

The Texas Medical Center Library

DigitalCommons@TMC

The University of Texas MD Anderson Cancer
Center UTHealth Graduate School of
Biomedical Sciences Dissertations and Theses
(Open Access)

The University of Texas MD Anderson Cancer
Center UTHealth Graduate School of
Biomedical Sciences

6-2022

Statistical Modeling of Longitudinal Medical Cost Data

Shikun Wang

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Applied Statistics Commons](#), [Longitudinal Data Analysis and Time Series Commons](#), [Statistical Methodology Commons](#), and the [Survival Analysis Commons](#)

Recommended Citation

Wang, Shikun, "Statistical Modeling of Longitudinal Medical Cost Data" (2022). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 1192.

https://digitalcommons.library.tmc.edu/utgsbs_dissertations/1192

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.



STATISTICAL MODELING OF LONGITUDINAL MEDICAL COST DATA

by

Shikun Wang, M.S.

APPROVED:

Liang Li

Liang Li, Ph.D.
Advisory Professor

Yu Shen

Yu Shen, Ph.D.
Co-advisory Professor

Ya-Chen Tina Shih

Ya-Chen Tina Shih, Ph.D.

Yisheng Li

Yisheng Li, Ph.D.

Jing Ning

Jing Ning, Ph.D.

Benjamin Smith

Benjamin Smith, M.D.

APPROVED:

Dean, The University of Texas
MD Anderson Cancer Center UTHHealth Graduate School of Biomedical Sciences

STATISTICAL MODELING OF LONGITUDINAL MEDICAL COST DATA

A

Dissertation

Presented to the Faculty of

the University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

by

Shikun Wang, M.S.

Houston, Texas

May, 2022

ACKNOWLEDGEMENT

Foremost, my deepest appreciation is to my advisors Dr. Liang Li and Dr. Yu Shen. I have been amazingly fortunate to have advisors who guide me hand in hand towards a confident and reliable researcher. They always support me when facing challenges in both career and life. My journey in statistical methodology and collaborative research in clinical sciences would not have been possible without their help. I would like to convey my sincere thanks to my advisory committee: Dr. Ya-chen Tina Shih, Dr. Jing Ning, Dr. Yisheng Li, and Dr. Benjamin Smith. They have provided domain knowledge, insightful questions, and helpful feedback for my projects. Their perpetual energy and enthusiasm in research keep inspiring me at different stages of my research.

I am also truly grateful to all faculties, staffs and peer students in the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences for their kind support and help. Particularly, a great gratitude to Dr. Daohai Yu, Dr. Yisheng Li and Dr. Paul Scheet, for bringing me to the program. I want to convey my deep thanks to Dr. Bhramar Mukherjee and Dr. Min Zhang of the University of Michigan for their incredible encouragement. I am also very grateful to Dr. Yan Wang and Dr. Yared Grumu, who had greatly helped me with professional development during my fellowship at the U.S. Food and Drug Administration.

Most importantly, none of this would have been possible without my parents. They have always encouraged me to explore my potential and pursue my dreams. Thanks for my friend Lihao Yin and my son Mimi Klaus for a constant source of love and support.

ABSTRACT

STATISTICAL MODELING OF LONGITUDINAL MEDICAL COST DATA

Shikun Wang, M.S.

Advisory Professors: Liang Li, Ph.D. and Yu Shen, Ph.D.

Projecting the future cancer care cost is critical in health economics research and policy making. An indispensable step is to estimate cost trajectories from an incident cohort of cancer patients using longitudinal medical cost data, accounting for terminal events such as death, and right censoring due to loss of follow-up. Since the cost of cancer care and survival are correlated, a scientifically meaningful quantity for inference in this context is the mean cost trajectory conditional on survival. Many standard approaches for longitudinal and survival analysis are not valid for the problem.

The research for my Ph.D. dissertation consists of three aims. In Aim 1, we developed a two-stage semiparametric likelihood-based method to estimate the conditional distribution of longitudinal medical cost trajectory given the time of terminal event. The cost data is assumed normal, which does not reflect the reality. So, for Aim 2, we developed a flexible model to address further challenges such as heteroscedasticity without imposing a cost data distributional assumption. In Aim 3, to conduct flexible and reliable inference on the estimated cost trajectory, we developed a longitudinal varying-coefficient single-index model, and computational optimization algorithm that is scalable to baseline feature inference with noise. For each of the aim, we provide theoretical and simulation-based justification for the proposed approach, and apply the methods to estimate cancer patient cost trajectories from the Surveillance, Epidemiology, and End Results (SEER)-Medicare linked database.

TABLE OF CONTENTS

Approval page	i
Title page	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF TABLES	vii
LIST OF ILLUSTRATIONS	viii
CHAPTER 1 : INTRODUCTION	1
CHAPTER 2 : STATISTICAL MODELING OF LONGITUDINAL MEDICAL COST TRAJECTORY: RENAL CELL CANCER CARE COST ANALYSES .	5
2.1 Introduction	5
2.2 Notation and Model	7
2.3 Estimation	12
2.4 Analyses of Renal Cell Cancer Cost Data from SEER-Medicare	16
2.5 Simulations	21
2.6 Discussion	28

CHAPTER 3 :	AN EXTENSION OF ESTIMATING EQUATIONS TO MODEL LONGITUDINAL MEDICAL COST TRAJECTORY WITH MEDICARE CLAIMS DATA LINKED TO SEER CANCER REGISTRY	32
3.1	Introduction	32
3.2	Notation and Model	36
3.3	Estimation	38
3.4	Analyses of Prostate Cancer Cost Data from SEER-Medicare	46
3.5	Simulations	51
3.6	Discussion	56
CHAPTER 4 :	LONGITUDINAL VARYING COEFFICIENT SINGLE-INDEX MODEL WITH CENSORED COVARIATES	60
4.1	Introduction	60
4.2	Model	64
4.3	Estimation	67
4.4	Analyses of Prostate Cancer Cost Data from SEER-Medicare	78
4.5	Simulations	85
4.6	Discussion	90
CHAPTER 5 :	SUMMARY	96
APPENDICES	98
BIBLIOGRAPHY	136
VITA	141

