 2\Xi_{i1}^* + 2\Xi_{i2}^*. \qquad (2.21)
\end{aligned}
$$

From $\frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} = o_p(1)$, Lemma 1 and (2.13), it follows that

$$
\begin{aligned}
\Xi_{i1}^* &= \int_t^L \frac{dM_i^c(u)}{K^2(u)} \Big(\frac{S(u) - \hat{S}(u)}{S(u)\hat{S}(u)} \Big)^2 \Big\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i)} \Big\|^2 \\
&= O_p(1) o_p(1) = o_p(1), \qquad (2.22)
\end{aligned}
$$

and

$$
\begin{aligned}
\Xi_{i2}^* &\leq 2 \int_t^L \frac{dM_i^c(u)}{K^2(u) S^2(u)} \Big[\Big\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{K(T_i)} - E(\mathbf{B_i}(\beta_0) I(T_i \geq u)) \Big\|^2 \\
&\qquad + \Big\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i \mathbf{B_i}(\beta_0) I(T_i \geq u)}{\hat{K}(T_i) K(T_i)} (K(T_i) - \hat{K}(T_i)) \Big\|^2 \Big] \\
&= o_p(1) + O_p(1) o_p(1) = o_p(1). \qquad (2.23)
\end{aligned}
$$

Therefore, from (2.18) - (2.23), we get that $\frac{1}{n} \sum_{i=1}^n ||B_i||^2 = o_p(1)$.

Finally, we have that

$$
\frac{1}{n} \sum_{i=1}^n ||\hat{\zeta}_i(\beta_0) - \zeta_i(\beta_0)||^2 = \frac{1}{n} \sum_{i=1}^n ||A_i + B_i||^2 \leq \frac{2}{n} \sum_{i=1}^n (||A_i||^2 + ||B_i||^2) = o_p(1).
$$

Then proof of Lemma 2 is complete.

Lemma 3.

(i) $\max_{1\le i\le n} n^{-1/2}||\hat{\zeta}_i(\beta_0)|| = o_p(1)$.

(ii) $n^{-1}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0)\hat{\zeta}_i(\beta_0)' \xrightarrow{p} V$, where $V = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\zeta_i(\beta_0)\zeta_i(\beta_0)'$.

(iii) $n^{-1/2}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0) \xrightarrow{d} N(0,V)$.

Proof of Lemma 3

(i) Since $\zeta_i(\beta_0)$'s are i.i.d. random vectors with zero mean and the positively definite variance-covariance matrix V, we have that $\max_{1\le i\le n}||\zeta_i(\beta_0)|| = o_p(\sqrt{n})$. As a result,

$$\max_{1\le i\le n} n^{-1/2}||\hat{\zeta}_i(\beta_0)|| \le \max_{1\le i\le n} n^{-1/2}||\hat{\zeta}_i(\beta_0)-\zeta_i(\beta_0)|| + \max_{1\le i\le n} n^{-1/2}||\zeta_i(\beta_0)|| = o_p(1). \tag{2.24}$$

(ii) Let $\tilde{V} = n^{-1}\sum_{i=1}^{n}\zeta_i(\beta_0)\zeta_i(\beta_0)'$ and $V^* = n^{-1}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0)\hat{\zeta}_i(\beta_0)'$. For any $p+1$ dimensional vector a,

$$\begin{aligned} a'(V^*-\tilde{V})a &= \frac{1}{n}\sum_{i=1}^{n}(a'(\hat{\zeta}_i(\beta_0)-\zeta_i(\beta_0)))^2 + \frac{2}{n}\sum_{i=1}^{n}(a'\zeta_i(\beta_0))(a'(\hat{\zeta}_i(\beta_0)-\zeta_i(\beta_0))) \\ &\equiv J_1 + J_2. \end{aligned}$$

Similar to the proof of Lemma 2, we can get that $J_1 = o_p(1)$ and $J_2 = o_p(1)$, hence $a'(V^*-\tilde{V})a = o_p(1)$. Therefore $V^* = \tilde{V}+o_p(1) = V+o_p(1)$, and Lemma 3(ii) is proved.

(iii) Similar to the proof of Lemma 2, we can obtain that $n^{-1/2}\sum_{i=1}^{n}(\hat{\zeta}_i(\beta_0)-\zeta_i(\beta_0)) = o_p(1)$. Lemma 3(iii) follows immediately from $n^{-1/2}\sum_{i=1}^{n}\zeta_i(\beta_0) \xrightarrow{d} N(0,V)$ and the following decomposition

$$n^{-1/2}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0) = n^{-1/2}\sum_{i=1}^{n}(\hat{\zeta}_i(\beta_0)-\zeta_i(\beta_0)) + n^{-1/2}\sum_{i=1}^{n}\zeta_i(\beta_0).$$

Proof of Theorem 2.2.1

Using Lemma 3, we can get that $||\lambda|| = O_p(n^{-1/2})$.
From $\max_{1\le i\le n} ||\hat{\zeta}_i(\beta_0)|| = o_p(n^{1/2})$, $\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\zeta}_i(\beta_0)}{1+\lambda'\hat{\zeta}_i(\beta_0)} = 0$, and

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\zeta}_i(\beta_0)}{1+\lambda'\hat{\zeta}_i(\beta_0)} &= \frac{1}{n}\hat{\zeta}_i(\beta_0)[1-\lambda'\hat{\zeta}_i(\beta_0) + \frac{(\lambda'\hat{\zeta}_i(\beta_0))^2}{1+\lambda'\hat{\zeta}_i(\beta_0)}] \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0) - (\frac{1}{n}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0)\hat{\zeta}_i(\beta_0)')\lambda \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\zeta}_i(\beta_0)(\lambda'\hat{\zeta}_i(\beta_0))^2}{1+\lambda'\hat{\zeta}_i(\beta_0)}. \qquad (2.25)
\end{aligned}$$

it follows that

$$\begin{aligned}
\lambda &= \Big(\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0)\hat{\zeta}_i(\beta_0)'\Big)^{-1}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0) + o_p(n^{-1/2}), \\
0 &= \sum_{i=1}^{n}\frac{\lambda'\hat{\zeta}_i(\beta_0)}{1+\lambda'\hat{\zeta}_i(\beta_0)} = \sum_{i=1}^{n}\lambda'\hat{\zeta}_i(\beta_0) - \sum_{i=1}^{n}(\lambda'\hat{\zeta}_i(\beta_0))^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{(\lambda'\hat{\zeta}_i(\beta_0))^3}{1+\lambda'\hat{\zeta}_i(\beta_0)} \\
&= \sum_{i=1}^{n}\lambda'\hat{\zeta}_i(\beta_0) - \sum_{i=1}^{n}(\lambda'\hat{\zeta}_i(\beta_0))^2 + o_p(1). \qquad (2.26)
\end{aligned}$$

Then we have that $\sum_{i=1}^{n}\lambda'\hat{\zeta}_i(\beta_0) = \sum_{i=1}^{n}(\lambda'\hat{\zeta}_i(\beta_0))^2 + o_p(1)$. Applying Taylor's expansion to the log likelihood ratio, we get that

$$\begin{aligned}
l(\beta_0) &= 2\sum_{i=1}^{n} log\{1+\lambda'\hat{\zeta}_i(\beta_0)\} \\
&= 2\sum_{i=1}^{n}(\lambda'\hat{\zeta}_i(\beta_0) - \frac{1}{2}(\lambda'\hat{\zeta}_i(\beta_0))^2) + o_p(1) \\
&= (n^{-1/2}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0))'(n^{-1}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0)\hat{\zeta}_i(\beta_0)')^{-1}(n^{-1/2}\sum_{i=1}^{n}\hat{\zeta}_i(\beta_0)) + o_p(1) \\
&\xrightarrow{d} \chi^2_{p+1}.
\end{aligned}$$

PART 3

NOVEL STATISTICAL METHODS FOR MEAN AND UPPER QUANTILE MEDICAL COSTS WITH CENSORED AND ZERO-INFLATED OBSERVATIONS

In practice, we observe that medical cost data contain a large proportion of zeros. For example, some patients were not aware of their illness or they were not willing to go to the hospital to take diagnostic tests. Such cases could happen especially when patients have nonlethal or chronic diseases or they are not in good financial condition. However, for policymakers or healthcare providers, those zero costs can not be ignored naively since those circumstances may change with the rapid development of the modern world.

For data distribution with many zeros and heavy long tails, researches prefer to assume such zero-inflated data follows an underlying distribution like delta-lognormal distribution [24] in which there is a probability δ of being zero and those positive values follows a lognormal distribution with mean μ and standard deviation σ. However, the distribution of the nonzero costs is usually unknown, which motivates us to propose nonparametric methods for the cost data analysis. Our main research interest is to make an inference of the mean and upper quantiles of censored and zero-inflated medical costs.

This part of the dissertation is organized as follows. In Section 3.1, we describe the methods used for making inferences for zero-inflated mean costs. In Section 3.2, we describe the methods used for making inferences for zero-inflated upper quantile costs. In Section 3.3, we performed simulation studies for mean and quantile costs and compare the coverage probabilities we achieved from different methods in different scenarios. In Section 3.4, we demonstrate the application of our method to a real-world dataset. Section 3.5 presents the conclusion and discussion.

3.1 Inference for Mean Costs

3.1.1 Notations and Assumptions

Before the introduction of our inference methods, we first need to clarify the notations that will be used. Similar to Part 1, for the i_{th} patient, its overall survival time can be denoted as T_i and C_i represents the censoring time. Censoring can occur when a patient leaves from a study, when a patient is lost to follow-up or for administrative reasons such as a patient's follow-up time is less than the survival time (Bang and Zhao, 2012[3]). We need to have a random censoring assumption, which means a patient's censoring time is independent of the survival time and the total cost. Our assumption usually holds when censoring occurs because of study termination, which is referred to as administrative censoring. Due to the existence of large censoring, medical cost for the patients are not completely observed in most cases.

The observed survival time of i_{th} patient is denoted by $X_i = min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$ will be the indicator of censoring status. Because of censoring, it is difficult to estimate the medical cost for the entire life without any time restrictions. Thus, a general approach is to consider the cost incurred by a patient up to a fixed time point L, where a reasonable amount of complete data is available over the time period $[0, L]$. Then, the modified time to event value T_i^L will be the minimal of T_i and L. For simplicity, we will use T_i instead of T_i^L for the rest of this part of the dissertation. Let $M(t)$ denote the total medical cost for a patient from the time that the patient entered the study ($t = 0$) to time t. Since we are interested in making inference for mean and quantiles of total medical costs, the observed total cost for i_{th} individual can be shown as $M_i \equiv M(X_i)$. If the patient experienced the event (death) before being censored, then $M_i \equiv M(X_i) = M(T_i)$. The observed entry for each individual of n cases should be represented as $\big\{(X_i, \Delta_i, M_i) : i = 1, \cdots, n\big\}$.

3.1.2 Estimating Overall Mean of Medical Costs

In our proposed scenarios, we need to consider both zero and positive medical costs. Considering a lot of total costs being zeros ($M_i = 0$), which can not provide much useful information, instead, we decide to use M_j^+ to indicate only positive costs in proposed estimators. And then, $\mu = E(M_i)$ and $\mu^+ = E(M_j^+)$ can be used to denote expectations of total cost and positive cost accordingly.

Let $\delta = P(M_i = 0) > 0$ be the percentage of zero costs. Let n_0 and n_1 be the number of zero costs and positive costs M_j^+'s respectively. We can assume $n_0 \sim \text{binomial}(n, \delta)$.

The overall mean value can be represented as $\mu = (1-\delta)\mu^+$ given δ and μ^+. However, it is hard to make inference for $(1-\delta)\mu^+$ due to such a product term. Hence, it leads us to think of making log transformation first to convert a product term to sum of two components such that we will have $\log \mu = \log((1-\delta)\mu^+) = \log(1-\delta) + \log(\mu^+)$. Let $\hat{\delta}$ and $\hat{\mu}^+$ be unbiased estimates for δ and μ^+. Then, we could apply MOVER (Method of Variance Estimate Recovery) algorithm to generate an approximated confidence interval for $\log(\mu)$ if symmetric confidence intervals for $\log(1-\delta)$ and $\log(\mu^+)$ can be constructed successfully. Here, we should hold the assumption that $\log(1-\hat{\delta})$ and $\log(\hat{\mu}^+)$ are independent.

3.1.3 MOVER Confidence Intervals for the Mean Cost

MOVER, first proposed by Graybill and Wang[20] to find a confidence limit for a linear combination of variance components, is a method to find a CI for a linear combination of parameters based on individual CIs of the parameters[72, 73]. The MOVER CI for the sum of two parameters can be described as follows. Let $\hat{\gamma}_1$ and $\hat{\gamma}_2$ be unbiased estimates of γ_1 and γ_2. And let (l_i, u_i) be a $1-\alpha$ CI for γ_i. Then, a $1-\alpha$ MOVER CI (L, U) for $\gamma_1 + \gamma_2$ can be expressed as

$$\begin{aligned} L &= \hat{\gamma_1} + \hat{\gamma_2} - \sqrt{(\hat{\gamma_1} - l_1)^2 + (\hat{\gamma_2} - l_2)^2} \\ &\textit{and} \\ U &= \hat{\gamma_1} + \hat{\gamma_2} + \sqrt{(\hat{\gamma_1} - u_1)^2 + (\hat{\gamma_2} - u_2)^2} \end{aligned} \tag{3.1}$$

Let $\gamma_1 = \log(1-\delta)$ and $\gamma_2 = \log(\mu^+)$. Our main task is to estimate γ_1, γ_2 and their confidence intervals (l_1, u_1) and (l_2, u_2). After that, we can obtain a confidence interval (L, U) for $\log \mu$. Finally, the MOVER CI for the mean medical cost μ will be given by $(exp(L), exp(U))$.

3.1.4 Fiducial Confidence Intervals for the Zero Proportion

First, we need to find a confidence interval for $\gamma_1 = \log(1-\delta)$. Enlighten by Weerahandi(1995)[63] and Hannig(2009)[23], we will apply the method based on fiducial quantity (also known as generalized pivotal quantity(GPQ)) to construct a confidence interval for $\gamma_1 = \log(1-\delta)$.

To construct a fiducial quantity-based confidence interval γ_1, we need to find a fiducial distribution for δ which is the proportion of zero medical costs. Let $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ with parameters a and b. Then the Beta distribution $B_{a,b}$ with pdf $\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$ can be used as a fiducial distribution for δ. Here we propose to use the following fiducial distributions for δ:

The first one is proposed by Li, Zhou and Tian (2013) [35]:

$$F_{\delta_{LZT}} = 0.5(B_{n_0,n_1+1} + B_{n_0+1,n_1}). \tag{3.2}$$

The second one is recently proposed by Most Hasan and Krishnamoorthy (2018)[24]:

$$F_{\delta_{HK}} = B_{n_0+0.5,n_1+0.5}. \tag{3.3}$$

The third one is based on the Wilson score confidence interval for a binomial proportion and proposed by Li *et al.*(2013)[35]:

$$F_{\delta_{Li}} = \frac{n_0 + Z^2/2}{n+Z^2} + \frac{Z}{n+Z^2}\{n_0(1-\frac{n_0}{n}) + \frac{Z^2}{4}\}^{1/2}, \tag{3.4}$$

where $Z \sim N(0,1)$.

Then, a 95% level fiducial confidence interval for δ, denoted as $(\hat{\delta}'_l, \hat{\delta}'_u)$, can be found by

using Monte Carlo simulation and any one of the three fiducial distributions for δ, where $\hat{\delta}'_l$ is the 2.5% quantile $\hat{F}_{\delta_{0.025}}$ of the fiducial quantities, and $\hat{\delta}'_u$ is the 97.5% quantile $\hat{F}_{\delta_{0.975}}$ of the fiducial quantities generated from $F_\delta = F_{\delta_{LZT}}$, $F_{\delta_{HK}}$, or $F_{\delta_{Li}}$ respectively.

3.1.5 Normal Approximation-based Confidence Intervals for the Zero Proportion

In addition to fiducial quantity based inference, a lot of scholars also proposed confidence intervals for binomial parameter based on normal approximation methods. Zou *et al.*[74] recently used the following $(1-\alpha)$ level normal approximation-based CI for $\delta' = 1-\delta$,

$$(\hat{\delta}'_l, \hat{\delta}'_u) = \frac{\hat{\delta}' + 0.5z^2_{1-\alpha/2}/n \pm \sqrt{[\hat{\delta}'(1-\hat{\delta}') + 0.25z^2_{1-\alpha/2}/n]/n}}{1 + z^2_{1-\alpha/2}/n}, \tag{3.5}$$

where $\hat{\delta}' = n_1/n$.

In the simulation study in Section 3.3, we apply above fiducial quantity-based methods and the normal approximation-based methods to make inferences for the zero cost proportion δ.

3.1.6 Normal Approximation based CIs for the Mean Positive Cost

Currently, available methods in medical cost analysis consider censoring but ignoring observed zero costs. In this section, we introduce normal approximation based methods available for making inference of mean positive medical costs with censored observations. Since the complete survival time and the total cost are not observed for each patient, we cannot naively estimate the mean medical cost by using the sample mean cost, which could generally cause an underestimation problem. Bang and Tsiatis (2000)[1] proposed the following simple weighted estimator for the mean medical cost of n nonzero observations:

$$\hat{\mu}_{BT} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i M_i}{\hat{K}(T_i)},$$

where $\hat{K}(t)$ is the Kaplan-Meier estimator for $K(t) = Pr(C_i > t)$ based on the data $\{(T_i, C_i, (1-\Delta_i) : i = 1, 2, \cdots, n\}$.

They have shown that the above estimator is consistent and asymptotically normal[1] with asymptotic variance estimator denoted as follows:

$$\begin{aligned}\hat{\sigma}^2_{BT} &= \frac{1}{n^2}\Bigg\{\sum_{i=1}^{n}\frac{\Delta_i(M_i-\hat{\mu}_{BT})^2}{\hat{K}(T_i)} \\ &\quad + \int_0^L \frac{dN^c(u)}{\hat{K}(u)^2}\Big[\hat{D}(M^2,u) - \hat{D}^2(M,u)\Big]\Bigg\},\end{aligned}$$

where $N^c(u) = \sum_{i=1}^n N_i^c(u) = \sum_{i=1}^n I(X_i \le u, \Delta_i = 0)$, $\hat{D}(M,u) = \dfrac{1}{n\hat{S}(u)}\displaystyle\sum_{i=1}^{n}\frac{\Delta_i M_i I(T_i \ge u)}{\hat{K}(T_i)}$, and $\hat{S}(u)$ is the Kaplan-Meier estimator for $S(u) = Pr(T_i > u)$. Finally, a normal approximation based confidence interval for mean positive medical costs can be obtained easily based on $\hat{\mu}$ and $\hat{\sigma}$.