LIST OF TABLES

TABLE 2.1	Simulation results to compare the descriptive, one-stage (1-stg) and two-stage (2-stg) methods under different sample size and censoring rate.	28
TABLE 2.2	Simulation results for the two-stage (2-stg) estimates for variance coefficients under different sample size and censoring rate. . . .	28
TABLE 3.1	Simulation results to compare GEE1-N, GEE1-Y, and GEE2-Y when the cost data follow Gamma distributions.	54
TABLE 3.2	Simulation results to compare GEE1-N, GEE1-Y, and GEE2-Y when the cost data follow zero-inflated Gamma distribution. . .	55
TABLE 4.1	Estimates of linear coefficients for medical cost trajectory from SEER-Medicare prostate cancer data.	94
TABLE 4.2	Bias, MSE and CP of linear coefficients in the simulation studies. Data generated from normal distribution and zero-inflated Gamma distribution	95
TABLE A.1	Compare the descriptive, one-stage (1-stg) and two-stage (2-stg) estimates under model misspecification.	106
TABLE B.1	Simulation results under different censoring proportions for the GEE1-N, GEE1-Y, and GEE2-Y estimates.	123
TABLE B.2	Simulation results under different zero proportions for the GEE1-N, GEE1-Y, and GEE2-Y estimates.	124
TABLE B.3	Simulation results under different levels of skewness for the GEE1-N, GEE1-Y, and GEE2-Y estimates.	125
TABLE C.1	Bias, MSE and CP of linear coefficients in the simulation studies.	134

LIST OF ILLUSTRATIONS

FIGURE 2.1	Renal cell cancer care cost analyses results for the log transformed cost trajectory estimates.	22
FIGURE 2.2	Renal cell cancer care cost analyses results for the estimated quantiles of quarterly costs at selected death times using two-stage estimates.	23
FIGURE 2.3	Renal cell cancer care cost analyses results for the estimated bivariate grid surface.	24
FIGURE 2.4	Simulation results for pointwise mean squared error (MSE) of the two-stage (2-stg) estimates.	29
FIGURE 3.1	Prostate cancer care cost descriptive analysis of the monthly medical costs.	34
FIGURE 3.2	Prostate cancer care cost analyses results for the monthly cost trajectory estimates.	48
FIGURE 3.3	Prostate cancer care cost analyses results for the monthly cost trajectory 2D heatmaps and 3D surfaces.	49
FIGURE 3.4	Simulation results to compare EM, GEE1-N, GEE1-Y, and GEE2-Y estimates.	57
FIGURE 4.1	Plots of the estimated mean curves of quarterly medical cost by quadratic splines with five equally spaced knots for the SEER-Medicare prostate cancer cohort by three racial groups	62

FIGURE 4.2	SEER-Medicare data application results of the cost Nomogram. Instructions for policy maker are as follows: first, locate the patient's baseline covaraites on the top axes. Draw a vertical line straight upwards to the points axis to determine how many points are contributed; second, sum the points achieved for each predictor and locate this sum on the single index axis below. Third, draw a vertical line straight down to find the patient's total cost (\$1000) assuming the time to death is 16, 24, 32 or over 40 quarters.	82
FIGURE 4.3	SEER-Medicare data application results of the cost trajectories of different index when survival time is 16, 24, 32 or over 40 quarters. Single index equals 0 for reference group. For patients who died within 40 quarters after cancer diagnosis (or for LTS), single index is -0.336 (0.013) for "old-old"; 0.266 (-0.055) for non-Hispanic Black; -0.07 (-0.042) for "old-old" non-Hispanic Black; 0.895 (0.13) for non-Hispanic Black who received radiotherapy as initial treatment; 1.527 (1.014) for non-Hispanic Black who have over 1 comorbidity score and received radiotherapy as initial treatment. Unmentioned conditions are the same as the reference group.	86
FIGURE 4.4	SEER-Medicare data application results for the estimated reference cost trajectories (a) when the time to death (quarter) equals $s = 16, 24, 32$, and over 40 (LTS) with 95% confidence intervals in shaded areas; (b) 2D heatmaps and (c) 3D "surface" with 95% confidence intervals with 95% upper and lower bounds in grey meshes.	87

FIGURE 4.5	Simulation results of estimated trajectories for Normal data (top) and zero-inflated Gamma data (bottom).	91
FIGURE 4.6	Simulation results of estimated power curve for Normal data (top) and zero-inflated Gamma data (bottom).	92
FIGURE A.1	Averaged log transformed quarterly cost stratified by year of survival of renal cell cancer patients.	108
FIGURE A.2	Kaplan-Meier estimates and histograms of log transformed cost of renal cell cancer patients.	109
FIGURE B.1	Prostate cancer care cost sensitivity analyses via one-stage (1-stg) estimate.	121
FIGURE C.1	Descriptive analysis of the quarterly medical costs for the SEER-Medicare prostate cancer cohort.	131
FIGURE C.2	Plots of the estimated mean curves of quarterly medical cost by quadratic splines with five equally spaced knots for the SEER-Medicare prostate cancer cohort by baseline covariates.	132
FIGURE C.3	Simulation results of estimated trajectories for Normal (top) or Gamma (bottom) data.	135

CHAPTER 1

INTRODUCTION

The American Cancer Society (ACS) shows that Overall Cancer Costs are Rising. Cancer represents a significant portion of total U.S. health care spending. Approximately \$183 billion was spent on cancer related health care in 2015, and this amount is projected to grow to \$246 billion by 2030, an increase of 34% (Mariotto et al., 2020). These high costs are paid by people with cancer and their families, employers, insurance companies and taxpayer-funded public programs like Medicare and Medicaid.

Estimating the cost of cancer care is of considerable interest to health policy makers in the United States and other countries. Accurate cost estimation not only helps policy makers better understand the financial burden of cancer to the society, but also contributes to the planning and allocation of health care budgets. Due to advanced data collection and management techniques, medical cost data are now routinely recorded by hospitals and insurance companies; this offers valuable data sources for health services research to enhance the estimation accuracy. Despite the policy significance of accurate cost estimation and increasingly accessible medical cost data, there has been limited progress in statistical methods for the estimation of longitudinal medical cost trajectories.

The conventional phase-of-care (POC) approach for cost estimation developed by the National Cancer Institute (NCI) defines three cancer care phases (Mariotto et al., 2011): an initial care phase which contains the first 12 months after cancer diagnosis; a terminal care phase which includes the last 12 months before deaths; and a continuing care phase which covers the period in between the initial and terminal care phases. The POC approach is oversimplified, and cannot describe the highly volatile patterns of

medical cost data.

Statistical modeling of medical cost trajectories is challenging, because it is nonlinear, involves a substantial proportion of patients with incomplete follow-up, and exhibits patient-level heterogeneity. Descriptive analyses in Brown et al. (2002) have shown that the average monthly medical cost for breast cancer, summarized based on the year of death, is “U-shaped”, since costs during the initial and terminal care phases tend to be higher than those in the continuing care phase. Patients in the initial care phase often receive active treatment, such as surgery, radiotherapy, chemotherapy, immunotherapy, and targeted therapy, and patients in the terminal care phase may receive salvage and/or palliative end-of-life care. During the continuing care phase, patients’ clinical visits become less frequent as most visits are for follow-up surveillance. Thus, the costs of initial and terminal care phases tend to be higher than those in the continuing care phase. An example of this U-shaped pattern can be found in a recent paper that applied the POC approach to report the costs of renal cell cancer using the SEER-Medicare data (Shih et al., 2019). The above patterns of medical cost data observed with real-world data analysis suggest that longitudinal cost trajectories should be modeled nonlinearly with a flexible functional form and the shape of these trajectories should be allowed to vary based on the survival time.

Besides nonlinearity, incomplete follow-up time is also common in medical cost data. The vast majority of medical cost data comes from claims databases, and the time span of the follow-up data depends on health insurance enrollment as well as the schedule to release data for research use. As a result of administrative censoring, a substantial proportion of patients may have incomplete longitudinal cost data and unknown death time. The approach of only analyzing data from uncensored subjects (Brown et al., 2002), while easy to implement, may produce biased cost estimates depending on the estimand stud-

ied. Even when selected estimands can be estimated unbiasedly under the independent censoring assumption, this approach will lead to inefficient cost trajectory estimates, particularly among subjects with longer survival times (Li et al., 2018; Kong et al., 2018). Therefore, proper adjustment for incomplete follow-up (i.e., right censoring) is necessary. Another related issue is that there are less subjects under follow-up with longer survival times; this sparsity at the tail of the survival distribution may lead to high variation not only in the estimated survival distribution, but also in cost estimates. The variation can be reduced by incorporating cost data from censored subjects.

Furthermore, proper modeling of the heterogeneity in medical cost data helps to improve the interpretability of cost estimations. For example, cancer patients with longer survival often incur lower incident costs (e.g., monthly cost or quarterly costs) in a given year as their overall health tends to be better, while patients with shorter survival may incur higher incident costs due to worse overall health and/or high disease severity. Not surprisingly, longer survival usually implies higher cumulative costs, which is not necessary true for incident costs. Therefore, an approach that views patients as a homogeneous group, and calculates the cost averages among those who are under follow-up without adjustment for survival, will lead to results with less interpretability and precision. A straightforward way to handle heterogeneity is to estimate cost trajectories by classifying patients into subgroups with respect to their survival time, and to model the within- as well as between-patient variations simultaneously. Several joint models for the analysis of longitudinal medical costs and survival are based on indirectly modeling the association between the shape of trajectory and the survival time through shared random effects (Liu et al., 2007, 2008; Liu, 2009). Fitting such models often involves intensive computation with possibly multi-dimensional integral with respect to the random effects, and for this reason, these models used very simple trajectory shapes such as linear shapes with a

small number of random effects. Therefore, these methods are not directly applicable to our application to estimate the mean cost trajectories. Chan and Wang (2010) studied a nonparametric method to estimate the mean cost trajectories counting from the terminal time backwards, but their estimation is not conditional on survival time.

Some recent methodology and software for the shared parameter model can be used to estimate non-linear cost trajectories by incorporating a large number of splines basis into its longitudinal sub-model (Rizopoulos, 2012). However, the increasing the number of spline coefficients results in increasing numerical dimension of integration with respect to the shared random effects, which makes the computation intensive, especially when the data come from large insurance claims databases such as SEER-Medicare. Since the shared parameter model is not designed for estimating nonlinear cost trajectory conditional on survival, it may involve modeling assumptions that are not essential but at-risk of model violation, such as the proportional hazard assumption in its survival sub-model. Furthermore, the typical shared parameter model does not directly model and estimate the cost trajectory, unlike the proposed model in this dissertation.

The remainder of the dissertation is organized as follows. In Chapter 2, we will concentrate on a two-stage semiparametric likelihood-based method to estimate the conditional distribution of longitudinal medical cost trajectory given the time of terminal event (Wang et al., 2020). In Chapter 3, we will a flexible marginal model to address further challenges such as heteroscedasticity without assuming a cost data distribution. In Chapter 4, we will further propose a longitudinal varying-coefficient single-index model and computational optimization algorithm for baseline feature inference. Chapter 5 is the Summary of this dissertation, which summarizes the innovative methods we proposed in this dissertation.

CHAPTER 2

STATISTICAL MODELING OF LONGITUDINAL MEDICAL COST TRAJECTORY: RENAL CELL CANCER CARE COST ANALYSES*

2.1. Introduction

Estimating the cost of cancer care is of considerable interest to health policy makers in the United States and other countries. Accurate cost estimation not only helps policy makers better understand the financial burden of cancer to the society, but also contributes to the planning and allocation of health care budgets. Due to advanced data collection and management techniques, medical cost data are now routinely recorded by hospitals and insurance companies; this offers valuable data sources for health services research to enhance the estimation accuracy. Despite the policy significance of accurate cost estimation and increasingly accessible medical cost data, there has been limited progress in statistical methods for the estimation of longitudinal medical cost trajectories.

Our motivating dataset is from the Surveillance, Epidemiology and End Results (SEER)-Medicare database. It consists of survival information and monthly measured medical costs from 15,282 renal cell cancer patients who were first diagnosed between January 1, 2003 and December 31, 2011. In preliminary analyses presented in Figure A.1, we found that the average quarterly medical costs by survival years exhibit a nonlinear trend with many local fluctuations. In addition, a substantial percentage of patients are right-censored (i.e. 55%), and the maximum censoring time is eight years, leading to

*This chapter is based upon “Wang, S., Shen, Y., Shih, Y. C. T., Xu, Y. and Li, L., 2020. *Statistical modeling of longitudinal medical cost trajectory: renal cell cancer care cost analyses. Biostatistics.*”, available online at: <https://doi.org/10.1093/biostatistics/kxab024>. This article has been accepted for publication in Biostatistics Published by Oxford University Press.

incomplete data in both cost and mortality.