Zhao and Tian (2001)[66] also proposed another more efficient estimator utilizing the patient's available cost histories instead of considering total cumulative cost only. Their estimator for the mean positive medical cost is defined as follows:

$$\begin{aligned}\hat{\mu}_{ZT} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i M_i}{\hat{K}(T_i)} + \frac{1}{n}\sum_{i=1}^{n}\int_0^L \frac{dN_i^c(u)}{\hat{K}(u)}\Big\{e\big\{M_i^H(u)\big\} \\ &\quad -\hat{D}^*\Big[e\big\{M_i^H(u)\big\},u\Big]\Big\},\end{aligned}$$

where $e\big\{M_i^H(u)\big\}$=$M_i(u)$ is a functional of the cost history $M_i^H(u)$, and $\hat{D}^*\Big[e\big\{M_i^H(u)\big\},u\Big] = \Big[\sum_{i=1}^n e\big\{M_i^H(u)\big\}Y_i(u)\Big]/Y(u)$ with $Y(u) = \sum_{i=1}^n Y_i(u) = \sum_{i=1}^n I(X_i \ge u)$.

Similarly, Zhao and Tian defined the asymptotic variance estimator as:

$$\begin{aligned}\hat{\sigma}^2_{ZT} &= \hat{\sigma}^2_{BT} - \frac{2}{n^2}\int_0^L \sum_{i=1}^{n} \frac{\Delta_i}{\hat{K}(T_i)}\Big[M_i - \hat{D}(M,u)\Big] \\ &\times\Big\{e\big\{M_i^H(u)\big\} - \hat{D}^*\Big[e\big\{M_i^H(u)\big\},u\Big]\Big\} \\ &\times\frac{I(T_i \geq u)dN^c(u)}{Y_(u)\hat{K}(u)} \\ &+\frac{2}{n^2}\int_0^L \sum_{i=1}^{n}\Big[e\big\{M_i^H(u)\big\} - \hat{D}^*\Big[e\big\{M_i^H(u)\big\},u\Big]\Big]^2 \\ &\times\frac{Y_i(u)dN^c(u)}{Y(u)\hat{K}^2(u)}.\end{aligned}$$

3.1.7 Empirical Likelihood based CIs for the Mean Positive Cost

For the nonzero mean medical cost, we are more interested in Bang and Tsiatis's simple weighted estimator (BT) since we observed that the improvement of the ZT estimator over the BT estimator is not significant but the ZT estimator requires much heavier calculation burden. Under our zero cost settings, we have the following estimator for mean medical cost $\tilde{\mu}_{BT}$ based on zero proportion δ and observed positive costs $\{(X_j^+, \Delta_j^+, M_j^+(u)) : 0 \leq u \leq X_j^+, j = 1, \cdots, n_1\}$:

$$\tilde{\mu}_{BT} = (1-\delta)\hat{\mu}_{BT} = (1-\delta)\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\Delta_j^+ M_j^+}{\hat{K}(T_j^+)} \tag{3.6}$$

where $\hat{\mu}_{BT}$ is the simple weighted estimator for mean of n_1 censored positive costs. we will have:

$$
\begin{aligned}
n^{\frac{1}{2}}(\tilde{\mu}_{BT}-\mu) &= n^{\frac{1}{2}}[(1-\delta)\hat{\mu}_{BT}-(1-\delta)\mu^{+}] \\
&= n^{\frac{1}{2}}(1-\delta)(\hat{\mu}_{BT}-\mu^{+}) \\
&= \sqrt{\frac{n}{n_1}}(1-\delta)\{n_1^{\frac{1}{2}}(\hat{\mu}_{BT}-\mu^{+})\} \\
&= \sqrt{\frac{n}{n_1}}(1-\delta)\{n_1^{-\frac{1}{2}}\sum_{j=1}^{n_1}[(M_j^{+}-\mu^{+}) \\
&- \int_0^L \frac{M_j^{+}-D(M^{+},u)}{K(u)}dM_j^{+c}(u)] + o_p(1)\} \\
&= \sqrt{\frac{n}{n_1}}(1-\delta)\{n_1^{-\frac{1}{2}}\sum_{j=1}^{n_1} g_j(\mu^{+}) + o_p(1)\}
\end{aligned}
\tag{3.7}
$$

In order to make inference for the positive mean medical cost, except for the normal approximation methods described in Section 3.1.6, research result from Jeyarajah and Qin(2017)[27] shows that when cost data are heavily censored and heavily skewed, EL-based inference will have better coverage probability than normal approximation approaches, especially when sample size is small.

For EL-based inference, from equation 3.7, let

$$
g_j(\mu^{+}) = (M_j^{+}-\mu^{+}) - \int_0^L \frac{M_j^{+}-D(M^{+},u)}{K(u)}dM_j^{+c}(u)
$$

to be the j-th influence function of $\hat{\mu}_{BT}$, we will have the following likelihood function[27].

$$
L_{IF}(\mu^{+}) = sup\{\prod_{j=1}^{n_1} p_j : p_1 \geq 0, ..., p_{n_1} \geq 0, \sum_{i=1}^{n_1} p_j\hat{g}_j(\mu^{+}) = 0\} \tag{3.8}
$$

Similarly to the procedure from Section 2.2.1, we can get the Influence Function-based Empirical Log-likelihood ratio statistics for μ^{+}:

$$
l_{IF}(\mu^{+}) = 2\sum_{j=1}^{n_1}\log(1+\lambda\hat{g}_j(\mu^{+})) \tag{3.9}
$$

where λ is the solution of the equation $\dfrac{1}{n}\sum_{j=1}^{n_1}\dfrac{\hat{g}_j(\mu^{+})}{1+\lambda\hat{g}_j(\mu^{+})} = 0$.

According to Theorem 1 from Jeyarajah and Qin(2017)[27], a $(1-\alpha)$ level empirical likelihood interval based on the influence functions (ELI) for μ^+ can be constructed as $\{\mu^+ : l_{IF}(\mu^+) \leq \chi^2_{1,1-\alpha}\}$

And our EL based CI for $\gamma_2 = \log(\mu^+)$ will be $(l_2^{EL}, u_2^{EL}) = (min\{\log(\mu^+) : l_{IF}(\mu^+) \leq \chi^2_{1,1-\alpha}\}, max\{\log(\mu^+) : l_{IF}(\mu^+) \leq \chi^2_{1,1-\alpha}\})$

3.1.8 The Symmetric Confidence Interval Adjustment

The construction of MOVER confidence interval (L, U) (equation 3.1) requires its components' CIs (l_i, u_i) be symmetric with respect to γ_i. Since empirical likelihood based confidence intervals (l_2^{EL}, u_2^{EL}) may not be symmetric, which may cause under coverage problem according to our simulation studies, especially for small sample scenarios. Inspired by Li *et al.*[34], we can construct a symmetric interval for γ_2 based on (l_2^{EL}, u_2^{EL}).

Let $\hat{\gamma_2} = \log(\hat{\mu}_{BT}) = \log \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\Delta_j^+ M_j^+}{\hat{K}(T_j^+)}$, $\Delta_l = \hat{\gamma_2} - l_2^{EL}$ and $\Delta_u = u_2^{EL} - \hat{\gamma_2}$. Then, a symmetric EL-based confidence interval for γ_2 is defined as

$$(l_{ELS}, u_{ELS}) = (\hat{\gamma_2} - \sqrt{(\Delta_l^2 + \Delta_u^2)/2}, \hat{\gamma_2} + \sqrt{(\Delta_l^2 + \Delta_u^2)/2}) \tag{3.10}$$

3.2 Inference for Upper Quantile Costs

3.2.1 Notations and Assumptions

The standard notations we used for quantile costs study are the same as what we introduced in section 3.1.1. In summary, the observed survival time of a patient is denoted by $X_i = min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$ is the indicator of censoring. The total cost at time u is denoted by $M_i(u)$. The observed data are $\big\{(X_i, \Delta_i, M_i(u)), 0 \leq u \leq X_i = 1, \cdots, n\big\}$.

3.2.2 Existing Methods for Inferences of Non-zero Quantile Costs with Censored Data

As far as we know, the inference method for censored and zero-inflated medical costs is still not available. Current approaches for quantile medical costs such as Zhao *et al.*'s [68] nonparametric methods can be used to estimate the median of positive medical costs

and confidence interval with censored observations, which can also be easily extended to any positive quantiles. In order to estimate the quantile medical cost, they proposed an estimator for the survival function $S(x) = P\{M_i(T_i) > x\}$ from total cost data based on the inverse probability weighting scheme which was originally proposed by Horvitz and Thompson (1952)[26] for analyzing survey data, and later used for handling various statistical problems in biostatistics.

Zhao *et al.*'s simple weighted (SW) estimator for the survival probability for cumulative cost x is defined as

$$\hat{S}_{SW}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I(M_i > x)}{\hat{K}(T_i)} \tag{3.11}$$

where $\hat{K}(u)$ is the Kaplan-Meier estimator[31] for the survival function $K(u) = P(C > u)$ of the censoring variable C based on the data $\{(X_i, (1-\Delta_i)),\ i = 1, 2, \cdots, n\}$. For such simple weighted estimator, in order to take all censored observations into consideration, each uncensored patient needs to represent on average $\frac{1}{K(T_i)}$ patients whose survival times are censored by weighting each uncensored patient with its probability of being uncensored.

According to Zhao and Tsiatis (1997)[67] and theories for counting process and missing data problems, Zhao *et al.* (2012) showed that the simple weighted estimator is consistent and asymptotically normal:

$$\frac{\hat{S}_{SW}(x) - S(x)}{\sigma_{SW}(x)} \xrightarrow{d} N(0,1)$$

where $\sigma^2_{SW}(x)$ represents the variance of $\hat{S}_{SW}(x)$.

The asymptotic variance $\sigma^2_{SW}(x)$ of $\hat{S}_{SW}(x)$ is still unknown, but it can be estimated by

$$\hat{\sigma}^2_{SW}(x) = \frac{1}{n}\hat{S}_{SW}(x)\Big\{1 - \hat{S}_{SW}(x)\Big\} + \frac{1}{n^2}\sum_{i=1}^{n}\frac{(1-\Delta_i)}{\hat{K}^2(C_i)}\Big[\hat{G}(B, C_i) - \hat{G}^2(B, C_i)\Big],$$

where

$$\hat{G}(B, u) = \frac{1}{n\hat{S}^T(u)}\sum_{j=1}^{n}\frac{\Delta_j B_j I(T_j \geq u)}{\hat{K}(T_j)}$$

with $B_j = I(M_j > x)$ and $\hat{S}^T(u)$ being the Kaplan-Meier estimator of $S^T(u) = P(T > u)$.

Let y_p be the p-th quantile of the medical cost where $0 < p < 1$, i.e., $S(y_p) = 1 - p$,

and we use $\hat{S}_{SW}(y_p)$ as the simple weighted estimator for $S(y_p)$. Then y_p can be consistently estimated by $\hat{y_p}$ which is the solution of the following equation:

$$\hat{y_p} = inf\{x : \hat{S}(x) \leq 1 - p\}.$$

For example, when $p = 0.5$, $\hat{y}_{0.5}$ is a consistent estimator for the median cost.

Zhao *et al.* (2012) proposed the following confidence interval for the quantile cost y_p based on the asymptotic normality of $\hat{S}(x)$:

$$R_\alpha = \left\{ x : \frac{[\hat{S}(x) - (1-p)]^2}{\hat{\sigma}^2(x)} \leq \chi^2_{1,1-\alpha} \right\},$$

where $\hat{\sigma}^2(x) = \hat{\sigma}^2_{SW}(x)$, and $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-th quantile of the χ^2 distribution with one degree of freedom.

However, if we want to make inferences based on the whole dataset, cases with zero costs due to various reasons usually do not have survival information on records. As a result, to achieve accurate inferences for quantiles using a normal approximation approach is not an option. Moreover, the confidence intervals constructed need complex variance estimates and may have poor small sample performances due to high skewness and heavy censoring.

3.2.3 Estimating Quantiles of Medical costs

Even though the method above can not be used directly in our zero-inflated costs setting, we can still take benefit of their ideas to handle our positive medical costs part. If the proportion of zero costs δ is known, we can have $P(M_i > y_p) = (1-\delta)P(M_i > y_p | M_i > 0)$ for $\forall y_p > 0$ and $y_p = inf\{y : P(M_i > y) = 1 - p\} = inf\{x : P(M_i > x | M_i > 0) = 1 - r\}$, where $r = max\{0, (p-\delta)/(1-\delta)\}$.

Since $(p-\delta)/(1-\delta)$ has to be within $(0,1)$, we only focus on the upper p-th quantile cost with p greater than the zero cost proportion δ given n_0 and n_1. Let $x_r = y_p$, where x_r is the r_{th} quantile of n_1 positive medical costs $M_j^+ (j = 1, ..., n_1)$. Then $x_r = inf\{x : P(M_i > x | M_i > 0) = P(M_j^+ > x) = 1 - r\}$. We can also get an estimate for the positive cost survival

probability $S_{M^+}(x_r)$ for cost x_r as:

$$\hat{S}_{M^+}(x_r) = \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\Delta_j^+ I(M_j^+ > x_r)}{\hat{K}(T_j^+)} \tag{3.12}$$

Since r depends on δ, while δ is unknown, in order to make inference for x_r, these provoke the motivation to develop empirical likelihood-based intervals for the quantile cost with censored and zero-inflated data.

3.2.4 Empirical Likelihood Method

Let $Y_j(u) = I(X_j \geq u)$, $Y(u) = \sum_j Y_j(u)$, $N_j^c(u) = I(X_j \leq u, \Delta_j^+ = 0)$, $N^c(u) = \sum_j N_j^c(u)$, and let $\lambda^c(u)$ be the hazard function of the censoring time C. The corresponding martingale process $M_j^{+c}(u)$ can be expressed as

$$M_j^{+c}(u) = N_j^c(u) - \int_0^u \lambda^c(t)Y_j(t)dt.$$

From Zhao and Tsiatis[67] and Robins and Rotnitzky[52], we get the following equations:

$$\frac{\Delta_j}{K(T_j)} = 1 - \int_0^\infty \frac{dM^{+c}(u)}{K(u)}, \tag{3.13}$$

$$\frac{\hat{K}(T_j) - K(T_j)}{K(T_j)} = -\int_0^{T_j}\frac{\hat{K}(u^-)dM^{+c}(u)}{K(u)Y(u)}, \tag{3.14}$$

$$n^{-1}Y(u) = \hat{K}(u^-)\hat{S}(u^-), \tag{3.15}$$

where $M^{+c}(u) = \sum_j M_j^{+c}(u)$, and $\hat{S}(u)$ being the Kaplan-Meier estimator of $S(u) = P(T_j(x_r) \geq u)$.

According to Zhao and Tsiatis (1997)[67], we have:

$$\begin{aligned} n_1^{\frac{1}{2}}(\hat{S}_{M^+}(x_r) - S_{M^+}(x_r)) &= n_1^{-\frac{1}{2}}\sum_{j=1}^{n_1}[B_j - S_{M^+}(x_r) - \int_0^L \frac{B_j - G(B,u)}{K(u)}dM_j^{+c}(u)] + o_p(1) \\ &= n_1^{-\frac{1}{2}}\sum_{j=1}^{n_1}\gamma_j(\delta, x_r) + o_p(1) \end{aligned} \tag{3.16}$$

where $S_{M^+}(x_r) = 1 - r = (1-p)/(1-\delta)$, $B_j = I(M_j^+ > x_r)$,$G(B,u) = E\{B_j I(T_j(y_p) \geq u)/S(u)\}$, and $\gamma_j(\delta, x_r) = (B_j - (1-r)) - \int_0^L \frac{B_j - G(B,u)}{K(u)} dM_j^{+c}(u)$ is the j-th influence function of $\hat{S}_{M^+}(x_r)$.

Since $K(u)$ and $G(B,u)$ are still unknown, we replace them by their respective estimates $\hat{K}(u)$ and $\hat{G}(B,u)$, and get the following estimated influence function:

$$\hat{\gamma}_j(\delta, x_r) = \big(B_i - (1-p)/(1-\delta)\big) - \int_0^L \frac{B_j - \hat{G}(B,u)}{\hat{K}(u)} dM_i^{+c}(u).$$

Based on these influence functions, we propose the following empirical likelihood function for the quantile medical cost y_p:

$$\begin{aligned} L_{IF}(\delta, y_p) = L_{IF}(\delta, x_r) = sup\{\delta^{n_0}(1-\delta)^{n_1} \prod_{j=1}^{n_1} p_j : p_1 \geq 0, ..., p_{n_1} \geq 0, \\ \sum_{j=1}^{n_1} p_j = 1, \sum_{j=1}^{n_1} p_j \hat{\gamma}_j(\delta, x_r) = 0\} \end{aligned} \tag{3.17}$$

We introduce multipliers λ and η, then the Lagrange function can be written as

$$\begin{aligned} G = n_0 \log \delta + n_1 \log(1-\delta) + \sum_{j=1}^{n_1} \log p_j \\ + \eta(\sum_{j=1}^{n_1} p_j - 1) - n_1 \lambda (\sum_{j=1}^{n_1} p_j \hat{\gamma}_j(\delta, x_r)) \end{aligned} \tag{3.18}$$

Setting the partial derivative of the Lagrangian function G with respect to p_j to 0 gives

$$\frac{\partial G}{\partial p_j} = \frac{1}{p_j} + \eta - n_1 \lambda \hat{\gamma}_j(\delta, x_r) = 0 \tag{3.19}$$

So

$$0 = \sum_{j=1}^{n_1} p_j \frac{\partial G}{\partial p_j} = n_1 + \eta \tag{3.20}$$

Thus we will have $\eta = -n_1$.

To determine $L_{IF}(\delta, y_p)$, according to equations (3.20), (3.21), using the Lagrange mul-

tipliers, we can easily get

$$p_j = \frac{1}{n_1\big(1 + \lambda\hat{\gamma}_j(\delta, x_r)\big)},\ j = 1, 2, \cdots, n_1,$$

where λ is the solution of the equation

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\hat{\gamma}_j(\delta, x_r)}{1 + \lambda\hat{\gamma}_j(\delta, x_r)} = 0. \tag{3.21}$$

A modified Newton's method can be used to solve the equation[51]. And once the value of λ is obtained, the profile log-likelihood is

$$l_{IF}(\delta, x_r) = n_0 \log\delta + n_1 \log(1 - \delta) - \sum_{j=1}^{n_1}\log(1 + \lambda\hat{\gamma}_j(\delta, x_r)) \tag{3.22}$$

We can define the likelihood ratio statistic as

$$R_{IF}(x_r) = 2\{\max_{\delta, x_r} l_{IF}(\delta, x_r) - \max_{\delta} l_{IF}(\delta, x_r)\}. \tag{3.23}$$

Let x_{r0} be the true value of x_r. Under some regularity conditions, we can prove that that $R_{IF}(x_{r0})$ converges in distribution to a χ^2 random variable with 1 degree of freedom as $n \to \infty$. Then, an EL-based confidence interval for x_{r0} can be constructed by using the asymptotic distribution of $R_{IF}(x_{r0})$.

3.3 Simulation Studies

This section reports the simulation results conducted to evaluate the small sample performances of the proposed nonparametric methods on the zero-inflated mean and median medical costs with censored observations.

Simulation settings for these studies were adopted from studies similar to Jeyarajah *et al.*(2019) [28] with moderate changes.