Li et al. (2018) proposed to estimate mean cost trajectory conditional on survival time, using an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) in a joint model of longitudinal costs and survival. This method adjusts for right censoring when estimating the longitudinal cost trajectory. By extending the methodology in Li et al. (2018), we propose a two-stage semiparametric approach to estimate the cost trajectory from a joint model of longitudinal medical costs and survival.

Our method address the data challenges mainly in four aspects. First, we observed that many cancer cost datasets have a substantial proportion of patients whose death times are beyond the maximum follow-up time. Hence, the survival distribution is not identifiable on its entire support. Our proposed approach incorporates these “long-term survivors” (LTS) in the joint model. Second, instead of using a flexible nonlinear parametric function to characterize the longitudinal mean cost trajectory by survival, we extend the model with more flexibility to allow the mean cost to have a distinct parameter at any cost period for any survival group. This flexibility is attractive from a practical perspective, because it minimizes the possible misspecification of the mean function of the cost trajectory when we want to apply the cost estimation method to a variety of cancer populations and datasets. Third, the price of gaining more flexibility and adding LTS to the model is a considerable increase in computational complexity and time, due to a large number of parameters in the model. To solve this problem, we propose a two-stage estimator, which is computationally simpler than the EM algorithm in Li et al. (2018) yet still retains adequate statistical efficiency. This computational advantage makes the proposed method much easier to use by practitioners. Fourth, we establish an asymptotic theory for the proposed two-stage estimator.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the notation and the model. In Section 2.3, we propose a two-stage estimation method, provide some theoretical justification and discuss practical adjustments of the fitting algorithm. We describe an application of our methodology to renal cell cancer cost data from SEER-Medicare in Section 2.4, and then compare the empirical performance of our proposed method with a descriptive method from Brown et al. (2002) and a one-stage method in a simulation study in Section 2.5. A discussion is provided in Section 2.6. The Appendix A details the proofs and additional analyses.

2.2. Notation and Model

Suppose we have independent and identically distributed data from n patients, denoted by $\{\mathbf{Y}_i, T_i = \min(\tilde{T}_i, C_i), \delta_i = 1\{\tilde{T}_i \leq C_i\}; i = 1, \dots, n\}$ where \tilde{T}_i denotes the time to death, C_i denotes the time to censoring, T_i and δ_i denote the observed follow-up time and censoring indicator respectively. We use $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ to denote subject i 's vector of incident medical costs, which are defined to be the cost corresponding to pre-specified fixed-length time intervals (e.g. monthly, quarterly or annually) up to T_i . These intervals (called measurement time of the incident cost) are indexed by $\mathbf{t}_i = (t_{i1} = 1, t_{i2} = 2, \dots, t_{in_i} = \lfloor T_i \rfloor)^T$, where $\lfloor x \rfloor$ denotes the largest integer that is smaller than x . $n_i \equiv \lfloor T_i \rfloor$ denotes the number of cost measurements. For notational convenience, we also express the incident cost as a function of the measurement time, i.e., $Y_{ij} = Y_i(t_{ij}), j = 1, \dots, n_i$. We assume that C_i and $(\tilde{T}_i, \mathbf{Y}_i)$ are independent. We use τ to denote the maximum follow-up time, which can be empirically defined as the maximum of uncensored survival times in the dataset. The index time $0 < t_{ij} \leq \tau$ as our primary interest is the cost distributions within the upper bound τ . In studies with limited follow-up duration, a substantial proportion of subjects are censored at the end of the study and hence the survival distribution is not identifiable beyond τ . In what follows, we will show the need

to consider separate modeling for the cost of patients who survive beyond τ .

2.2.1. Two-part Model for Longitudinal Cost Trajectory

Our goal is to estimate the mean incident cost trajectory as a function of time, conditional on a patient's time to death, i.e., $\mu(t, \tilde{T}_i) = E\{\mathbf{Y}_i(t)|\tilde{T}_i\}$ for $t = 1, 2, \dots, \lfloor \tilde{T}_i \rfloor$. We want to use cost data from both the censored and uncensored subjects. For that purpose, a joint distribution of longitudinal costs and survival will facilitate the estimation to utilize the cost data from censored subjects. In the following, we will show a computationally tractable method of decomposing the joint distribution into the marginal distribution of survival and the conditional distribution of cost given survival, $f(\mathbf{Y}_i|\tilde{T}_i; \boldsymbol{\theta})$. Any appropriate parametric model can be used for this conditional distribution, and our proposed estimation procedure and theoretical justification can apply. Since preliminary analyses of our motivating dataset on renal cell cancer cost showed that, after log-transformation, the quarterly incident costs are approximately normally distributed, we will present our methodology under a multivariate normal cost model given survival. Other properly specified parametric models can be used if justified by model diagnosis. Details on model diagnosis are presented in Appendix A.

Our proposed model is a two-part mixture model in the sense that we specify a model for $f\{Y(t)|\tilde{T}\}$ for those with $\tilde{T} \leq \tau$ and another for those with $\tilde{T} > \tau$. Since the distribution of \tilde{T} is unidentifiable beyond τ , this motivates us to model $f\{Y(t)|\tilde{T} > \tau\}$ instead of $f\{Y(t)|\tilde{T}, \tilde{T} > \tau\}$, which is justified in our motivating dataset. We observe a notable proportion of patients who survived beyond 8 years, and their mean quarterly cost was \$4,161; for patients who died within 8 years, their mean quarterly cost was \$15,014. This evidence suggests that the cost distribution for LTS could be very different

from the rest of the population. We propose the following two-part model:

$$\begin{aligned} \mathbf{Y}_i|\tilde{T}_i; \boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\phi}) &\sim MVN\left(\mathbf{D}_i(\tilde{T}_i)\boldsymbol{\beta}, \boldsymbol{\Sigma}_i(\boldsymbol{\phi})\right), \tilde{T}_i \leq \tau, \\ \mathbf{Y}_i|\boldsymbol{\theta}_2 = (\boldsymbol{\xi}, \boldsymbol{\psi}) &\sim MVN\left(\mathbf{G}_i\boldsymbol{\xi}, \boldsymbol{\Sigma}_i(\boldsymbol{\psi})\right), \tilde{T}_i > \tau, \end{aligned} \quad (2.1)$$