We assume that all observations were independently and identically distributed. The

positive total cost for each observation contains three cost components in different periods. For the j_{th} observation, $M_j(0)$ is the initial diagnostics cost at the beginning of the study; b_j is the follow up annual cost with T_j as the survival time, and d_j is the terminal death cost. Hence, the total cost is given by:

$$M_j = M_j(0) + b_j T_j + d_j I(T_j \leq L),\ j = 1, \cdots, n,$$

where $L = 10$ is a fixed time to end of experiment. The following two scenarios of simulation are used to generate cost data.

Scenario 1: $M_j(0), b_j$ and d_j, are generated from the following models :

$$M_j(0) \sim exp(N(0,1) * 1 + 5) + 1000$$

$$b_j \sim U[0,1] * 1000 + 1000$$

$$d_j \sim exp(N(0,1) * 1.5 + 6) + 1000$$

where $N(0,1)$ represents a random number from the standard normal distribution, and $U[0,1]$ represents a random number from the uniform distribution on [0,1].

Scenario 2: $M_j(0), b_j$ and d_j, are generated similarly as in scenario 1 with modifications on parameter values:

$$M_j(0) \sim exp(N(0,1) * 1.5 + 5) + 1000$$

$$b_j \sim U[0,1] * 3000 + 1000$$

$$d_j \sim exp(N(0,1) * 1.8 + 6) + 1000$$

Scenarios 1 is designed to generate cost data with moderate skewness and variability, while scenario 2 generates cost data with higher skewness and variability. Survival time T_j was generated from $U[0,10]$ distribution on years and an exponential distribution with a

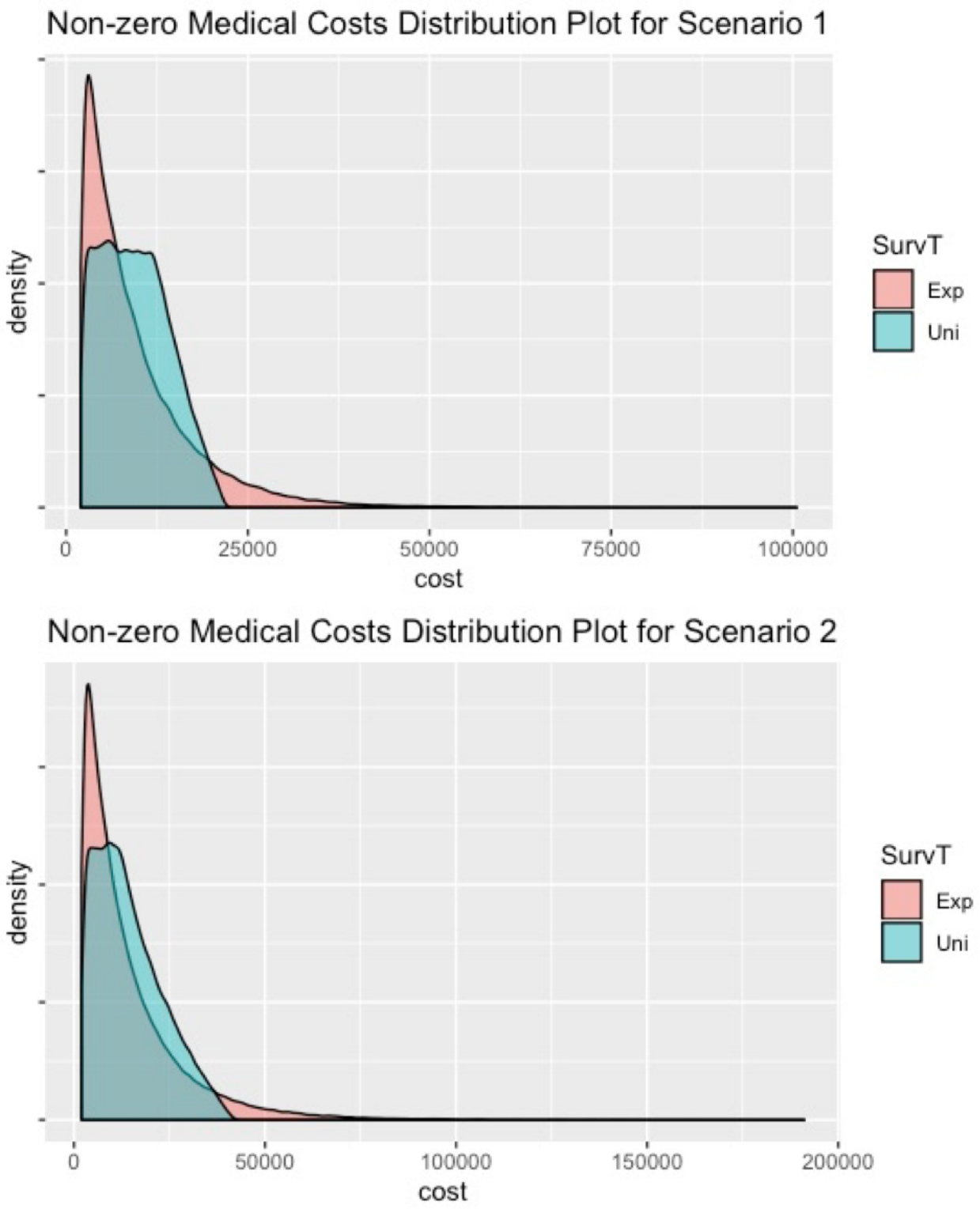

Figure 3.1 Medical cost distributions in two simulation scenarios (Exp: exponentially distributed survival time; Uni: Uniformly distributed survival time)

mean of 5 years for each of these two scenarios. Figure 3.1 provides a general view of the cost distributions in two scenarios that we generated without taking zero costs into account.

Let the number of zero costs follows binomial distribution given small sample sizes equal to n=50, 100, 200 but with successful probability parameter p=0.1, 0.2, 0.3.

Then, the true mean costs given 0.1, 0.2, 0.3 zero costs probabilities are \$8,561.75, \$7,610.45, and \$6,659.14 given uniform survival distribution and \$8,530.42, \$7582.59, and \$6634.77 given exponential distribution respectively in scenario 1. The estimated skewness parameters are 0.116, 0.204, 0.376 for uniform survival distribution and 2.096, 2.108, 2.191 for exponential survival distribution.

In scenario 2, the true mean costs given 3 different zero probabilities are \$13,075.82, \$11,622.94, and \$10,172.67 given uniform survival distribution and \$13,028.56, \$11580.93, and \$10133.33 given exponential distribution respectively. The corresponding skewness parameters were 0.608, 0.665, 0.798 and 2.476, 2.505, 2.601 respectively.

Considering generating censoring variable C_j from Uniform$[0, 20]$ and Uniform$[0, 12.5]$, while the corresponding two levels of censorship will be light censoring, resulting in 25-30% censoring, and the heavy censoring, resulting in 40-45% censoring.

Table 3.1 and Table 3.2 present the coverage probabilities and the median lengths of these intervals at a 95% confidence level in scenario 1 with different survival distributions. While in each table, we specify subcategories as censoring type, sample size, zero percentage, inferential methods for zero parameters, and normal approximation methods (SW) or Empirical Likelihood methods (ELI) for positive mean costs. From Table 3.1, we observed that when the sample size is as small as less than 200, the proposed confidence intervals using the MOVER methods by combining HK and ELI have overall better coverage probabilities than intervals using a normal approximation. From Table 3.2, we can see that when survival distribution is exponential, MOVER confidence intervals using the SW estimator seems not working well given under all conditions we set, while the coverage probabilities from ELI based methods are much closer to the nominal level when sample size reaches 200.

From Table 3.3 and Table 3.4, which are simulation studies in scenario 2, we observe

that the overall trend is similar to what we got in scenario 1. Since cost data distribution in scenario 2 has higher skewness, we can see that the coverage probabilities for zero-inflated mean costs are a little bit lower than what we got from scenario 1 when choices of methods are the same. But we can still observe that even when censoring is heavy, our ELI based nonparametric methods still have noticeable coverage accuracy nearly close to 95% nominal level.

Table 3.5 to Table 3.8 present coverage probabilities for upper quantile zero-inflated medical costs in two same scenarios using the empirical likelihood inferential approaches described in Section 3.2. We can easily tell that in both scenarios 1 and 2, the higher the upper quantile, the higher coverage probabilities. More specifically, for 70% or 80% quantiles, as long as sample size reach at least 200, coverage probabilities generated from EL based methods can always reach a 95% nominal level. However, for 90% or higher quantiles, the coverage probabilities are much lower. Probably, it is due to the reason that for small sample size, impact factors like zero proportion, right skewness could result in a long right tail, especially under exponential survival distribution condition. Also, a large proportion of censoring cost data will not provide enough information to generate an accurate estimation. In order to verify our assumption, we decide to increase the sample size to 1000 in simulation experiments under exponential survival distribution in both scenarios. The simulation results are shown in Table 3.9, we can see that compared to sample size equals to 200, coverage probabilities are improved by around 30%, which is acceptable.

In summary, our simulation studies show that the proposed confidence intervals for zero-inflated mean perform overall better than existing normal approximation-based intervals. As sample size increases, the performances of all the types of confidence intervals improve. And for upper 70% and 80% quantile costs, our proposed methods perform overall quite well in terms of coverage probabilities. Last but not least, if we want to make inference for 90% or higher quantiles, under scenarios similar to our simulation settings, we recommend to use our methods to make inference when the sample size is greater than 1000.

Table 3.1 Coverage probabilities and (median lengths) of 95% confidence intervals for the mean cost with uniform survival distribution in Scenario 1

Censoring Type		Zero Percentage	δ=0.1	δ=0.2	δ=0.3
		Survival Dist	Uni	Uni	Uni
Sample	CI Estimation	True Mean	8561.75	7610.45	6659.14
Size	for δ	Skewness	0.116	0.204	0.376
Light Censoring					
n=50	Zou	SW	0.942(2670)	0.944(2446)	0.968(2264)
		ELI	0.944(1748)	0.922(1872)	0.924(1793)
	LZT	SW	0.904(2646)	0.962(2500)	0.962(2368)
		ELI	0.946(1836)	0.924(1998)	0.924(1895)
	Li	SW	0.972(2862)	0.968(2807)	0.994(2745)
		ELI	0.952(1806)	0.940(2039)	0.922(1926)
	HK	SW	0.978(2811)	0.978(2789)	0.986(2779)
		ELI	0.952(1924)	0.936(2002)	0.928(1970)
n=100	Zou	SW	0.955(1865)	0.952(1763)	0.986(1612)
		ELI	0.968(2055)	0.954(2134)	0.952(2167)
	LZT	SW	0.962(1863)	0.982(1798)	0.990(1684)
		ELI	0.962(2018)	0.952(2251)	0.958(2134)
	Li	SW	0.960(2006)	0.972(2010)	0.990(1973)
		ELI	0.970(2103)	0.964(2322)	0.958(2219)
	HK	SW	0.980(2011)	0.992(1992)	0.988(1982)
		ELI	0.974(2146)	0.962(2396)	0.964(2243)
n=200	Zou	SW	0.982(1329)	0.966(1234)	0.970(1148)
		ELI	0.976(2112)	0.970(2334)	0.958(2256)
	LZT	SW	0.958(1335)	0.974(1276)	0.988(1212)
		ELI	0.972(2134)	0.972(2379)	0.964(2380)
	Li	SW	0.988(1418)	0.966(1423)	0.992(1405)
		ELI	0.978(2285)	0.966(2496)	0.962(2411)
	HK	SW	0.948(1415)	0.998(1395)	0.990(1316)
		ELI	0.980(2375)	0.970(2532)	0.964(2473)
Heavy Censoring					
n=50	Zou	SW	0.888(2805)	0.832(2586)	0.878(2303)
		ELI	0.864(1805)	0.844(1966)	0.824(1911)
	LZT	SW	0.920(2766)	0.892(2665)	0.902(2477)
		ELI	0.860(1883)	0.852(1972)	0.832(1935)
	Li	SW	0.928(2923)	0.902(2861)	0.930(2826)
		ELI	0.866(1904)	0.848(1954)	0.830(2129)
	HK	SW	0.930(2973)	0.894(2856)	0.940(2843)
		ELI	0.862(1948)	0.852(2092)	0.838(2183)
n=100	Zou	SW	0.868(1980)	0.882(1812)	0.922(1678)
		ELI	0.934(2430)	0.926(2536)	0.924(2356)
	LZT	SW	0.890(2005)	0.946(1901)	0.940(1778)
		ELI	0.936(2461)	0.930(2521)	0.928(2278)
	Li	SW	0.948(2136)	0.968(2089)	0.958(2092)
		ELI	0.934(2502)	0.928(2380)	0.934(2419)
	HK	SW	0.928(2105)	0.966(2082)	0.968(2058)
		ELI	0.938(2509)	0.934(2416)	0.938(2522)
n=200	Zou	SW	0.930(1400)	0.942(1304)	0.930(1216)
		ELI	0.952(2425)	0.964(2476)	0.942(2491)
	LZT	SW	0.922(1419)	0.968(1342)	0.952(1262)
		ELI	0.954(2310)	0.962(2418)	0.948(2394)
	Li	SW	0.924(1492)	0.954(1474)	0.962(1446)
		ELI	0.958(2568)	0.962(2643)	0.944(2648)
	HK	SW	0.932(1488)	0.978(1471)	0.992(1439)
		ELI	0.962(2581)	0.968(2658)	0.948(2858)

Table 3.2 Coverage probabilities and (median lengths) of 95% confidence intervals for the mean cost with exponential survival distribution in Scenario 1

Censoring Type		Zero Percentage	δ=0.1	δ=0.2	δ=0.3
		Survival Dist	Exp	Exp	Exp
Sample	CI Estimation	True Mean	8530.42	7582.59	6634.77
Size	for δ	Skewness	2.096	2.108	2.191
Light Censoring					
n=50	Zou	SW	0.868(2821)	0.858(2631)	0.828(2331)
		ELI	0.894(1748)	0.882(1734)	0.844(1856)
	LZT	SW	0.886(2823)	0.876(2705)	0.908(2490)
		ELI	0.898(1881)	0.896(1876)	0.866(1878)
	Li	SW	0.840(2936)	0.904(2868)	0.940(2786)
		ELI	0.910(1873)	0.910(1840)	0.862(2176)
	HK	SW	0.922(3041)	0.898(2888)	0.898(2757)
		ELI	0.912(1871)	0.914(1929)	0.878(2105)
n=100	Zou	SW	0.898(2036)	0.866(1860)	0.922(1722)
		ELI	0.922(2504)	0.922(2461)	0.920(2319)
	LZT	SW	0.902(2057)	0.922(1954)	0.896(1759)
		ELI	0.924(2686)	0.920(2975)	0.918(2247)
	Li	SW	0.892(2169)	0.918(2080)	0.908(1989)
		ELI	0.920(2715)	0.922(2565)	0.922(2596)
	HK	SW	0.894(2139)	0.908(2078)	0.902(1964)
		ELI	0.924(1856)	0.928(2526)	0.926(2574)
n=200	Zou	SW	0.866(1448)	0.882(1328)	0.912(1221)
		ELI	0.952(2975)	0.950(2973)	0.948(2505)
	LZT	SW	0.898(1463)	0.924(1365)	0.920(1271)
		ELI	0.958(2968)	0.954(3012)	0.946(2512)
	Li	SW	0.902(1508)	0.912(1463)	0.912(1400)
		ELI	0.956(3011)	0.952(3094)	0.954(2921)
	HK	SW	0.916(1477)	0.910(1452)	0.908(1421)
		ELI	0.958(3101)	0.958(3074)	0.954(2889)
Heavy Censoring					
n=50	Zou	SW	0.752(2921)	0.694(2640)	0.766(2451)
		ELI	0.756(1975)	0.742(1983)	0.702(1898)
	LZT	SW	0.710(2888)	0.712(2671)	0.810(2580)
		ELI	0.754(1897)	0.744(1796)	0.710(1888)
	Li	SW	0.766(3062)	0.724(2935)	0.784(2780)
		ELI	0.758(1984)	0.742(1807)	0.718(1894)
	HK	SW	0.822(3099)	0.756(2921)	0.782(2778)
		ELI	0.768(1959)	0.752(1889)	0.724(2063)
n=100	Zou	SW	0.810(2141)	0.770(1953)	0.738(1782)
		ELI	0.892(2302)	0.884(2226)	0.844(2268)
	LZT	SW	0.848(2182)	0.788(1993)	0.798(1848)
		ELI	0.898(2506)	0.880(2459)	0.842(2272)
	Li	SW	0.788(2204)	0.776(2045)	0.818(1995)
		ELI	0.896(2701)	0.890(2714)	0.860(2461)
	HK	SW	0.842(2269)	0.738(2101)	0.746(1970)
		ELI	0.906(2551)	0.898(2781)	0.866(2509)
n=200	Zou	SW	0.766(1497)	0.728(1394)	0.726(1244)
		ELI	0.936(2915)	0.932(3264)	0.922(2827)
	LZT	SW	0.788(1536)	0.762(1441)	0.846(1314)
		ELI	0.936(3110)	0.930(3240)	0.926(2905)
	Li	SW	0.796(1602)	0.744(1467)	0.822(1468)
		ELI	0.942(3072)	0.934(3388)	0.930(3271)
	HK	SW	0.766(1595)	0.748(1544)	0.758(1447)
		ELI	0.940(3166)	0.938(3492)	0.932(3203)

Table 3.3 Coverage probabilities and (median lengths) of 95% confidence intervals for the mean cost with uniform survival distribution in Scenario 2

Censoring Type		Zero Percentage	δ=0.1	δ=0.2	δ=0.3
		Survival Dist	Uni	Uni	Uni
Sample	CI Estimation	True Mean	13075.82	11622.94	10172.67
Size	CI for δ	Skewness	0.608	0.665	0.798
Light Censoring					
n=50	Zou	SW	0.922(4969)	0.922(4616)	0.940(4217)
		ELI	0.920(3624)	0.922(2795)	0.924(2852)
	LZT	SW	0.958(5094)	0.952(4628)	0.944(4310)
		ELI	0.918(3550)	0.926(2751)	0.930(2865)
	Li	SW	0.952(5161)	0.982(5036)	0.962(4748)
		ELI	0.922(3553)	0.930(2790)	0.928(3009)
	HK	SW	0.948(5131)	0.956(5038)	0.972(4797)
		ELI	0.924(3641)	0.934(2894)	0.928(3212)
n=100	Zou	SW	0.918(3548)	0.914(3293)	0.968(2969)
		ELI	0.948(3912)	0.944(3144)	0.932(3299)
	LZT	SW	0.928(3584)	0.962(3317)	0.952(3094)
		ELI	0.950(4008)	0.954(3224)	0.938(3290)
	Li	SW	0.978(3696)	0.978(3544)	0.978(3428)
		ELI	0.954(3046)	0.954(3370)	0.942(3573)
	HK	SW	0.954(3727)	0.950(3599)	0.980(3405)
		ELI	0.956(3144)	0.952(3233)	0.946(3503)
n=200	Zou	SW	0.928(2499)	0.905(2328)	0.954(2111)
		ELI	0.960(4271)	0.952(4442)	0.938(4291)
	LZT	SW	0.924(2555)	0.926(2402)	0.938(2194)
		ELI	0.964(4130)	0.958(4460)	0.940(4228)
	Li	SW	0.946(2671)	0.938(2564)	0.958(2433)
		ELI	0.964(4201)	0.960(4546)	0.944(4350)
	HK	SW	0.938(2606)	0.962(2565)	0.988(2438)
		ELI	0.968(4493)	0.962(46019)	0.952(4547)
Heavy Censoring					
n=50	Zou	SW	0.862(5231)	0.904(4796)	0.918(4389)
		ELI	0.824(2549)	0.806(3546)	0.784(3753)
	LZT	SW	0.858(5345)	0.890(4835)	0.922(4513)
		ELI	0.836(2905)	0.814(3607)	0.792(3758)
	Li	SW	0.846(5400)	0.886(5165)	0.936(4908)
		ELI	0.856(2754)	0.848(3709)	0.812(3843)
	HK	SW	0.856(5491)	0.894(5136)	0.940(5005)
		ELI	0.862(2792)	0.858(3734)	0.808(3702)
n=100	Zou	SW	0.932(3787)	0.914(3427)	0.924(3116)
		ELI	0.934(3835)	0.922(3973)	0.924(3231)
	LZT	SW	0.934(3784)	0.900(3571)	0.924(3267)
		ELI	0.934(3806)	0.924(3938)	0.928(3284)
	Li	SW	0.912(3963)	0.924(3775)	0.974(3564)
		ELI	0.942(4220)	0.928(4114)	0.926(3538)
	HK	SW	0.898(3935)	0.932(3748)	0.958(3609)
		ELI	0.940(4218)	0.932(4250)	0.934(3683)
n=200	Zou	SW	0.914(2725)	0.898(2447)	0.934(2248)
		ELI	0.940(4289)	0.936(4266)	0.934(2382)
	LZT	SW	0.868(2757)	0.928(2532)	0.928(2300)
		ELI	0.944(4290)	0.938(4451)	0.932(2419)
	Li	SW	0.920(2820)	0.918(2684)	0.966(2572)
		ELI	0.942(4248)	0.942(4523)	0.938(2778)
	HK	SW	0.904(2797)	0.924(2705)	0.964(2561)
		ELI	0.952(4361)	0.948(4479)	0.936(2455)

Table 3.4 Coverage probabilities and (median lengths) of 95% confidence intervals for the mean cost with exponential survival distribution in Scenario 2

Censoring Type		Zero Percentage	δ=0.1	δ=0.2	δ=0.3
		Survival Dist	Exp	Exp	Exp
Sample	CI Estimation	True Mean	13028.56	11580.93	10133.33
Size	CI for δ	Skewness	2.476	2.505	2.601
Light Censoring					
n=50	Zou	SW	0.888(5233)	0.896(4759)	0.828(4311)
		ELI	0.924(4612)	0.912(4714)	0.904(3856)
	LZT	SW	0.912(5289)	0.884(4838)	0.874(4378)
		ELI	0.928(4534)	0.920(4779)	0.910(3880)
	Li	SW	0.868(5407)	0.876(5065)	0.906(4807)
		ELI	0.932(4585)	0.918(4796)	0.912(4010)
	HK	SW	0.892(5480)	0.880(5001)	0.946(4788)
		ELI	0.934(4675)	0.922(4832)	0.922(4273)
n=100	Zou	SW	0.868(3750)	0.902(3441)	0.916(3089)
		ELI	0.940(4955)	0.926(5134)	0.922(4267)
	LZT	SW	0.896(3768)	0.878(3464)	0.878(3186)
		ELI	0.944(5018)	0.932(5251)	0.926(4234)
	Li	SW	0.864(3866)	0.906(3680)	0.930(3430)
		ELI	0.950(5003)	0.944(5322)	0.930(4519)
	HK	SW	0.914(3912)	0.922(3685)	0.936(3444)
		ELI	0.952(5146)	0.952(5296)	0.926(4543)
n=200	Zou	SW	0.826(2677)	0.868(2458)	0.886(2191)
		ELI	0.946(5248)	0.942(5272)	0.932(4513)
	LZT	SW	0.890(2696)	0.888(2483)	0.864(2251)
		ELI	0.948(5136)	0.942(5398)	0.934(4595)
	Li	SW	0.888(2764)	0.902(2622)	0.918(2449)
		ELI	0.948(5206)	0.946(5239)	0.944(4826)
	HK	SW	0.898(2773)	0.898(2612)	0.918(2466)
		ELI	0.954(5324)	0.950(5302)	0.942(4870)
Heavy Censoring					
n=50	Zou	SW	0.738(5391)	0.704(4743)	0.744(4380)
		ELI	0.724(3588)	0.726(3666)	0.708(3711)
	LZT	SW	0.716(5421)	0.738(4865)	0.746(4455)
		ELI	0.720(3624)	0.714(3672)	0.722(3735)
	Li	SW	0.814(5661)	0.746(5211)	0.810(4792)
		ELI	0.732(3772)	0.728(3754)	0.732(3829)
	HK	SW	0.714(5475)	0.792(5180)	0.756(4799)
		ELI	0.742(3795)	0.744(3792)	0.738(3883)
n=100	Zou	SW	0.806(3989)	0.796(3629)	0.782(3285)
		ELI	0.914(4811)	0.912(3989)	0.914(4246)
	LZT	SW	0.804(3977)	0.782(3691)	0.836(3360)
		ELI	0.922(4849)	0.920(4028)	0.918(4286)
	Li	SW	0.844(4141)	0.818(3864)	0.814(3496)
		ELI	0.920(4125)	0.918(4144)	0.922(4498)
	HK	SW	0.826(4110)	0.824(3834)	0.844(3593)
		ELI	0.920(4203)	0.914(4206)	0.922(4537)
n=200	Zou	SW	0.828(2884)	0.840(2631)	0.796(2324)
		ELI	0.924(5005)	0.916(4394)	0.922(4487)
	LZT	SW	0.826(2902)	0.812(2634)	0.828(2423)
		ELI	0.922(5183)	0.918(4467)	0.920(4499)
	Li	SW	0.838(2941)	0.822(2778)	0.846(2586)
		ELI	0.926(5104)	0.926(4623)	0.926(4725)
	HK	SW	0.812(2982)	0.848(2796)	0.862(2595)
		ELI	0.932(5148)	0.924(4640)	0.928(4800)

Table 3.5 Coverage probabilities and (median lengths) of 95% confidence intervals for different upper quantiles with uniform survival distribution using EL method in Scenario 1

	Zeros%	δ=0.1	δ=0.2	δ=0.3
Censoring Type	Surv Dist	Uni	Uni	Uni
	True 70%	11637.92	11049.07	10294.60
Sample Size	True 80%	13365.47	12885.95	12319.94
	True 90%	15712.28	15370.83	14949.24
Light Censoring				
	70%	0.951(5051.15)	0.972(5680.73)	0.942(6528.07)
n=50	80%	0.844(3939.13)	0.874(4525.95)	0.862(4452.26)
	90%	0.662(3205.35)	0.711(3772.94)	0.702(3920.08)
	70%	0.978(5647.23)	0.979(4581.60)	0.945(4733.85)
n=100	80%	0.902(3179.53)	0.906(4291.12)	0.962(4488.98)
	90%	0.795(3195.46)	0.811(3385.64)	0.846(3782.42)
	70%	0.996(5932.43)	0.987(4665.35)	0.952(4758.89)
n=200	80%	0.972(4919.04)	0.982(3998.67)	0.967(4579.30)
	90%	0.844(3669.32)	0.828(3264.74)	0.855(3912.49)
Heavy Censoring				
	70%	0.903(5376.24)	0.928(6779.83)	0.946(6294.96)
n=50	80%	0.812(4559.43)	0.762(5022.09)	0.804(4750.07)
	90%	0.645(3413.47)	0.664(3258.26)	0.678(4269.51)
	70%	0.961(6529.85)	0.980(8806.35)	0.983(6397.49)
n=100	80%	0.863(5041.57)	0.904(5225.42)	0.945(5548.90)
	90%	0.782(3838.35)	0.645(3028.92)	0.787(3903.22)
	70%	0.982(6493.32)	0.988(9008.04)	0.992(5149.37)
n=200	80%	0.966(5311.77)	0.959(5344.25)	0.984(5066.58)
	90%	0.823(3523.88)	0.801(3260.72)	0.825(4652.33)

Table 3.6 Coverage probabilities and (median lengths) of 95% confidence intervals for different upper quantiles with exponential survival distribution using EL method in Scenario 1

	Zeros%	δ=0.1	δ=0.2	δ=0.3
Censoring Type	Surv Dist	Exp	Exp	Exp
	True 70%	10159.77	9260.26	8264.43
Sample Size	True 80%	13264.31	12342.60	11327.51
	True 90%	18607.36	17709.86	16672.73
Light Censoring				
	70%	0.982(7141.49)	0.980(8288.51)	0.956(7787.57)
n=50	80%	0.866(5879.92)	0.876(7304.58)	0.842(6593.13)
	90%	0.412(4310.95)	0.442(4933.03)	0.522(5260.52)
	70%	0.991(7353.85)	0.983(7030.55)	0.982(5934.03)
n=100	80%	0.962(5734.72)	0.981(6752.78)	0.942(6135.94)
	90%	0.422(4252.37)	0.508(4327.65)	0.626(5054.42)
	70%	0.992(7215.92)	0.988(7101.39)	0.989(6322.84)
n=200	80%	0.988(5038.65)	0.983(6648.75)	0.952(6005.62)
	90%	0.455(4809.41)	0.516(4442.08)	0.662(5445.88)
Heavy Censoring				
	70%	0.975(7358.16)	0.964(8045.94)	0.960(7879.33)
n=50	80%	0.689(6374.73)	0.667(6712.11)	0.644(7096.32)
	90%	0.348(4235.75)	0.323(4290.27)	0.266(4662.39)
	70%	0.989(9884.97)	0.986(8395.05)	0.982(7909.66)
n=100	80%	0.898(8699.34)	0.922(8171.56)	0.924(7513.86)
	90%	0.362(6396.49)	0.383(7002.85)	0.345(5742.74)
	70%	0.987(9692.23)	0.987(7752.72)	0.986(7882.26)
n=200	80%	0.942(8802.48)	0.979(9210.99)	0.939(7617.03)
	90%	0.408(6667.35)	0.465(8675.71)	0.502(5265.38)

Table 3.7 Coverage probabilities and (median lengths) of 95% confidence intervals for different upper quantiles with uniform survival distribution using EL method in Scenario 2

	Zeros%	δ=0.1	δ=0.2	δ=0.3
Censoring Type	Surv Dist	Uni	Uni	Uni
	True 70%	17338.14	16082.71	14602.97
Sample Size	True 80%	21343.37	20215.14	18926.07
	True 90%	26878.27	26049.98	25115.66
Light Censoring				
	70%	0.923(10861.78)	0.948(10618.60)	0.966(12499.95)
n=50	80%	0.911(8548.82)	0.848(9242.31)	0.912(10069.92)
	90%	0.745(8244.95)	0.689(7840.53)	0.726(9249.23)
	70%	0.956(10853.13)	0.947(10044.01)	0.934(9859.88)
n=100	80%	0.912(9277.03)	0.902(10749.63)	0.917(10227.24)
	90%	0.726(6721.32)	0.734(8518.87)	0.786(7206.292)
	70%	0.966(10426.23)	0.962(10068.53)	0.957(10135.73)
n=200	80%	0.921(9696.09)	0.919(9458.36)	0.924(9416.25)
	90%	0.778(7856.10)	0.752(7824.15)	0.747 (7423.64)
Heavy Censoring				
	70%	0.828(11220.68)	0.813(11902.81)	0.943(13822.83)
n=50	80%	0.625(11539.49)	0.622(9837.01)	0.705(11041.65)
	90%	0.529(9449.291)	0.346(8259.96)	0.482(7733.47)
	70%	0.937(13382.53)	0.918(11911.99)	0.928(11332.53)
n=100	80%	0.823(11849.94)	0.838(10623.09)	0.812(12352.37)
	90%	0.577(7858.99)	0.468(7784.18)	0.495(8519.43)
	70%	0.942(12358.08)	0.945(11890.02)	0.936(11425.35)
n=200	80%	0.879(10197.56)	0.886(9869.53)	0.882(9997.20)
	90%	0.731(7998.15)	0.718(8041.74)	0.722(8325.56)

Table 3.8 Coverage probabilities and (median lengths) of 95% confidence intervals for different upper quantiles with exponential survival distribution using EL method in Scenario 2

	Zeros%	δ=0.1	δ=0.2	δ=0.3
Censoring Type	Surv Dist	Uni	Uni	Uni
	True 70%	15082.34	13565.75	11882.97
Sample Size	True 80%	20423.63	18808.48	17023.99
	True 90%	30099.09	28431.65	26599.22
Light Censoring				
	70%	0.928(12138.26)	0.943(12801.77)	0.941(12374.08)
n=50	80%	0.796(11883.43)	0.896(11733.40)	0.916(12713.61)
	90%	0.496(9996.69)	0.508(10833.87)	0.548(11471.09)
	70%	0.953(11559.28)	0.958(10096.68)	0.948(10725.68)
n=100	80%	0.884(11873.02)	0.901(12915.98)	0.927(12353.37)
	90%	0.498(10209.34)	0.516(10056.62)	0.536 (11191.47)
	70%	0.959(12209.23)	0.966(10134.37)	0.963(10025.32)
n=200	80%	0.895(12235.40)	0.926(11321.28)	0.938(12315.39)
	90%	0.508(10179.33)	0.511(10886.36)	0.545(10804.05)
Heavy Censoring				
	70%	0.866(14609.53)	0.878(13053.86)	0.913(14852.1)
n=50	80%	0.692(14656.29)	0.676(11798.76)	0.689(13077.05)
	90%	0.325(12967.80)	0.329(11215.08)	0.383(12592.35)
	70%	0.905(13081.28)	0.919(12414.32)	0.917(11720.2)
n=100	80%	0.729(13463.52)	0.742(13779.02)	0.757(14990.73)
	90%	0.334(11614.45)	0.362(11749.35)	0.409(12122.2)
	70%	0.921(12678.27)	0.928(11850.39)	0.933(10274.55)
n=200	80%	0.763(13223.16)	0.808(13451.70)	0.798(12709.56)
	90%	0.442(12005.32)	0.426(11194.65)	0.456(12038.19)

Table 3.9 Coverage probabilities and (median lengths) of 95% confidence intervals for 90% quantiles with exponential survival distribution under sample size 1000

Scenarios		δ=0.1	δ=0.2	δ=0.3
Scenario 1	True 90%	18607.36	17709.86	16672.73
Light Censoring	90%	0.794(4953.26)	0.821(5006.28)	0.835(5365.10)
Heavy Censoring	90%	0.726(4390.05)	0.733(4820.66)	0.752(5008.33)
Scenario 2	True 90%	30099.09	28431.65	26599.22
Light Censoring	90%	0.736(10335.35)	0.757(10096.04)	0.789(11256.58)
Heavy Censoring	90%	0.689(8992.16)	0.714(9709.35)	0.729(10137.23)

3.4 Real Data Analysis

We use all the methods discussed above to analyze the same CIDS dataset as we used in section 2.5. But this time, we not only use 430 positive costs but also consider the rest 229 zero cost observations. For simplicity, if we assume the number of zero costs follows a binomial distribution. Then, the estimated zero-proportion $\hat{\delta}$ is 0.35. To have a better insight into the dataset in addition to what we showed in section 2.5. The survival curve and costs distributions are displayed in Figure 3.2.

Table 3.10 Estimation and 95% confidence intervals of mean and upper quantile costs for CIDS dataset

Parameters	Estimated Costs	95% Confidence Intervals
Mean	47477.59	(36474.02, 61861.34)
70% quantile	74257.44	(58415.10, 90099.78)
80% quantile	94153.61	(77023.37, 111283.85)
90% quantile	112598.97	(94612.85, 130584.99)

Regarding the inference for zero-inflated mean costs, we first calculate the mean cost of the positive costs using Bang and Tsiatis' estimator. Then the estimated positive mean cost is \$73042.44, and the corresponding EL-based 95% level confidence interval is (\$56485.65, \$94892.34). From simulation studies, we found that if using HK methods to make inference for the zero proportion, generally, we will have higher coverage probabilities. As a result, considering HK's zero proportion confidence intervals (0.315, 0.380), after applying

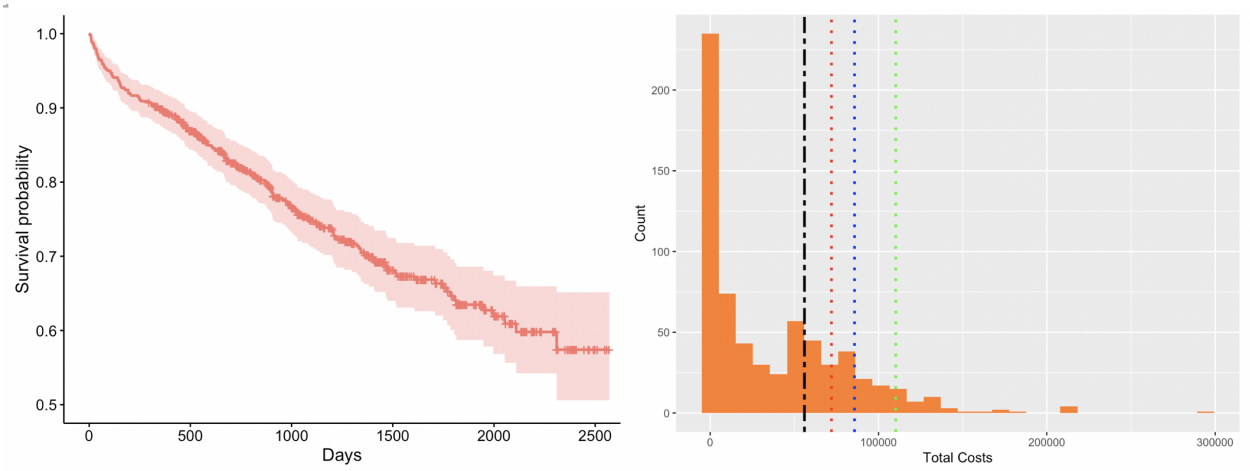

Figure 3.2 KM survival curves for CIDS cases with 95% confidence intervals (I); Histogram of total cost distribution (II) (dashed lines represent mean(black), 70%(red), 80%(blue), and 90%(green) quantiles for uncensored cases)

MOVER method, the estimated zero-inflated mean with an adjusted 95% confidence interval is \$47477.59 (\$36474.02, \$61861.34).

Then we try to make inference for zero-inflated upper 70%, 80%, and 90% quantile costs using the proposed empirical likelihood intervals. Results from data analysis are presented in Table 3.10. The estimated 70%, 80%, and 90% quantile costs of CIDS patients are \$74257.44, \$94153.61 and \$112598.97 respectively, and the corresponding 95% level confident intervals are (\$58415.10, \$90099.78), (\$77023.37, \$111283.85), and (\$94612.85, \$130584.99).