where $\mathbf{D}_i(\tilde{T}_i)$ is a matrix of basis functions that depends on both the measurement time \mathbf{t}_i and the (possibly censored) time to death \tilde{T}_i ; $\boldsymbol{\beta}$ is a vector of unknown parameters representing the mean cost at each quarter conditional on survival; $\boldsymbol{\Sigma}_i(\boldsymbol{\phi})$ is the $n_i \times n_i$ covariance matrix of \mathbf{Y}_i when the survival time is less than τ ; \mathbf{G}_i is a matrix of basis functions that is dependent on measurement time \mathbf{t}_i but not the survival time \tilde{T}_i ; $\boldsymbol{\xi}$ is a vector of unknown parameters representing the mean cost at each quarter among all the subjects in the LTS group; and $\boldsymbol{\Sigma}_i(\boldsymbol{\psi})$ is the $n_i \times n_i$ covariance matrix of \mathbf{Y}_i for the LTS group. We use $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ to denote the parameters in the two models in (2.1), and $\boldsymbol{\theta}$ denotes their union.

2.2.2. Adjustment for Right Censoring

We construct a pseudo likelihood (Besag, 1975) function to incorporate the cost data from censored subjects into the estimation. Let $\eta(\cdot)$ be the survival function of \tilde{T} , \bar{F}_C be the survival function of the time to censoring C , $f_{\tilde{T}}$ be the marginal density of \tilde{T} , f_C be the marginal density of C , and $l(\boldsymbol{\theta}; \mathbf{Y}_i, \tilde{T}_i) = \log f(\mathbf{Y}_i|\tilde{T}_i; \boldsymbol{\theta})f_{\tilde{T}}(\tilde{T}_i)$ be the log-likelihood of subject i . Then from equation (2.1), we have $f(\mathbf{Y}_i|\tilde{T}_i; \boldsymbol{\theta}) = f(\mathbf{Y}_i|\tilde{T}_i; \boldsymbol{\theta}_1)^{1_{\{\tilde{T}_i \leq \tau\}}} f(\mathbf{Y}_i|\boldsymbol{\theta}_2)^{1_{\{\tilde{T}_i > \tau\}}}$.

In an ideal situation where the time to death is known or censored at τ for every subject, the likelihood function for the complete data $(\mathbf{Y}_i, \tilde{T}_i)$ of subject i can be factored into $f(\mathbf{Y}_i, \tilde{T}_i; \boldsymbol{\theta}) = \{f(\mathbf{Y}_i|\tilde{T}_i; \boldsymbol{\theta}_1)f_{\tilde{T}}(\tilde{T}_i)\}^{1_{\{\tilde{T}_i \leq \tau\}}} \{f(\mathbf{Y}_i|\boldsymbol{\theta}_2)\eta(\tau)\}^{1_{\{\tilde{T}_i > \tau\}}}$. Then the log full

likelihood function is given by

$$L_n^0(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n l(\boldsymbol{\theta}; \mathbf{Y}_i, \tilde{T}_i).$$

In the presence of right censoring, given the assumption that censoring times are independent of survival time and cost, the likelihood for subject i is

$$\begin{aligned} f(\mathbf{Y}_i, T_i, \delta_i; \boldsymbol{\theta}) &= \left\{ f(\mathbf{Y}_i | T_i; \boldsymbol{\theta}_1) f_{\tilde{T}}(T_i) \bar{F}_C(T_i) \right\}^{\delta_i 1\{T_i \leq \tau\}} \left\{ f(\mathbf{Y}_i | \boldsymbol{\theta}_2) \eta(\tau) f_C(\tau) \right\}^{(1-\delta_i) 1\{T_i = \tau\}} \\ &\times \left\{ \left\{ \int_{T_i}^{\tau} f(\mathbf{Y}_i | u; \boldsymbol{\theta}_1) f_{\tilde{T}}(u) du + f(\mathbf{Y}_i | \boldsymbol{\theta}_2) \eta(\tau) \right\} f_C(T_i) \right\}^{(1-\delta_i) 1\{T_i < \tau\}}. \end{aligned} \quad (2.2)$$

The first term in (2.2) represents the likelihood contribution from uncensored subjects and LTS who are censored at τ . The second term in (2.2) stands for the likelihood contribution from subjects being right-censored prior τ , who may or may not survive beyond τ . The functions associated with censoring times C will not contribute information to estimating $\boldsymbol{\theta}$, thus can be dropped from the likelihood.

We first propose a one-stage estimator $\hat{\boldsymbol{\theta}}$, which is estimated by maximizing the log-likelihood function for uncensored data plus the censored data from LTS, who were administratively censored at the last follow-up time τ . The log-likelihood function is proportional to

$$n^{-1} \sum_{i=1}^n \left\{ \delta_i 1\{T_i \leq \tau\} \log f(\mathbf{Y}_i | T_i; \boldsymbol{\theta}_1) + (1 - \delta_i) 1\{T_i = \tau\} \log f(\mathbf{Y}_i | \boldsymbol{\theta}_2) \right\}. \quad (2.3)$$

The one-stage estimation can be viewed as a regression of longitudinal costs on the uncensored subjects and LTS, with the observed survival as covariates. The one-stage estimator yields a consistent and asymptotically normal estimator under usual

independent censoring assumption for the censoring time. While valid, it is inefficient because it does not use cost data from censored subjects. In real data analysis when the data is sufficiently large, the one-stage estimator is applicable as a simple and reasonable solution to estimating the longitudinal medical cost trajectory. To properly utilize observations from subjects who are right censored before τ in the likelihood function, we propose a two-stage estimator. It is motivated by the following relationship:

$$E \left\{ l(\boldsymbol{\theta}; \mathbf{Y}_i, \tilde{T}_i) | \mathbf{Y}_i, T_i, \eta \right\} = \frac{\int_{T_i}^{\infty} l(\boldsymbol{\theta}; \mathbf{Y}_i, u) dF_{\tilde{T}}(u | \mathbf{Y}_i; \boldsymbol{\theta}, \eta)}{1 - F_{\tilde{T}}(T_i | \mathbf{Y}_i; \boldsymbol{\theta}, \eta)}$$