3.5 Discussion

In this part of the dissertation, we propose EL-based confidence intervals for the mean and upper quantiles medical costs with censored and zero costs data. Simulation studies have shown that for the mean cost, the proposed MOVER confidence intervals using Hasan and Krishnamoorthy (HK) and EL-based methods, in general, have better coverage accuracy than the other combinations using normal approximation approaches. Especially, as long as the sample size is greater than 200, even under heavy censoring condition, in both scenar-

ios, compared with the existing normal approximation-based intervals which need complex variance estimation, HK fiducial quantity plus influence function-based empirical likelihood intervals can have a good coverage probability and acceptable interval length. It shows that when making inference for zero-inflated mean cost, our proposed nonparametric method has an overall better small sample performance under heavy censoring and heavy skewness. And also, since there are no existing methods to make inference for quantile censored medical costs with a nonnegligible proportion of zero costs, we propose to use an empirical likelihood-based inference using influence functions and zero proportion. The result shows that the small sample performance of empirical likelihood-based inference for upper quantiles of zero-inflated medical costs is really good in terms of coverage probability under almost all scenarios given different censoring or zero proportions. In future studies, we hope we can improve the coverage probabilities for 90% or even higher quantiles given exponential survival distribution. Finally, for the same CIDS dataset described in first part of the dissertation, considering a 35% of zero cost proportion, we provide nonparametric confidence intervals for zero-inflated mean and 70%, 80% and 90% upper quantiles, which could be a valuable supportive information if we want to accurately evaluate patients' cost distribution.

PART 4

BREAST CANCER LOCAL RECURRENCE PREDICTION USING A NOVEL QUANTITATIVE CENTROSOMAL AMPLIFICATION SCORE

The last part of this dissertation focus on how to quantify centrosome amplification information of DCIS patients to construct a new CA score. Then, we figured out an optimal threshold of CAS that can be used to stratify patients into significant different recurrence-free survival groups while patients with high CAS will have a much higher risk of DCIS recurrence. It is organized as follows. In Section 4.1, we describe some current findings of DCIS recurrence-free survival prediction. In Section 4.2, we describe the data source and preparation for our investigation process. Also, the description of data structure and simple demographic information will be shown. In Section 4.3, we will be describing the notations and equations we used for CASi, CASm, and CAStotal that will be used in the model. In Section 4.4, we describe our statistical analysis procedures by comparing the univariate and multivariate performance of CAS for local recurrence prediction by using the training set and validate the results using the testing set. In Section 4.5, we evaluate the robustness of model performance for potential new data. In Section 4.6, we discuss the result and future studies.

4.1 Existing Methods

Current predictors of recurrence risk for DCIS are based on routinely used clinico-pathological parameters. However, such simple approaches do not show consistency and repeatability in predicting recurrence risks ([5, 43]). In addition, these tools do not integrate informative molecular predictors and underestimate DCIS heterogeneity. Solin *et al.*[56] proposed Oncotype Dx Breast DCIS score, based on proliferation gene group score, progesterone receptor(PR), and GSTM1, performed for patients with DCIS treated with surgical

excision without radiation. They defined a proliferation group score (PGS), which equals to $(Ki67 + STK15 + Survivin + CCNB1 + MYBL2)/5$. Solin *et al.*[56] defined the unscaled $DCIS\ score_{\mu}$ as $DCIS\ score_{\mu} = 0.31 \times PGS - 0.08 \times PR - 0.09 \times GSTM1$. Finally, the rescaled DCIS score in the range from 0 to 100 is as following:

$$DCIS\ score = 66.7 \times DCIS\ score_{\mu} + 10.0$$

Then they specified patients' risk categories based on DCIS score as low, intermediate and high-risk groups. Even though this cancer-related genes expression based assay has some values in predicting local recurrence, the poor stratification of different risk levels between patients in their cohorts made this prognostic value defective as it is lack of ability to be a powerful tool.

As we mentioned in Section 1.5, since we are aware of the interesting correlation between centrosome amplification and early-stage cancers, we want to find out a reasonable way to quantify the row centrosome amplification information. Choi *et al.*[10] applied centrosome amplification information in predicting head and neck cancer survival risk. They proposed following plain quantities: 1. percentage of nuclei who have an abnormal number of centrosomes as numerical CA%; 2. percentage of nuclei who have an aberrant large volume of centrosome exist as structural CA%; 3. total percentage of either numerical or structural centrosome amplification as Total CA%. Their results show that compared to other pathological characteristics and cancer-related biomarkers, for cancer patients with the negative HPV test result, total centrosome amplification percentage score can be used as a good predictor for long term overall survival risk, while patients with higher centrosome amplification percentage score could have much worse survival status.

However, due to the reason that that quantitation is rustic and has a limited performance for HPV negative patients only, we want to develop a novel well-designed standard score that has a logical structure and can be widely applied to different types of cancers like breast, pancreas, colon or prostate.

4.2 Notations and Quantification Formulas

The core innovation that we did in this paper is to quantify centrosome amplification as following procedures and formulas. Centrosomes in breast tissue were categorized into independent identifiable centrosomes or iCTRs and megacentrosomes or mCTRs. iCTR numbers and boundaries were clear to identify, and their volumes fall into the range of centrosome volumes found in normal breast tissue stained. After measuring volumes of centrosomes from normal samples from healthy individuals, the volume range for a normal centrosome was determined accordingly. We evaluated centrosome volumes in these samples as described in the analysis section. We chose the smallest and largest values of individual centrosome volume from normal tissue as the normal volume range for healthy breast tissue. The mean volume of centrosomes in normal breast cells ranged from 0.2-0.74 μm^3. Centrosomes with volumes greater than 0.74 μm^3 were categorized as megacentrosomes. More frankly, mCTRs are centrosomes with unusual large volumes and are considered to represent structurally amplified centrosomes. The numbers and volumes of iCTR and mCTR associated with each nucleus were recorded. Graph illustrations of numerical and structural amplification can be checked from Figure 4.1. We also brought up concepts like 'Severity' and 'Frequency' to indicate the source of amplification both numerically or structurally. A more scientific boundary between high and low severity or frequency still needs to be quantified and validated.

Next, we need to explain the notations used in our proposed formulas. For CAS_i, R_c is the greatest number of centrosomes exist in a normal breast cell. The common acceptable number is 2. For a given cell, let N_i be the number of iCTRs that exist, which is most likely greater than 2 in abnormal cells. Thus, $(N_i - R_c)$ indicates the number of excess centrosomes presented with numerical amplification. R_c is the range of values for the number of centrosomes present in a nucleus of a normal cell, which is 2 here. p_i is the percentage of such nuclei with iCTRs greater than 2. β_i is a scaling factor to ensure that both CAS_i(numerical) and CAS_m(structural) to be weighted equally when used for constructing CAS_{total}, which is the sum of CAS_i and CAS_m.

The so-called severity component of CAS_i, $mean\left(\frac{N_i - R_c}{R}\right)$, represented in (4.1), quantifies how severe the numerical amplification is. For example, it measures the average deviance level that the numerical amplification exceeds the baseline value of R_c in nuclei which could possibly carry more than three iCTRs. As a result, it is easy to notice that cancer cells with only one or two iCTRs will not make contributions to this component. We first assign indicators for those nuclei with the number of centrosomes greater than range R_c for the number of normal centrosomes we used. Second, among those nuclei with a large number of centrosomes, we take the average number of centrosomes greater than R_c. Then, a standardization procedure, also can be understood as a linear transformation, using a mean redundant number of centrosomes divided by a fixed range of R_c, was applied to create a severe numerical amplification measurement. Finally, given percentage of abnormal nuclei p_i and scaling factor β_i, the scaled frequency component of the CAS_i score can be calculated as p_i/β_i.

$$\begin{aligned} CASi &= mean\left(\frac{N_i - R_c}{R_c}\right)\frac{percentage\,(N_i > R_c)}{\beta_i} \\ &= \left(\frac{\sum_{i=1, N_i > R_c}^{N}(N_i - R_c)}{\sum_{i=1}^{N} I(N_i > R_c)}\frac{1}{R_c}\right)\frac{p_i}{\beta_i} \end{aligned} \tag{4.1}$$

For the CAS_m formula, shown in (4.2), V_{im} is the volume of the m_{th} mCTR in the i_{th} nucleus. Then, p_m is the percentage of cells with mCTRs, where a mCTR is defined as a large centrosome whose volume exceeds the critical value V_c that we predefined. Such a threshold V_c for a given tissue is the maximum volume of a normal centrosome in that tissue, which is 0.74 μm^3 for breast tissue. Similarly to β_i, β_m is a scaling factor used to ensure that both CAS_i and CAS_m contribute equally towards CAS_{total}. $\delta_{V_{im}}$ is the standard deviation of the volume of mCTRs. And simply speaking, N is the total number of nuclei.

For each mCTR, a standard score was computed based on the formula below, which evaluates the extent to which the volume of that mCTR exceeded the critical value, which is the standardized evaluation for $V_{im} - V_c$. It is computed by measuring the relativeness of

deviation to the baseline, which is achieved by dividing by the standard deviation of $\delta_{V_{im}}$.

In the next step, that value we got, also known as the standardized severity of structural amplification per nucleus, was multiplied by the number of mCTRs in the given cell. Finally, all values were averaged to obtain the severity score for structural amplification. Similar to CAS_i, the scaled frequency component of CAS_m can be constructed by using p_m, percentage of overall centrosome volumes greater than V_c, divided by scaling factor β_m. As claimed before, the components, CAS_i and CAS_m, contribute equally to the CAS_{total} score. To ensure such equal contribution, the scaling factors β_i and β_m we used in this paper adjusted the value range of CAS_i and CAS_m to be ranged approximately from 0 to 3. The value of CAS_i scaling factor β_i used here is 0.1 for breast tissue and value of β_m used here is 0.148.

$$\begin{aligned} CASm &= mean\left(\frac{V_{im}-V_c}{\sigma_{V_{im}}}\right)\frac{percentage\,(V_{im}>V_c)}{\beta_m} \\ &= \left(\frac{\sum_{i=1}^{N}\sum_{m=1}^{N_i}(V_{im}-V_c)I(V_{im}>V_c)}{\sigma_{V_{im}}}\right)\frac{p_m}{\beta_m} \end{aligned} \tag{4.2}$$

4.3 Data Preparation

4.3.1 Cancer Tissue Sample Preparation

In a 24-year study period since 1988, the breast cancer tissue sections of patients diagnosed with DCIS that we used in this paper were obtained from Nottingham City Hospital in the United Kingdom. Our training set contains consecutive pure DCIS patients with or without adjuvant radiotherapy([61, 60, 41]. In order to reduce the confounding effects on the study outcome, our experiment team members make sure that our available tissue samples should show free surgical margins greater than 2mm. All cases were histologically reviewed, and diagnoses were confirmed by three independent pathologists. Each of the tissue sample data was matched to its clinicopathologic variables such as age at diagnosis, tumor size, nu-

clear grade, presence of comedo type necrosis, and information about adjuvant radiotherapy treatment, recurrence-free survival time period defined by the time frame started after the initial treatment to first time local recurrence in the form of DCIS. Date of initial diagnosis, date of surgery, date of recurrence, and patient status at last contact were recorded and used for generating survival time([60]). No patients received adjuvant systemic therapy.

4.3.2 Data Structure for Centrosome Records

Raw image data were processed using volume rendering software to determine the volume of each centrosome inside the tissue sample. To exclude nonspecific signals, common background subtraction was applied to all images. This parameter was derived by first measuring the average diameter of approximately one hundred centrosomes available in tissue samples, and then we use this measurement as the background subtraction threshold. Finally, data from all optical sections were ordered to enable volume measurement for each centrosome. The final data of volumes of all centrosomes were then compared to a maximum intensity projection image and centrosomes for each cell were quantified based on closeness to their associated nuclei. The number and volume of all centrosomes associated with each nucleus in the tumor area were recorded. Sample data structure for each patient is presented in Table 4.1, while the number of centrosome counts and volume shown in this table are casually created for illustration only.

Table 4.1 Sample Centrosome Score Record Data Structure Illustration

Nuclei ID	Centrosome Count	Volume 1	Volume 2	Volume 3	Volume 4	Volume 5
1	1	1.02				
2	3	0.22	0.47	0.31		
3	2	0.13	0.16			
4	1	0.41				
5	5	0.26	0.21	0.32	0.19	0.24
6	4	1.24	0.95	0.88	0.47	
7	2	0.28	0.43			
...						

4.4 Statistical Analysis

4.4.1 Discovery Procedure

We found that in the initial training set, among those 133 patients(details in Table 4.2), 28 patients developed local recurrence. The median age at diagnosis was 58 years with a range from 41 to 84, and the median follow-up was 132 months, while patients' recurrence-free survival times range from 14 to 333 months. Out of 133 patients, around 42% (n=55) received radiotherapy. Higher nuclear grade, the presence of comedo necrosis and the use of radiotherapy were clinicopathological parameters that showed proportional differences between recurring and no local recurrence patient subgroups (Table 4.2). However, only high grade and comedo necrosis showed associations with recurrence-free survival in a univariable cox regression analysis (Table 4.3). The interesting thing is that none of these clinicopathological variables showed any significant association with recurrence-free survival in multivariate analyses (Table 4.4), which indicates the limitation of using those traditional clinicopathological variables to predict local recurrence for DCIS in our training set. And also, since there is an extremely imbalanced comprise of tumor grade level, in order to reduce the potential strong misleading effect, in the rest of studies, we decided to use only high-grade cases for analysis due to a dominant number of high-grade cases. Table 4.4 and 4.5 show similar results for high grade only cases, which includes 118 out of 133 cases with 21 of which have a local recurrence. Then, we decide to use a larger extended dataset to support our prediction result. Those demographic frequency distributions are shown in Table 4.6. It is comprised of 177 high-grade DCIS patients, out of which 33 patients presented with local recurrence. The median age of these patients was 57 years, and the median follow-up was 100 months. Out of 177 patients, about 33.3% (n=59) received radiotherapy. Age and presence of the comedo necrosis showed significant proportional differences between the local recurrence and no recurrence subgroups.

Centrosome numbers and volumes, evaluated and scored for numerical (CAS_i) and structural (CAS_m) centrosomal aberrations were integrated using our algorithm to generate

Table 4.2 Overall Clinical Charateristics Diagnosis for Training Set

Baseline Charateristics	**No Recurrence**	**Local Recurrence**	**p-value**
Patient Age,n(%)			
Age>50	87(82.86)	22(78.57)	0.6003
Age≤50	18(17.14)	6(21.43)	
Tumor Size,n(%)			
Size>16	51(48.57)	15(53.57)	0.6382
Size≤16	57(51.43)	13(46.43)	
Grade,n(%)			
High	97(92.38)	21(75.00)	0.0098
Mid and Low	8(7.62)	7(25.00)	
Comedo Necrosis,n(%)			
No	14(13.33)	8(28.57)	0.0538
Yes	91(86.67)	21(71.43)	
Radiotherapy,n(%)			
No	57(54.29)	21(75.00)	0.0480
Yes	48(45.71)	7(25.00)	
Receptor Status,n(%)			
ER-PR-HER2 Positive	3(2.86)	2(7.14)	0.6826
ER-PR Positive HER2 Negative	19(18.10)	7(25.00)	
HER2 Positive	9(8.57)	3(10.71)	
TNBC	9(8.57)	1(3.57)	
Missing	65(61.90)	15(53.57)	

Table 4.3 Univariate and Multivariate Cox Regression for Training Set

Variables	**Levels**	**Univariate Analysis**			**Multivariate Analysis**		
		HR	**p-value**	**95% CI**	**HR**	**p-value**	**95% CI**
CAStotal	High vs Low	6.337	<0.001	(2.196,18.287)	7.869	<0.001	(2.709,22.857)
Age	>50 vs ≤ 50	0.697	0.437	(0.280,1.733)	0.767	0.599	(0.284,2.068)
Grade	High vs Med&Low	0.317	0.009	(0.134,0.752)	0.257	0.072	(0.081,0.823)
Comedo Necrosis	Yes vs No	2.043	0.088	(0.899,4.640)	1.635	0.271	(0.681,3.926)
Radiotherapy	No vs Yes	1.946	0.128	(0.826,4.583)	1.470	0.403	(0.596,3.628)
Receptor Status	ER-PR-HER2 Positive	2.425	0.240	(0.553,10.640)	2.329	0.323	(0.435,12.456)
	ER-PR Positive HER2 Negative	1.719	0.194	(0.759,3.893)	2.044	0.163	(0.748,5.581)
	HER2 Positive	1.480	0.534	(0.430,5.089)	2.458	0.214	(0.595,10.151)
	TNBC	0.638	0.663	(0.084,4.821)	0.969	0.977	(0.120,7.835)

a composite CAS_{total} value for each sample of the training set (Figure 4.2). Interestingly, DCIS patients that developed local recurrence showed significantly higher CASi relative to no recurrence patients. These patients with local recurrence showed greater CAS_i severity and higher CAS_i frequency of numerical CA compared to no recurrence patients. Analysis of structural CA revealed that CAS_m was significantly higher for the local recurrence subgroup relative to no recurrence subgroup. DCIS with local recurrence exhibited greater CASm severity and CAS_m frequency of structural CA compared to no recurrence DCIS. Cumulatively, for high tumor grade cases, a summation of CAS_i and CAS_m generated CAS_{total}, which was significantly higher for DCIS patients with local recurrence relative to no recur-

Table 4.4 Overall Clinical Charateristics Diagnosis for High Grade Training Set

Baseline Charateristics	**No Recurrence**	**Local Recurrence**	**p-value**
Patient Age,n(%)			
Age>50	85(87.63)	15(71.43)	0.0612
Age≤50	12(12.37)	6(28.57)	
Tumor Size,n(%)			
Size>16	43(44.33)	11(52.38)	0.5019
Size≤16	54(55.67)	10(47.62)	
Comedo Necrosis,n(%)			
No	10(10.31)	4(19.05)	0.2615
Yes	87(89.69)	17(80.95)	
Radiotherapy,n(%)			
No	49(50.52)	15(71.43)	0.0811
Yes	48(49.48)	6(28.57)	
Receptor Status,n(%)			
ER-PR-HER2 Positive	3(3.09)	2(9.52)	0.2537
ER-PR Positive HER2 Negative	19(19.59)	7(33.33)	
HER2 Positive	9(9.28)	3(14.29)	
TNBC	9(9.28)	1(4.76)	
Missing	57(58.76)	8(38.10)	

Table 4.5 Univariate and Multivariate Cox Regression for High Grade Training Set

Variables	**Levels**	**Univariate Analysis**			**Multivariate Analysis**		
		HR	**p-value**	**95% CI**	**HR**	**p-value**	**95% CI**
CAStotal	High vs Low	9.453	0.0025	(2.197,40.666)	9.742	0.0026	(2.215,42.844)
Age	>50 vs ≤ 50	0.419	0.0734	(0.162,1.086)	0.353	0.0669	(0.116,1.075)
Comedo Necrosis	Yes vs No	1.684	0.3488	(0.566,5.007)	0.795	0.7266	(0.220,2.869)
Radiotherapy	No vs Yes	1.980	0.1581	(0.767,5.111)	2.582	0.0668	(0.936,7.119)
Receptor Status	ER-PR-HER2 Positive	3.545	0.1105	(0.749,16.774)	0.652	0.6949	(0.077 ,5.513)
	ER-PR Positive HER2 Negative	2.232	0.1210	(0.809,6.155)	1.495	0.4770	(0.494,4.527)
	HER2 Positive	2.575	0.1627	(0.682,9.719)	3.707	0.0841	(0.838,16.396)
	TNBC	0.998	0.9984	(0.125,7.993)	0.652	0.6949	(0.077,5.513)

rence patients.