where $F_{\tilde{T}}(u | \mathbf{Y}_i; \boldsymbol{\theta}, \eta) = f(\tilde{T}_i \leq u | \mathbf{Y}_i; \boldsymbol{\theta}, \eta) = 1 - \int_u^{\tau} f(\mathbf{Y}_i | t; \boldsymbol{\theta}_1) d\eta(t) - f(\mathbf{Y}_i | \boldsymbol{\theta}_2) \eta(\tau)$ is the estimated conditional cumulative density function of time to death \tilde{T}_i given observed medical cost data \mathbf{Y}_i . The two-stage estimator $\tilde{\boldsymbol{\theta}}$ is defined to be the maximizer of the following function:

$$\begin{aligned} L_n(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) \propto n^{-1} \sum_{i=1}^n \left\{ \delta_i 1\{T_i \leq \tau\} \log f(\mathbf{Y}_i | T_i; \boldsymbol{\theta}_1) + (1 - \delta_i) 1\{T_i = \tau\} \log f(\mathbf{Y}_i | \boldsymbol{\theta}_2) \right. \\ \left. + (1 - \delta_i) 1\{T_i < \tau\} \frac{\int_{T_i}^{\infty} l(\boldsymbol{\theta}; \mathbf{Y}_i, u) dF_{\tilde{T}}(u | \mathbf{Y}_i; \hat{\boldsymbol{\theta}}, \hat{\eta})}{1 - F_{\tilde{T}}(T_i | \mathbf{Y}_i; \hat{\boldsymbol{\theta}}, \hat{\eta})} \right\}. \end{aligned} \quad (2.4)$$

In this pseudo likelihood function, the conditional expectation term includes the consistent estimators $\hat{\boldsymbol{\theta}}$ from the one-stage estimator, and $\hat{\eta}(\cdot)$, which is the Kaplan-Meier estimator of the survival function of \tilde{T} , estimated up to τ . Define the true value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_0$, which has finite dimension. Under some regularity conditions, we have shown in Appendix A that the estimators from the one-stage and the two-stage procedures, respectively, are consistent and asymptotically normal. Furthermore, the two-stage estimator is more efficient than the one-stage estimator, and the efficiency gain depends on the accuracy of estimation on the right tail of $F_{\tilde{T}}(C | \mathbf{Y}; \hat{\boldsymbol{\theta}}, \hat{\eta})$ beyond the censoring time C .

for censored subjects.

2.3. Estimation

2.3.1. Basis Function for Modeling Non-linear Cost Trajectory

In this section we discuss the choice of basis in equation (2.1) to model the mean trajectories, i.e., $\mathbf{D}_i(\tilde{T}_i)$ and \mathbf{G}_i . First, we let the time to death \tilde{T}_i take values in pre-specified integer time unit (e.g. month, quarter or year), and τ becomes the maximum feasible integer time to death. Therefore, the mean function of cost trajectory $\mu(t, \tilde{T}_i), t = 1, 2, \dots, n_i \leq \tilde{T}_i, \tilde{T}_i = 1, 2, \dots, \tau$ can be reviewed as a bivariate discrete grid surface over a triangular area. The matrix of basis is denoted as $\mathbf{D}_i(\tilde{T}_i)$. The k -th row vector is equal to $D(t_{ik}, \tilde{T}_i) = \left(1\{t_{ik} = 1, \tilde{T}_i = 1\}, 1\{t_{ik} = 1, \tilde{T}_i = 2\}, 1\{t_{ik} = 2, \tilde{T}_i = 2\}, \dots, 1\{t_{ik} = 1, \tilde{T}_i = \tau\}, 1\{t_{ik} = 2, \tilde{T}_i = \tau\}, \dots, 1\{t_{ik} = \tau, \tilde{T}_i = \tau\}\right), k = 1, 2, \dots, n_i$. This is the tensor product of $\left(1\{t_{ik} = 1\}, 1\{t_{ik} = 2\}, \dots, 1\{t_{ik} = \tau\}\right)$ and $\left(1\{\tilde{T}_i = 1\}, 1\{\tilde{T}_i = 2\}, \dots, 1\{\tilde{T}_i = \tau\}\right)$ with the removal of elements at which $t_{ik} > \tilde{T}_i$. Similarly, the k -th row of basis for the mean trajectory of the LTS group \mathbf{G}_i is denoted as $G(t_{ik}) = \left(1\{t_{ik} = 1\}, 1\{t_{ik} = 2\}, \dots, 1\{t_{ik} = \tau\}\right)$. This choice of basis functions achieves maximum flexibility, because it gives a distinct parameter to every t by \tilde{T} combination ($t \leq \tilde{T}$).

2.3.2. Penalized Likelihood

The proposed tensor product basis representation of the bivariate surface $\mu(t, \tilde{T})$ is very flexible but also introduces a large number of parameters. Penalized estimation, as shown below, is helpful in improving numerical stability and reducing variation. It also helps to form a “smooth” surface in the sense that the mean costs are closer at the adjacent (t, \tilde{T}) value. Some datasets may have sparse survival data at the tail of the follow-up time. As a result, there is not enough cost data in that region to estimate

that part of the trajectory properly. Penalized estimation enables borrowing information from other regions of the surface to ensure improved estimation of the cost trajectory at tails. We describe penalized estimation below. This method is applicable to both the one-stage and two-stage estimators.

We enhance our parameter estimation approaches by adding $\frac{\lambda_\beta}{2}(\boldsymbol{\beta}^T \boldsymbol{\Omega}_\beta \boldsymbol{\beta})$ and $\frac{\lambda_\xi}{2}(\boldsymbol{\xi}^T \boldsymbol{\Omega}_\xi \boldsymbol{\xi})$ to the likelihood function in equation (2.3) and equation (2.4), where λ_β and λ_ξ are prespecified penalty parameters. Since our choice of basis allows the mean incident cost at any time for any survival group to have a distinct parameter, it is natural to introduce difference (or higher order) penalties on both the measurement time direction and the survival time direction. To improve model performance on boundary points of the triangular area, for $\tilde{T} \leq \tau$, we add penalty on the difference of cost at both initial quarters and terminate quarters when patients usually actively receive treatments.