After that, if we apply the same methodology to the testing set, we calculated CAS and found that DCIS cases with local recurrence exhibited higher CAS_{total} relative to no recurrence patients, while the result from Wilcoxon rank sum test of distribution differences indicating significance with p-value less than 0.0001 (Figure 4.5). Furthermore, similar trends were observed for other components of CAS as found in the training set; the mean values of CAS_i and CAS_m, including their severity and frequency components, were higher in the patient subgroup with local recurrence than in the no recurrence subgroup with p-values all less than 0.05. Collectively, our data strongly suggest a clear difference in centrosome aberrations between DCIS tumor tissues of patients with and without local recurrence.

In order to stratify the patients into high or low CAS groups and make comparisons,

Table 4.6 Overall Clinical Charateristics Diagnosis for High Grade Testing Set

Baseline Charateristics	**No Recurrence**	**Local Recurrence**	**p-value**
Patient Age,n(%)			
Age>50	116(80.56)	21(63.64)	0.0361
Age≤50	28(19.44)	12(36.36)	
Tumor Size,n(%)			
Size>16	83(57.64)	18(54.55)	0.7461
Size≤16	61(42.36)	15(45.45)	
Comedo Necrosis,n(%)			
No	15(10.42)	8(24.24)	0.0331
Yes	129(89.58)	25(75.76)	
Radiotherapy,n(%)			
No	93(64.58)	25(75.76)	0.2194
Yes	51(35.42)	8(24.24)	
Receptor Status,n(%)			
ER-PR-HER2 Positive	5(3.47)	4(12.12)	
ER-PR Positive HER2 Negative	31(21.53)	10(30.30)	
HER2 Positive	21(14.58)	5(15.15)	0.1716
TNBC	14(9.72)	2(6.06)	
Missing	73(50.69)	12(36.36)	

Table 4.7 Univariate and Multivariate Cox Regression for High Grade Testing Set

Variables	**Levels**	**Univariate Analysis**			**Multivariate Analysis**		
		HR	**p-value**	**95% CI**	**HR**	**p-value**	**95% CI**
CAStotal	High vs Low	6.098	<.0001	(2.513,14.795)	7.223	<.0001	(2.887,18.072)
Age	> 50 vs ≤ 50	0.475	0.0404	(0.234, 0.968)	0.426	0.0388	(0.189,0.957)
Comedo Necrosis	Yes vs No	2.137	0.0617	(0.963,4.740)	1.455	0.4210	(0.583, 3.631)
Radiotherapy	No vs Yes	1.459	0.3527	(0.658,3.238)	1.764	0.1807	(0.768,4.048)
Receptor Status	ER-PR-HER2 Positive	3.421	0.0333	(1.102,10.621)	1.413	0.5972	(0.392,5.089)
	ER-PR Positive HER2 Negative	1.701	0.2146	(0.735,3.938)	1.090	0.8588	(0.420,2.832)
	HER2 Positive	1.504	0.4435	(0.530,4.270)	2.013	0.2032	(0.685, 5.912)
	TNBC	1.029	0.9700	(0.230,4.602)	1.070	0.9300	(0.237,4.829)

we need to figure out a threshold for CAS. The optimal threshold we used here for CAS_{total} is 1.436, which comes from using a scan algorithm of all possible CAS values and then select the best one that leads to minimal p-value from the log-rank test for Kaplan-Meier survival curve differences. More specifically, our null hypothesis is that H_0: There is no difference between the two group survival curves. For group i, let O_{it} be the observed number of events at t, E_{it} be the observed number of events at t. Then our log-rank test statistic will be $\sum_{i=1}^{2} \frac{(\sum_t O_{it} - \sum_t E_{it})^2}{\sum_t E_{it}}$, which follows a χ^2 distribution with 1 degree of freedom.

Upon stratification of all training set patients into low and high CAS groups (Figure 4.6), we found that DCIS patients with high CAS_i were associated with poorer recurrence-free survival relative to those with low CAS_i. Similarly, high CAS_m was associated with poorer recurrence-free survival compared to low CAS_m. Finally, as our greatest interest target,

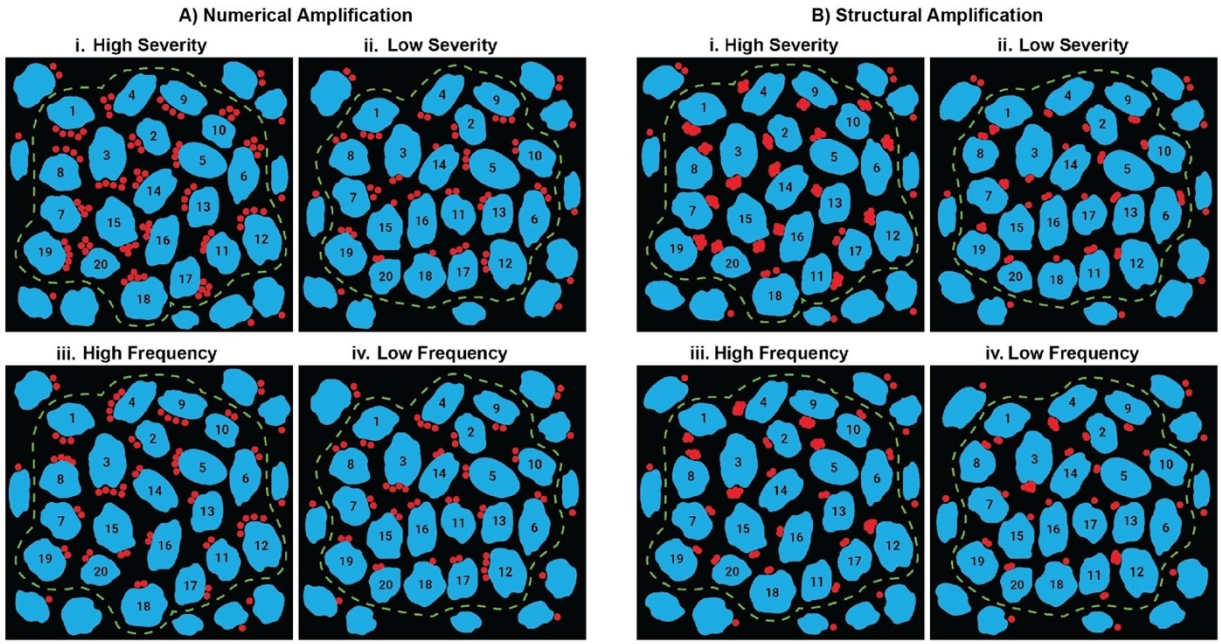

Figure 4.1 Schematic depicting numerical centrosome amplification and structural centrosome amplification

CAS_{total} stratified the high-risk and low-risk DCIS patients with high significance. We found that 90.4% of patients with local recurrence were in the high CAS_{total} group. In a univariate cox regression model, we can see that only CAS_{total} remained significantly associated with recurrence-free survival. Moreover, the association of recurrence risk with CAS_{total} remained significant even after accounting for potential confounders, including comedo necrosis, age and radiotherapy status while we noticed that in Table 4.5, only the hazard ratio between CAS_{total} High vs Low is as high as 9.742 with p-value equals 0.0026 in multivariate proportional hazard regression model.

To verify whether CAS_i, CAS_m, and CAS_{total} could be used to stratify patients in the testing set, we used pre-determined CAS cutoffs from the training set (Figure 4.3. We found that high CAS_i, CAS_m, and CAS_{total} were associated with poorer recurrence-free survival compared to low CAS_i, CAS_m, and CAS_{total}, respectively. Of the patients with local recurrence, approximately 81% of patients with local recurrence were classified into

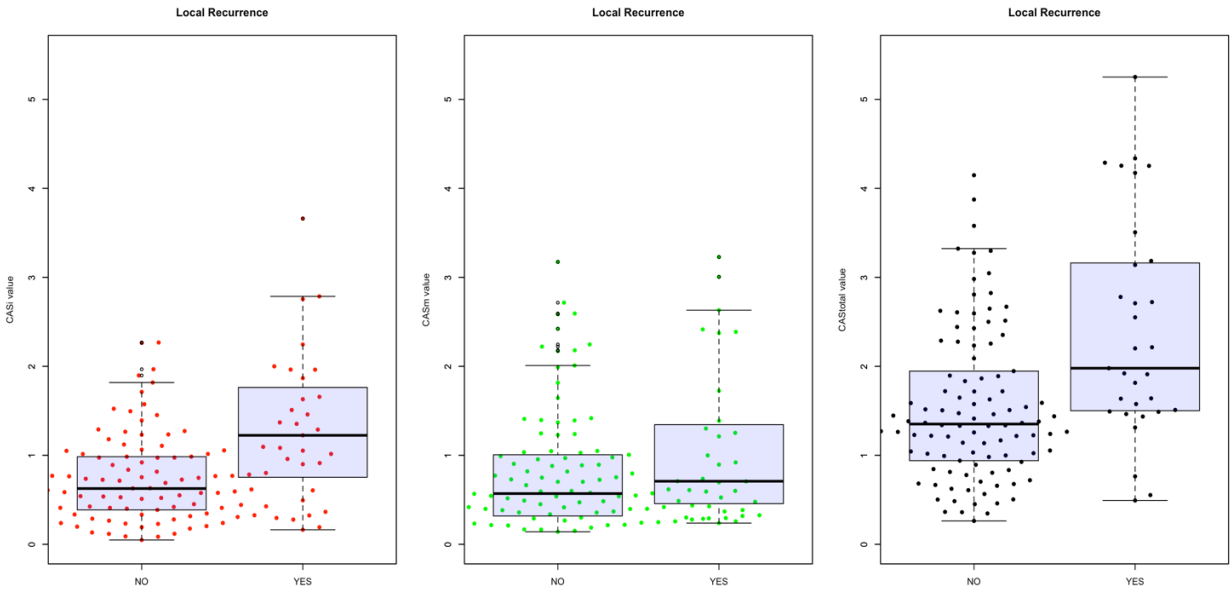

Figure 4.2 Distribution of CAS expression fro different recurrence status for training set

the high CAS_{total}(>1.436) subgroups. In both univariate and multivariate analyses after adjusting for potentially confounding effects of factors like age, grade, comedo necrosis, and radiotherapy, CAS_{total} was the strongest and most significant independent predictor of recurrence-free survival. For example, hazard ratios for CAS_{total} were higher than hazard ratios of all other clinicopathologic factors (Table 4.7). Additionally, in both the training set and testing set, the 10-year estimated risk of local recurrence increased continuously as the CAS increased. Those results collectively show that CAS can robustly predict 10-year local recurrence risk for DCIS patients from two different datasets.

In order to further verify that CAS could also be an indicator of any treatment effect. We decided to do more trials on CAS expression to investigate its ability to stratify high and low-risk groups after radiotherapy. In our radiotherapy group, we have in total of 59 patients, where 8 of those have DCIS recurrence. While for the no radiotherapy group, there are 25 out of 118 cases have local recurrence. From Figure 4.7, we can see that without radiotherapy, it is crystal clear to see that based on the optimal threshold we set before, all CAS_i, CAS_m

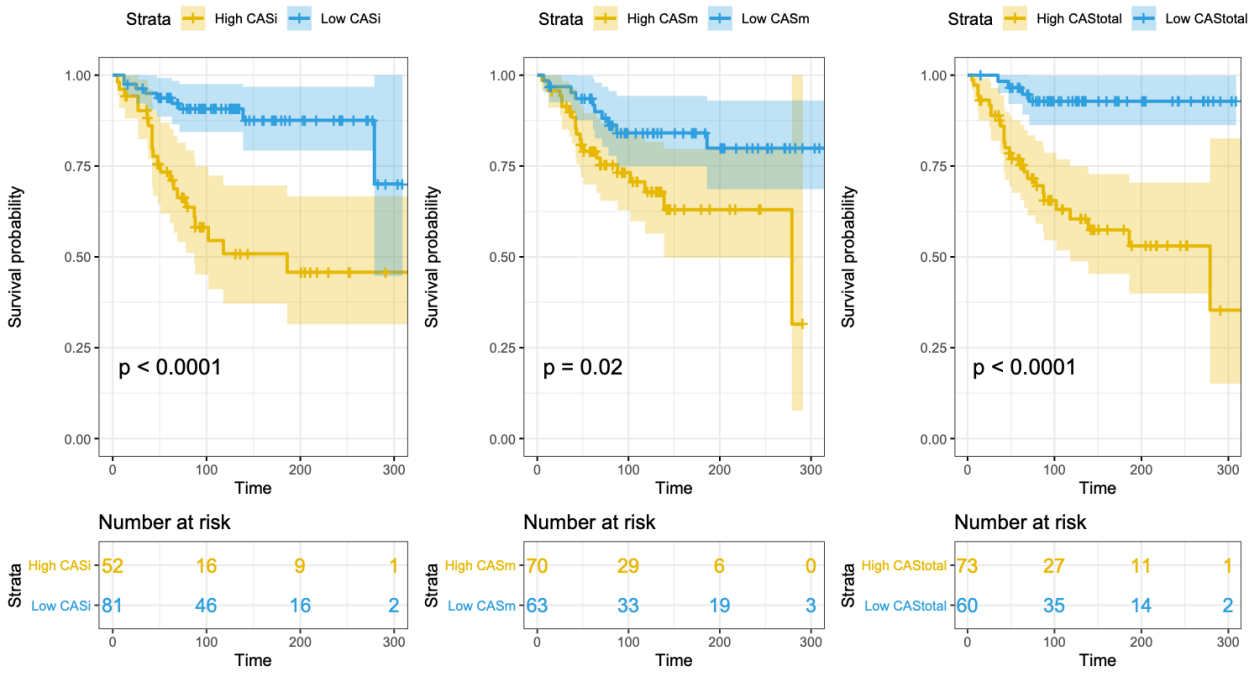

Figure 4.3 Kaplan meier survival plot for recurrence free survival for training set

and CAS_{total} can perfectly stratify recurrence-free survival difference. Especially, for patients with low CAS_{total}, after 10 years, the probability of local recurrence can still be greater than 0.9, however, a 10-year recurrence-free survival rate for patients in CAS_{total} high group will jump to only 0.5. Then, after radiotherapy treatment, from Figure 4.8, we noticed that radiotherapy truly has a positive effect on improving patients' recurrence-free survival rate. Under such conditions, CAS_m has a weak contribution to stratifying survival differences. However, CAS_i (p-value=0.00098) and CAS_{total} (p-value=0.055) are still working here with log-rank tests of KM survival curve differences being significant. In general, no matter whether patients' received radiotherapy treatment or not, patients with high CAS_{total} values always have a higher recurrence rate. Apparently, radiotherapy treatment worked as an effective treatment here since recurrence-free survival probabilities for both high CAS and low CAS groups are improved, and the stratification differences between groups were reduced.

Next, we evaluated the clinical significance of CAS by examining the associations of CAS with traditionally employed clinicopathological variables (i.e., age, tumor size, comedo necrosis, and radiotherapy). Our data shows that CAS_{total} provides clinically relevant prognostic information over and beyond what is provided by current clinicopathologic parameters alone. Given that high CAS is associated with more aggressive disease phenotypes, we not only observed the association of high CAS_{total} with higher recurrence rates but also found that CAS_{total} segments patient subgroups more deeply than traditional clinicopathologic parameters. For example, the recurrence rate forest plot (Figure 4.4) for high-grade DCIS patients in the training set showed that patients with comedo necrosis (red) are at high risk of recurrence with estimated rate 0.59 compared to the estimated recurrence rate 0.33 for patients who did not present comedo necrosis. When we further stratified these DCIS patients with comedo necrosis into high (green) and low (blue) CAS groups, we observed that inside comedo necrosis subgroup, the recurrence rate for the high CAS group (green) was 0.83 and recurrence rate for the low CAS subgroup (blue) was 0.10. Thus, we believe that CAS was able to more deeply segment the patients with comedo necrosis into high and low-risk local recurrence groups. Similar trends were clear to observe for tumor size, radiotherapy, and age subgroups. All those results suggest that centrosome amplification plays a nonnegligible role in tumor progression detection.

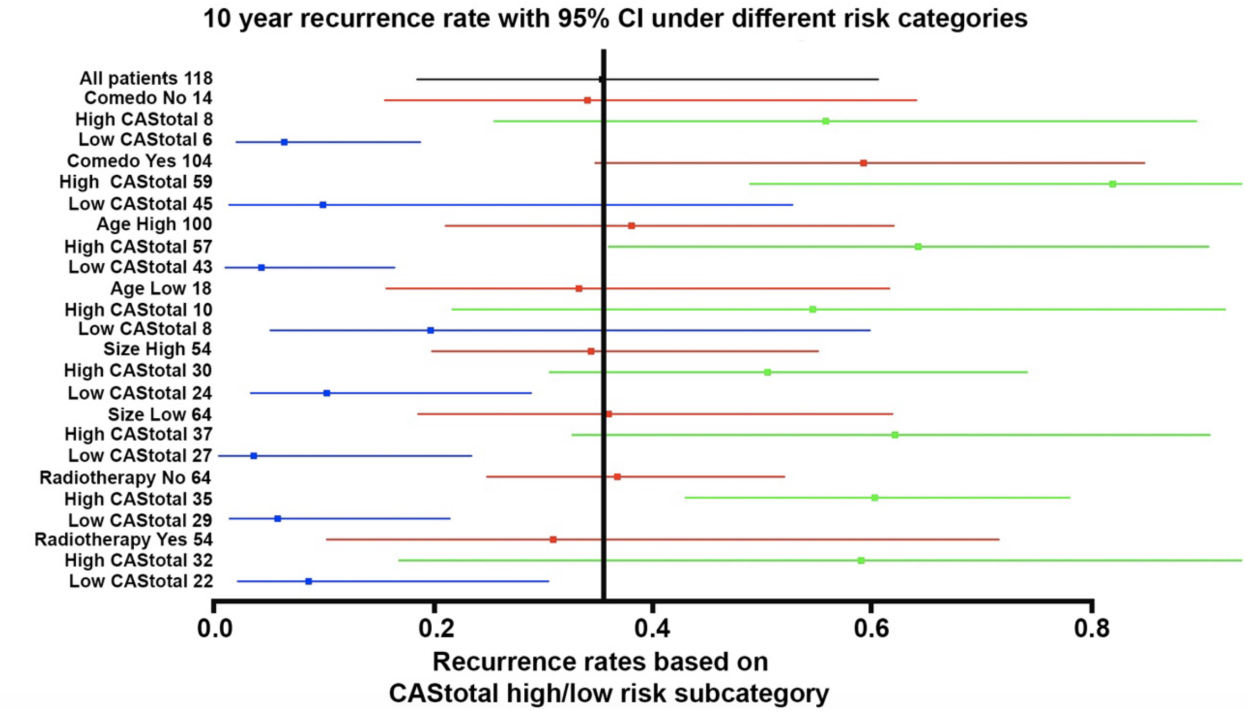

Figure 4.4 Forest plot showing 95% confidence interval for 10-year recurrence rate prediction

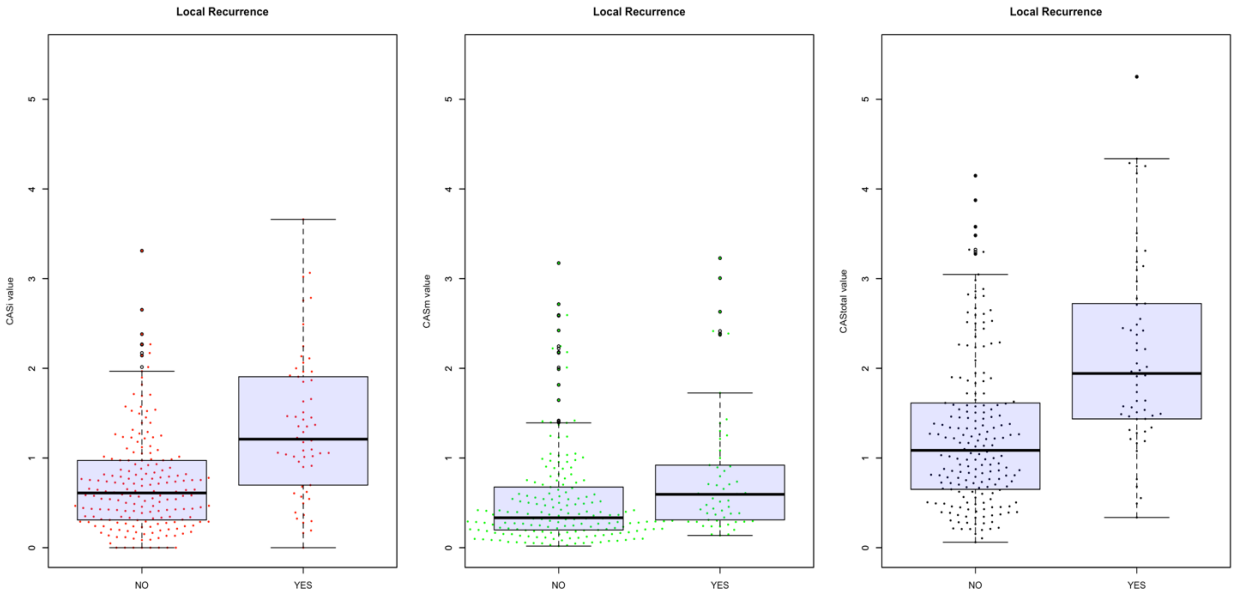

Figure 4.5 Distribution of CAS expression for different recurrence status for testing set(High Grade only)

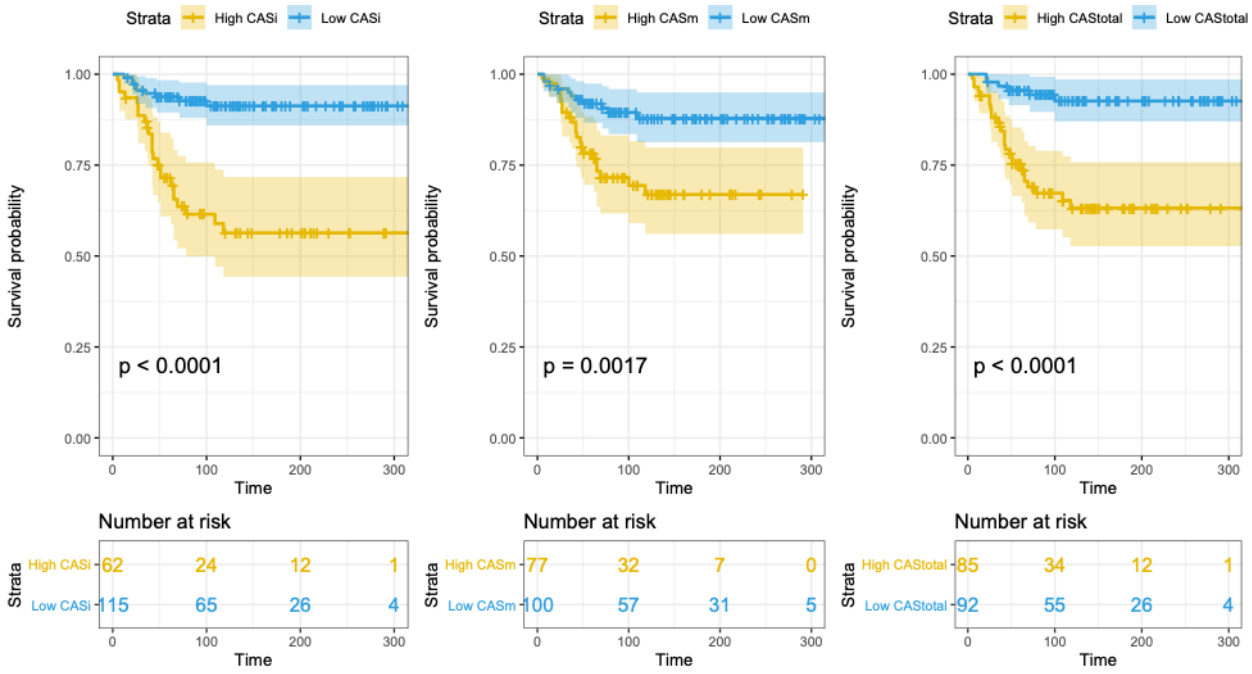

Figure 4.6 Kaplan meier survival plot for recurrence free survival for testing set (High Grade only)

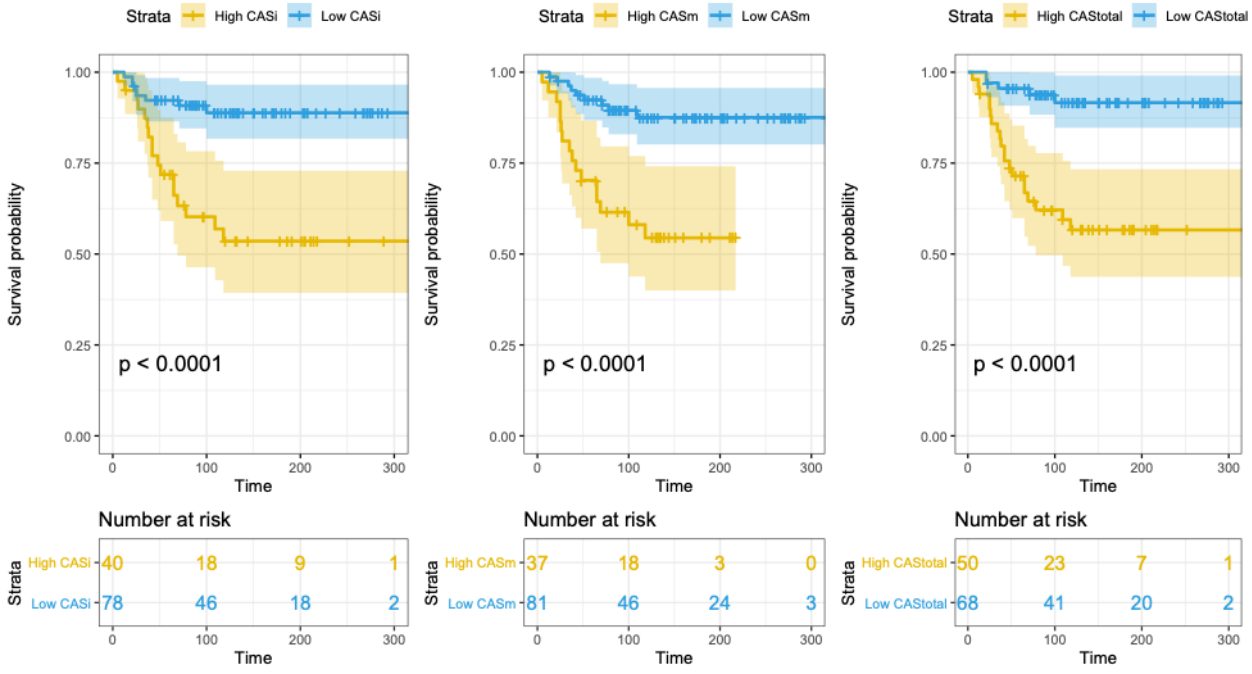

Figure 4.7 Kaplan meier survival plots for cases without radiotherapy

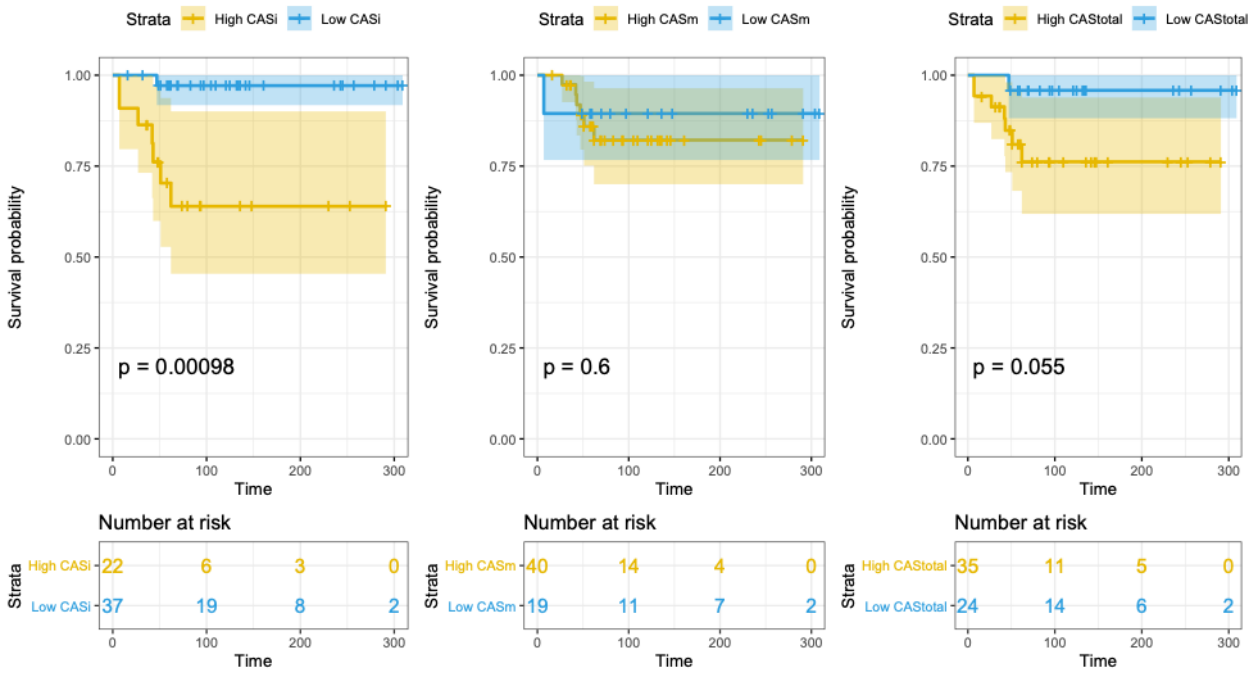

Figure 4.8 Kaplan meier survival plots for cases received radiotherapy

4.4.2 Robustness Diagonosis

In order to reduce the randomness effect of sample set selection to avoid the casual irreproducible findings. We decide to do a diagnosis analysis based on resampled datasets to support the robustness of our findings within the whole combined dataset we have. We first merged all available cases on hand to construct a candidate pool, which owns in total 252 cases within the original unadjusted study period considering both high and low tumor grade status, and then we can use the bootstrap technique to extract 500 small equal-sized resample datasets. For each resample set, there will be 100 patients included.

Due to the reason that our main focus of our investigation is the ability of CAStotal in stratifying patients' recurrence-free survival differences. So we decided to show 500 CAStotal high/low paired recurrence-free survival curves with 95% confidence limits in a collective plot (Figure 4.9), while blue area indicates the possible recurrence-free survival probability for any patients with CAS_{total} low and yellow area represents the possible probability for CAS_{total} high group. In Figure 4.9, we also put a dashed horizontal reference line which shows that

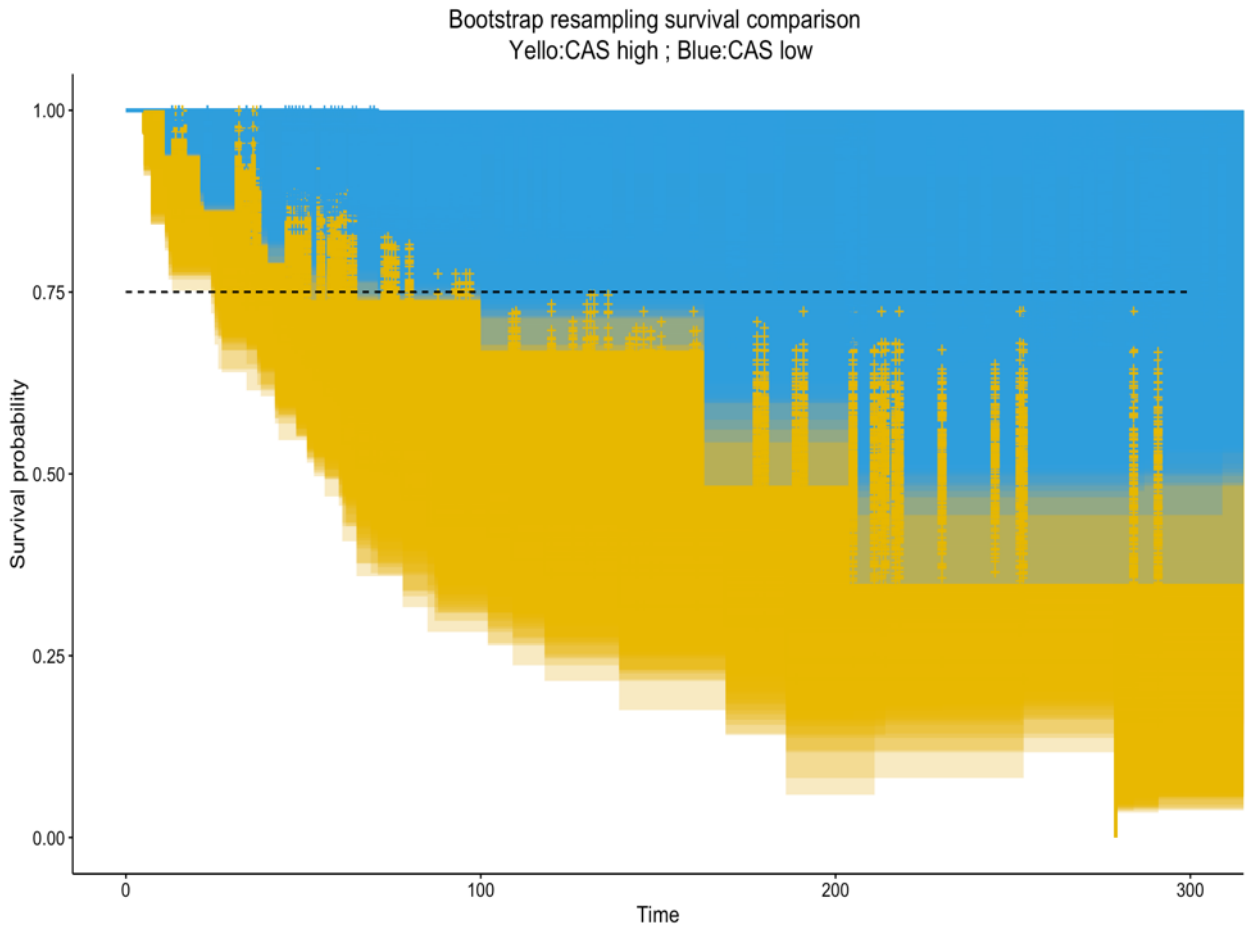

Figure 4.9 Kaplan meier survival plots generated from 500 bootstrap samples using CAStotal as stratification variable

after 10 years, 0.75 survival probability is still a clear boundary line for survival probability in CAS_{total} high or low-risk groups.

Additionally, for those 500 resample datasets, we perform univariate cox regression on CAS_{total} only (High vs Low). Interestingly, we found that even the least hazard ratio we can achieve is more than 2 given all the models we built are significant with a p-value less than 0.05 (Figure 4.10). What's more, while in a range from 2 to 20, hazard ratios between CAS_{total} high and low-risk groups appear to be around 5 to 6 in most of the settings.

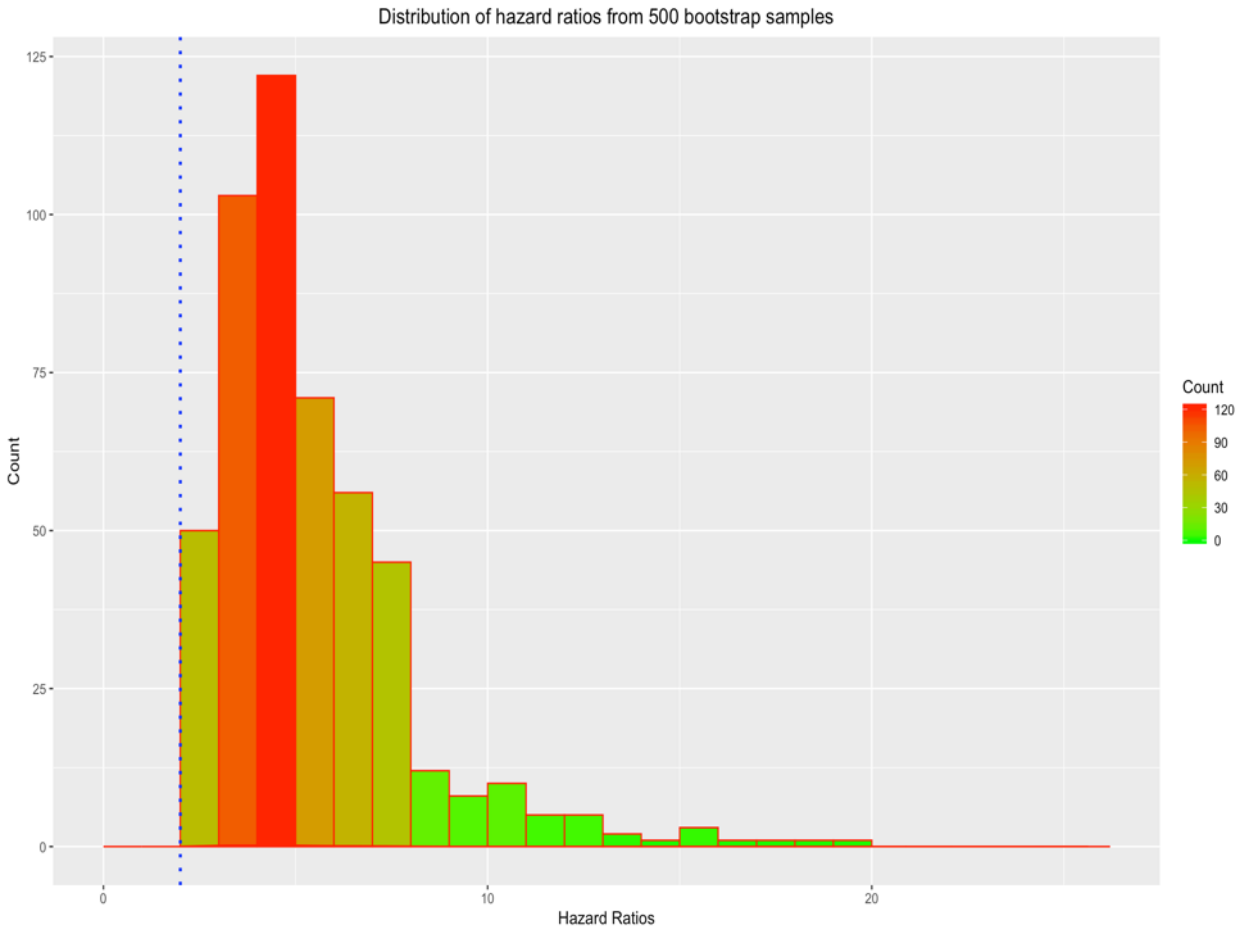

Figure 4.10 Distribution of Hazard Ratios of CAStotal High vs Low from bootstrap trials

Such diagnosis analysis above together strongly support the reliability of the models we built and the findings we got. We do believe that for any future studies, as long as we can follow the rules to collect new tumor tissues and calculate CAS_{total} value as shown in the method section, even a small sample around 100 cases is enough to be used as the stratification and prediction indicator for DCIS recurrence-free survival status.