Note that strictly speaking, the estimation of the mean cost trajectory function $\mu(t, \tilde{T})$ is not a smoothing problem on a bivariate function because both t and \tilde{T} are discrete and the number of discrete time points are bounded. The smoothness of the trajectory is hence undefined because its second derivative with respect to t and \tilde{T} is undefined. Nonetheless, it is reasonable to expect $\mu(t, \tilde{T})$ to be closer at more adjacent t or \tilde{T} values. For this reason, we extended the idea of penalized spline estimation (Eilers and Marx, 1996) as our main strategy to estimate $\mu(t, \tilde{T})$. For convenience of presentation, we followed the terminology from the penalized spline smoothing literature, but noted that the precise definitions of those terms are different in our discrete estimation problem. As an example, the “smoothness” penalty is essentially a ridge penalty, which is important to control the variation of the estimated bivariate surface, because our model for $\mu(t, \tilde{T})$ employs a large number of parameters, i.e., a distinct parameter at each t by \tilde{T} combination. The ridge penalty produces a visually smooth surface when the estimated

$\mu(t, \tilde{T})$ is plotted as a function of t and \tilde{T} .

The outline of our algorithm for the one-stage estimator $\hat{\boldsymbol{\theta}}$ is as follows. First, the initial values of $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\xi}}^{(0)}$ are estimated by the following penalized least squares estimator

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(0)} &= \left\{ n^{-1} \sum_{i=1}^n \delta_i \mathbf{D}_i(T_i)^T \mathbf{D}_i(T_i) + \lambda_{\beta} \boldsymbol{\Omega}_{\beta} \right\}^{-1} \left\{ n^{-1} \sum_{i=1}^n \delta_i \mathbf{D}_i(T_i)^T \mathbf{Y}_i \right\} \\ \hat{\boldsymbol{\xi}}^{(0)} &= \left\{ n^{-1} \sum_{i=1}^n (1 - \delta_i) 1\{T_i = \tau\} \mathbf{G}_i^T \mathbf{G}_i + \lambda_{\xi} \boldsymbol{\Omega}_{\xi} \right\}^{-1} \left\{ n^{-1} \sum_{i=1}^n (1 - \delta_i) 1\{T_i = \tau\} \mathbf{G}_i^T \mathbf{Y}_i \right\}.\end{aligned}$$

We then update $\boldsymbol{\theta}$ iteratively until a prespecified convergence criterion is reached. At iteration step m , the variance-covariance parameter estimators $\hat{\boldsymbol{\phi}}^{(m)}$ and $\hat{\boldsymbol{\psi}}^{(m)}$ are calculated from residuals given the mean components from the previous iteration, analogously to the generalized least squares estimation (Pinheiro and Bates, 2006). The updated estimators of $\hat{\boldsymbol{\beta}}^{(m)}$ and $\hat{\boldsymbol{\xi}}^{(m)}$ are given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m)} &= \left\{ n^{-1} \sum_{i=1}^n \delta_i \mathbf{D}_i(T_i)^T \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\phi}}^{(m)}) \mathbf{D}_i(T_i) + \lambda_{\beta} \boldsymbol{\Omega}_{\beta} \right\}^{-1} \\ &\quad \times \left\{ n^{-1} \sum_{i=1}^n \delta_i \mathbf{D}_i(T_i)^T \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\phi}}^{(m)}) \mathbf{Y}_i \right\} \\ \hat{\boldsymbol{\xi}}^{(m)} &= \left\{ n^{-1} \sum_{i=1}^n (1 - \delta_i) 1\{T_i = \tau\} \mathbf{G}_i^T \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\psi}}^{(m)}) \mathbf{G}_i + \lambda_{\xi} \boldsymbol{\Omega}_{\xi} \right\}^{-1} \\ &\quad \times \left\{ n^{-1} \sum_{i=1}^n (1 - \delta_i) 1\{T_i = \tau\} \mathbf{G}_i^T \boldsymbol{\Sigma}_i^{-1}(\hat{\boldsymbol{\psi}}^{(m)}) \mathbf{Y}_i \right\}.\end{aligned}$$

To calculate the two-stage estimator $\tilde{\boldsymbol{\theta}}$, we first estimate the residual lifetime distribution of \tilde{T} for each censored subject, conditional on the observed cost data. The parameter in this conditional distribution is estimated by its one-stage estimator in Equation (2.4). For a censored subject, \tilde{T} may take any values on a discrete scale from the censored

time to τ , or $> \tau$. For the i -th subject with a censoring time at T_i , its residual survival probability at the discrete times in $(T_i, \tau]$ and at the interval (τ, ∞) can be characterized by probability weights $\hat{w}_{ij}, j = T_i + 1, \dots, \tau, \tau + 1$ where

$$\hat{w}_{ij} = \begin{cases} \frac{f(\mathbf{Y}_i|\tilde{T}_i = j; \hat{\boldsymbol{\theta}}_1)(\hat{\eta}(j-1) - \hat{\eta}(j))}{\sum_{k=T_i+1}^{\tau} f(\mathbf{Y}_i|\tilde{T}_i = k; \hat{\boldsymbol{\theta}}_1)(\hat{\eta}(k-1) - \hat{\eta}(k)) + f(\mathbf{Y}_i|\hat{\boldsymbol{\theta}}_2)\hat{\eta}(\tau)}, & j = T_i + 1, \dots, \tau \\ \frac{f(\mathbf{Y}_i|\hat{\boldsymbol{\theta}}_2)\hat{\eta}(\tau)}{\sum_{k=T_i+1}^{\tau} f(\mathbf{Y}_i|\tilde{T}_i = k; \hat{\boldsymbol{\theta}}_1)(\hat{\eta}(k-1) - \hat{\eta}(k)) + f(\mathbf{Y}_i|\hat{\boldsymbol{\theta}}_2)\hat{\eta}(\tau)}, & j = \tau + 1 \end{cases}$$

Then we set the initial values to be the one-stage estimator $\hat{\boldsymbol{\theta}}$, and update $\boldsymbol{\theta}$ iteratively through the same algorithm for the one-stage estimate. At each iteration step m' , the updated parameters are