Here, we have proved the robustness of our model based on the whole dataset we have collected. In the future, we still need out-of-bag multi-site testing datasets to do further validation with new patients' records while a diverse range of demographic characteristics exists. We hope to see the recurrence risk prediction ability of CAStotal can be consistent under all different scenarios.

4.5 Discussion

It has been demonstrated that DCIS exhibits great heterogeneity while its natural history is short of study. Models, complicated by prognostic evidence of patient age, tumor margins, grade, and size, with low accuracy for the prediction of local recurrence risk, results in over or under treatment. Amplified centrosomes are present in premalignant cells and increase as the disease evolves to abnormal structure, highlighting the potential involvement of centrosome amplification in tumor transformation and progression ([38]). Centrosome amplification, associated with high-grade DCIS, is demonstrated as a hallmark of cancers and is observable in >80% of breast tumors including preinvasive lesions([6]).

It has been previously shown that high levels of CAS are associated with poor progression-free survival in invasive breast tumors, and CAS is higher in the aggressive TNBC subtype compared to grade matched non-TNBCs ([50, 42]). This notion was further validated by analysis of the CA20 gene score, which is based on genes associated with centrosome amplification ([45]). Recent studies have reported that higher CAS induces high-grade features in breast cancers. As a result, CA has been associated with tumor evolution ([15]).

Although studies have reported that breast cancers exhibit structurally amplified centrosomes, they have not yet established the prognostic value of this structural CA ([21]).

This may be due, in part, to the cross-sectional approaches used in these studies, which have limitations to accurately capture the three-dimension size of the centrosome. Moreover, most studies ([12]) examining CA in breast cancers have not rigorously evaluated confounding effects of other clinicopathologic variables on the prognostic value of centrosome amplification.

Our new methodology uses quantitative centrosomal phenotyping and novel algorithms to measure both numerical and structural centrosomal aberrations in DCIS tumors. For each sample, a continuous CAS was computed that can categorize patients as having a high or low risk of local recurrence. Findings from our retrospective study of DCIS cases showed that patients with local recurrence exhibited higher CAS_{total} relative to no local recurrence patients. Our study is the first to show that organelle level differences could be used to distinguish DCIS patients with local recurrence from no recurrence patients and that high levels of both numerical and structural CA are associated with increased local recurrence in DCIS patients. Our results suggest that abnormal centrosomal homeostasis in DCIS potentially accelerate disease progression. We have demonstrated that CAS_{total} is significantly and independently associated with poor recurrence-free survival. In DCIS patient subsets, defined based on their clinical and histopathological parameters, stratification by CAS_{total} prognostically augmented several clinicopathologic parameters in determining the rate of recurrence. Among subsets of DCIS patients treated with surgery or those receiving additional adjuvant radiotherapy, CAS_{total} identified patients with a high risk of local recurrence. Thus, CAS_{total} can be used as a clinical tool to identify patients who can be safely treated with surgery alone, and those who will benefit from the inclusion of radiotherapy. Our centrosomal profiling methodology, which dichotomizes DCIS patients into high and low-risk categories, enables clear therapeutic decision making, and can substantially augment individualized management of DCIS based upon risk conferred by the patients centrosomal complement.

CAS, as the standardized expression of the severity and frequency of numerical and structural CA. Our study, the first to robustly quantify CA in pure DCIS samples, has contributed evidence supporting a model of CA driven DCIS progression into invasive breast

cancer. Our findings are consistent with previous findings that TNBC, the most aggressive subtype of breast cancer, exhibits the highest CAS among all breast cancer subtypes ([50, 15]). Centrosome profiling can complement clinicopathologic and genomic evaluation to provide a comprehensive description of disease status. An approach for future research is to profile CA in all the stages of tumor progression starting from abnormal progression to invasive and metastatic disease to evaluate if CAS can function as a biomarker for tumor evolution.

Compared to current commercially available Oncotype DCIS score, our quantitative centrosomal phenotyping methodology, capturing prognostic information from a broader space of biological pathways that are free in the biology of DCIS, is more broadly applicable and could be refined for other cancer types with CAS.

PART 5

CONCLUSIONS

In this dissertation, we have developed novel statistical methods to make inference for quantiles and mean of medical costs with censored data and zero costs with/without covariates. We also propose a new quantification method based on centrosome amplification and use it to construct a brand new predictor to predict cancer tumor progression status.

First, we propose EL-based methods to construct confidence regions for the median medical cost regression coefficients. Our simulation results show that the proposed influence function based empirical likelihood confidence regions have better coverage accuracy than the normal approximation-based regions for the regression coefficients. We also construct EL-based confidence intervals for the median cost at given covariates based on numerical approaches. Based on our simulation studies we recommend IFEL and JEL intervals for the median medical cost with covariates.

Next, we propose an integrated nonparametric method to construct confidence intervals for the mean and median of the zero-inflated medical costs with censored data. Simulation results show that the proposed influence function-based empirical likelihood confidence intervals perform well with censored and skewed cost data. Hence, we recommend using the proposed EL-based confidence intervals for mean and quantiles of censored medical costs given a large number of zeros, particularly when heavy censoring and severely skewness exist.

Finally, we propose to construct a novel quantity called CAS based on centrosome amplification information in DCIS breast cancer tissue. Then based on the training set, we find an optimal threshold that can be used to stratify patients into high CAS or low CAS groups. After that, we use the testing set to validate the results. Our research shows that compared with other existing demographic characteristics or biomarker predictors, patients with high CAS have a significantly higher chance to experience local tumor progression. We

also make inference on a 10-year local recurrence rate for patients in CAStotal High and Low groups given demographic conditions with 95% confidence intervals.

REFERENCES

[1] Bang, H. and Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika*, 87(2):329–343.

[2] Bang, H. and Tsiatis, A. A. (2002). Median regression with censored cost data. *Biometrics*, 58(3):643–649.

[3] Bang, H. and Zhao, H. (2012). Average cost-effectiveness ratio with censored data. *Journal of biopharmaceutical statistics*, 22(2):401–415.

[4] Benson, J. R. and Wishart, G. C. (2013). Predictors of recurrence for ductal carcinoma in situ after breast-conserving surgery. *The Lancet Oncology*, 14(9):e348–e357.

[5] Boland, G., Chan, K., Knox, W., Roberts, S., and Bundred, N. (2003). Value of the van nuys prognostic index in prediction of recurrence of ductal carcinoma in situ after breast-conserving surgery. *British journal of surgery*, 90(4):426–432.

[6] Chan, J. Y. (2011). A clinical overview of centrosome amplification in human cancers. *International journal of biological sciences*, 7(8):1122.

[7] Chen, H., Chen, J., and Chen, S.-Y. (2010). Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling. *Canadian Journal of Statistics*, 38(4):582–597.

[8] Chen, J., Chen, S.-Y., and Rao, J. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *Canadian Journal of Statistics*, 31(1):53–68.

[9] Chen, J., Liu, L., Shih, Y.-C. T., Zhang, D., and Severini, T. A. (2016). A flexible model for correlated medical costs, with application to medical expenditure panel survey data. *Statistics in medicine*, 35(6):883–894.

[10] Choi, D. H., Mittal, K., Wei, G., Melton, B. D., Griffith, C., Klimov, S., Reid, M., Golusinki, P., Rida, P., and Aneja, R. (2018). Hypoxia induced centrosome amplification as a surrogate marker in hpv negative oropharyngeal squamous cell carcinomas. In *Laboratory Investigation*, volume 98, pages 474–474.

[11] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[12] D'Assoro, A. B., Barrett, S. L., Folk, C., Negron, V. C., Boeneman, K., Busby, R., Whitehead, C., Stivala, F., Lingle, W. L., and Salisbury, J. L. (2002a). Amplified centrosomes in breast cancer: a potential indicator of tumor aggressiveness. *Breast cancer research and treatment*, 75(1):25–34.

[13] D'Assoro, A. B., Lingle, W. L., and Salisbury, J. L. (2002b). Centrosome amplification and the development of cancer. *Oncogene*, 21(40):6146.

[14] de Almeida, B. P., Vieira, A. F., Paredes, J., Bettencourt-Dias, M., and Barbosa-Morais, N. L. (2019). Pan-cancer association of a centrosome amplification gene expression signature with genomic alterations and clinical outcome. *PLoS Computational Biology*, 15(3).

[15] Denu, R. A., Zasadil, L. M., Kanugh, C., Laffin, J., Weaver, B. A., and Burkard, M. E. (2016). Centrosome amplification induces high grade features and is prognostic of worse outcomes in breast cancer. *BMC cancer*, 16(1):47.

[16] Esserman, L. and Yau, C. (2015). Rethinking the standard for ductal carcinoma in situ treatment. *JAMA oncology*, 1(7):881–883.

[17] Freedman, G. M. and Fowble, B. L. (2000). Local recurrence after mastectomy or breast-conserving surgery and radiation. *Oncology*, 14(11).

[18] Fukasawa, K. (2005). Centrosome amplification, chromosome instability and cancer development. *Cancer letters*, 230(1):6–19.

[19] Godinho, S. and Pellman, D. (2014). Causes and consequences of centrosome abnormalities in cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1650):20130467.

[20] Graybill, F. A. and Wang, C.-M. (1980). Confidence intervals on nonnegative linear combinations of variances. *Journal of the American Statistical Association*, 75(372):869–873.

[21] Guo, H.-q., Gao, M., Ma, J., Xiao, T., Zhao, L.-l., Gao, Y., and Pan, Q.-j. (2007). Analysis of the cellular centrosome in fine-needle aspirations of the breast. *Breast Cancer Research*, 9(4):R48.

[22] Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review/Revue Internationale de Statistique*, pages 109–127.

[23] Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19(2):491.

[24] Hasan, M. S. and Krishnamoorthy, K. (2018). Confidence intervals for the mean and a percentile based on zero-inflated lognormal data. *Journal of Statistical Computation and Simulation*, 88(8):1499–1514.

[25] He, S. and Liang, W. (2014). Empirical likelihood for right censored data with covariables. *Science China Mathematics*, 57(6):1275–1286.

[26] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

[27] Jeyarajah, J. and Qin, G. (2017). Empirical likelihood-based confidence intervals for mean medical cost with censored data. *Statistics in medicine*, 36(25):4061–4070.

[28] Jeyarajah, J., Wei, G., and Qin, G. (2019). Influence function-based empirical likelihood for inference of quantile medical costs with censored data. *Statistical methods in medical research*, page 0962280219880573.

[29] Jing, B.-Y., Yuan, J., and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(487):1224–1232.

[30] Johnson, B. A. (2015). On huberized calibration regression for censored medical cost data. *Statistics in biosciences*, 7(2):367–378.

[31] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[32] Kapur, V., Blough, D. K., Sandblom, R. E., Hert, R., de Maine, J. B., Sullivan, S. D., and Psaty, B. M. (1999). The medical cost of undiagnosed sleep apnea. *Sleep*, 22(6):749–755.

[33] Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

[34] Li, C., Chen, J., and Qin, G. (2019). Partial youden index and its inferences. *Journal of Biopharmaceutical Statistics*, 29(2):385–399.

[35] Li, X., Zhou, X., and Tian, L. (2013). Interval estimation for the mean of lognormal data with excess zeros. *Statistics & Probability Letters*, 83(11):2447–2453.

[36] Lin, D. (2000). Linear regression analysis of censored medical costs. *Biostatistics*, 1(1):35–47.

[37] Lin, D. (2003). Regression analysis of incomplete medical cost data. *Statistics in medicine*, 22(7):1181–1200.

[38] Lopes, C. A., Mesquita, M., Cunha, A. I., Cardoso, J., Carapeta, S., Laranjeira, C., Pinto, A. E., Pereira-Leal, J. B., Dias-Pereira, A., Bettencourt-Dias, M., et al. (2018). Centrosome amplification arises before neoplasia and increases upon p53 loss in tumorigenesis. *The Journal of cell biology*, 217(7):2353–2363.

[39] Marmot, M. G., Altman, D., Cameron, D., Dewar, J., Thompson, S., and Wilcox, M. (2013). The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*, 108(11):2205.

[40] McBride, M., Rida, P. C., and Aneja, R. (2015). Turning the headlights on novel cancer biomarkers: Inspection of mechanics underlying intratumor heterogeneity. *Molecular aspects of medicine*, 45:3–13.

[41] Miligy, I. M., Gorringe, K. L., Toss, M. S., Al-Kawaz, A. A., Simpson, P., Diez-Rodriguez, M., Nolan, C. C., Ellis, I. O., Green, A. R., and Rakha, E. A. (2018). Thioredoxin-interacting protein is an independent risk stratifier for breast ductal carcinoma in situ. *Modern Pathology*, 31(12):1807.

[42] Mittal, K., Choi, D. H., Ogden, A., Donthamsetty, S., Melton, B. D., Gupta, M. V., Pannu, V., Cantuaria, G., Varambally, S., Reid, M. D., et al. (2017). Amplified centrosomes and mitotic index display poor concordance between patient tumors and cultured cancer cells. *Scientific reports*, 7:43984.

[43] Momtahen, S., Curtin, J., and Mittal, K. (2016). Current chemotherapy and potential new targets in uterine leiomyosarcoma. *Journal of clinical medicine research*, 8(3):181.

[44] Nigg, E. A. (2007). Centrosome duplication: of rules and licenses. *Trends in cell biology*, 17(5):215–221.

[45] Ogden, A., Rida, P. C., and Aneja, R. (2017). Prognostic value of ca20, a score based on centrosome amplification-associated genes, in breast tumors. *Scientific reports*, 7(1):262.

[46] Owen, A. et al. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

[47] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

[48] Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

[49] Page, D. L., Dupont, W. D., Rogers, L. W., and Landenberger, M. (1982). Intraductal carcinoma of the breast: follow-up after biopsy only. *Cancer*, 49(4):751–758.

[50] Pannu, V., Mittal, K., Cantuaria, G., Reid, M. D., Li, X., Donthamsetty, S., McBride, M., Klimov, S., Osan, R., Gupta, M. V., et al. (2015). Rampant centrosome amplification underlies more aggressive disease course of triple negative breast cancers. *Oncotarget*, 6(12):10487.

[51] Qin, J. (2017). *Biased Sampling, Over-identified Parameter Problems and Beyond*. Springer.

[52] Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer.

[53] Sherwood, B., Wang, L., and Zhou, X.-H. (2013). Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in medicine*, 32(28):4967–4979.

[54] Shi, X. (1984). The approximate independence of jackknife pseudo-values and the bootstrap methods. *Journal of Wuhan Institute Hydra-Electric Engineering*, 2:83–90.

[55] Silverstein, M. J., Lagios, M. D., Recht, A., Allred, D. C., Harms, S. E., Holland, R., Holmes, D. R., Hughes, L. L., Jackman, R. J., Julian, T. B., et al. (2005). Image-detected breast cancer: state of the art diagnosis and treatment. *Journal of the American College of Surgeons*, 201(4):586–597.

[56] Solin, L. J., Gray, R., Baehner, F. L., Butler, S. M., Hughes, L. L., Yoshizawa, C., Cherbavaz, D. B., Shak, S., Page, D. L., Sledge Jr, G. W., et al. (2013). A multigene expression assay to predict local recurrence risk for ductal carcinoma in situ of the breast. *Journal of the National Cancer Institute*, 105(10):701–710.

[57] Stoltzfus, J. C., Nishijima, D., and Melnikow, J. (2012). Why quantile regression makes good sense for analyzing economic outcomes in medical research. *Academic Emergency Medicine*, 19(7):850–851.

[58] Tian, L. (2005). Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Statistics in medicine*, 24(20):3223–3232.

[59] Topol, E. (2015). *The patient will see you now: the future of medicine is in your hands*. Basic Books.

[60] Toss, M. S., Miligy, I., Al-Kawaz, A., Alsleem, M., Khout, H., Rida, P. C., Aneja, R., Green, A. R., Ellis, I. O., and Rakha, E. A. (2018a). Prognostic significance of tumor-infiltrating lymphocytes in ductal carcinoma in situ of the breast. *Modern Pathology*, 31(8):1226.

[61] Toss, M. S., Miligy, I. M., Gorringe, K. L., AlKawaz, A., Khout, H., Ellis, I. O., Green, A. R., and Rakha, E. A. (2018b). Prolyl-4-hydroxylase a subunit 2 (p4ha2) expression is a predictor of poor outcome in breast ductal carcinoma in situ (dcis). *British journal of cancer*, 119(12):1518.

[62] Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614.

[63] Weerahandi, S. (1995). Generalized confidence intervals. In *Exact statistical methods for data analysis*, pages 143–168. Springer.

[64] Willan, A. R., Lin, D., and Manca, A. (2005). Regression methods for cost-effectiveness analysis with censored data. *Statistics in medicine*, 24(1):131–145.

[65] Young, T. A. (2005). Estimating mean total costs in the presence of censoring. *Pharmacoeconomics*, 23(12):1229–1242.

[66] Zhao, H. and Tian, L. (2001). On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics*, 57(4):1002–1008.

[67] Zhao, H. and Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*, 84(2):339–348.

[68] Zhao, H., Zuo, C., Chen, S., and Bang, H. (2012). Nonparametric inference for median costs with censored data. *Biometrics*, 68(3):717–725.

[69] Zhou, M. (1992). Asymptotic normality of thesynthetic data'regression estimator for censored survival data. *The Annals of Statistics*, pages 1002–1021.

[70] Zhou, X.-H., Qin, G., Lin, H., and Li, G. (2006). Inferences in censored cost regression models with empirical likelihood. *Statistica Sinica*, 16:1213–1232.

[71] Zhou, X.-H. and Tu, W. (2000). Confidence intervals for the mean of diagnostic test charge data containing zeros. *Biometrics*, 56(4):1118–1125.

[72] Zou, G. and Donner, A. (2008). Construction of confidence limits about effect measures: a general approach. *Statistics in medicine*, 27(10):1693–1702.

[73] Zou, G., Huo, C. Y., and Taleban, J. (2009a). Simple confidence intervals for lognormal means and their differences with environmental applications. *Environmetrics: The official journal of the International Environmetrics Society*, 20(2):172–180.

[74] Zou, G. Y., Taleban, J., and Huo, C. Y. (2009b). Confidence interval estimation for lognormal data with application to health economics. *Computational Statistics & Data Analysis*, 53(11):3755–3764.