$$\begin{aligned} \tilde{\boldsymbol{\beta}}^{(m')} &= \left\{ n^{-1} \sum_{i=1}^n \left\{ \delta_i \mathbf{D}_i(T_i)^T \boldsymbol{\Sigma}_i^{-1}(\tilde{\boldsymbol{\phi}}^{(m')}) \mathbf{D}_i(T_i) \right. \right. \\ &\quad \left. \left. + (1 - \delta_i) 1\{T_i < \tau\} \sum_{j=1}^{\tau} \hat{w}_{ij} \mathbf{D}_i(j)^T \boldsymbol{\Sigma}_i^{-1}(\tilde{\boldsymbol{\phi}}^{(m')}) \mathbf{D}_i(j) \right\} + \lambda_{\boldsymbol{\beta}} \boldsymbol{\Omega}_{\boldsymbol{\beta}} \right\}^{-1} \\ &\quad \times \left\{ n^{-1} \sum_{i=1}^n \left\{ \delta_i \mathbf{D}_i(T_i)^T + (1 - \delta_i) 1\{\tilde{T}_i < \tau\} \sum_{j=1}^{\tau} \hat{w}_{ij} \mathbf{D}_i(j)^T \right\} \boldsymbol{\Sigma}_i^{-1}(\tilde{\boldsymbol{\phi}}^{(m')}) \mathbf{Y}_i \right\} \\ \tilde{\boldsymbol{\xi}}^{(m')} &= \left\{ n^{-1} \sum_{i=1}^n (1 - \delta_i) \left\{ 1\{T_i = \tau\} + 1\{T_i < \tau\} \hat{w}_{i,\tau+1} \right\} \mathbf{G}_i^T \boldsymbol{\Sigma}_i^{-1}(\tilde{\boldsymbol{\psi}}^{(m')}) \mathbf{G}_i + \lambda_{\boldsymbol{\xi}} \boldsymbol{\Omega}_{\boldsymbol{\xi}} \right\}^{-1} \\ &\quad \times \left\{ n^{-1} \sum_{i=1}^n (1 - \delta_i) \left\{ 1\{T_i = \tau\} + 1\{T_i < \tau\} \hat{w}_{i,\tau+1} \right\} \mathbf{G}_i^T \boldsymbol{\Sigma}_i^{-1}(\tilde{\boldsymbol{\psi}}^{(m')}) \mathbf{Y}_i \right\}. \end{aligned}$$

The penalty parameters $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\xi}}$ are selected by 10-fold cross validation (CV) in the following steps.

1. Split the data randomly into 10 subsets, each subset having an equal number of subjects;
2. For $l = 1, 2, \dots, 10$, use the l -th subset as the test set, and the remaining 9 subsets

as the training set. Fit the model using the training set and calculate the following score in the test set:

$$S_l = \frac{\sum_{j=1}^{J_l} (\delta_j + (1 - \delta_j)1\{T_i = \tau\}) \|\mathbf{Y}_j - \hat{\mathbf{Y}}_j\|^2}{\sum_{j=1}^{J_l} (\delta_j + (1 - \delta_j)1\{T_i = \tau\}) n_j}$$

where $\{\mathbf{Y}_j, \mathbf{t}_j, T_j, \delta_j | j = 1, \dots, J_l\}$ is a collection of data from the test set. $\hat{\mathbf{Y}}_j$ is estimated from the fitted model to the training set;

3. The cross-validation score for $\lambda_\beta, \lambda_\xi$ is $CV(\lambda_\beta, \lambda_\xi) = 10^{-1} \sum_{l=1}^{10} S_l$;

4. Choose λ_β and λ_η to minimize the $CV(\lambda_\beta, \lambda_\xi)$.

Because the asymptotic variance of $\hat{\boldsymbol{\theta}}$ has a very complicated form, we instead used bootstrap to estimate the 95% pointwise confidence interval of the mean cost trajectory function $\mu(t, \tilde{T})$ at different (t, \tilde{T}) 's. In the b -th bootstrap resampling ($b = 1, 2, \dots, B$), we randomly sample n patients with replacement from the original dataset, and estimate the model parameters $\hat{\boldsymbol{\theta}}^{(b)}$ as well as $\hat{\mu}^{(b)}(t, \tilde{T})$ by running the bootstrap dataset through the aforementioned estimation procedure. The bootstrap variance estimator of the point estimator $\hat{\mu}(t, \tilde{T})$ is the sample variance of all the $\hat{\mu}^{(b)}(t, \tilde{T})$'s from a total of $B = 1000$ bootstrap samples. The 95% pointwise confidence interval of $\mu(t, \tilde{T})$ is calculated by Wald method.

2.4. Analyses of Renal Cell Cancer Cost Data from SEER-Medicare

We analyzed the longitudinal medical cost data of a cohort extracted from the SEER-Medicare database. The study cohort consisted of 15,282 patients with a renal cell cancer diagnosis between January 1, 2003 and December 31, 2011. These patients were 65 years old or above at the time of diagnosis and had continuous Medicare Parts A and B enrollment throughout the follow-up period. Among them, 55% of the patients were right

censored, and all censoring occurred beyond the first 12 months after cancer diagnosis. There were 12,339 (80.7%) patients in our study cohort who were first diagnosed with local or regional stage cancer, and 2,943 (19.3%) were first diagnosed with distant stage cancer. The five-year survival probability is much higher for local or regional stage cancer patients (67.8%) than distant stage patients (6.25%). The goal of this data application was to estimate the mean medical costs at the population level in order to study the patterns and trends of cost for patients in different renal cell cancer stages.

The cost includes both Medicare Part A and Part B payments. When constructing the analytical sample from the SEER-Medicare data, we have adjusted all the costs to 2017 US dollars by using the medical care component of Consumer Price Index. When aggregated into quarterly costs, the cost data clearly have a skewed distribution, but the distribution becomes close to Gaussian after a $\log(1 + x)$ transformation is applied to reduce the skewness (Figure A.2). This justifies the application of our proposed model on the transformed data. The goal of this analysis was to estimate the cost trajectory of patients based on stage at diagnosis. For comparison purposes, we also report results from a published descriptive method (Brown et al., 2002), which grouped patients by year to death, then calculated the mean quarterly costs among subjects in each group; only uncensored subjects were used in that analysis. We will refer to the latter approach as the descriptive method (desc) in this chapter.

For implementation of the proposed method, we pre-specified $\tau = 8$ years (32 quarters) and all patients who survived beyond τ were administratively censored. In this situation, 11% of patients were censored at τ . Very few patients died between τ and 40 quarters, the maximum follow-up time, particularly for the distant stage cancer group. Therefore, administrative censoring does not cause much loss of information, but it does avoid the excessive variability in the corresponding cost trajectory estimates caused by

sparse survival data at the tails. All patients censored at τ were still incorporated in the analysis through the model for the LTS group in equation (2.1). Our preliminary analyses suggested that the cost trajectory of the LTS group appears to take an L-shape (Figure A.1). We chose a compound symmetry correlation structure to describe the latent mechanism that drives the fluctuation of costs for each individual. We will show in simulation studies later that the mean trajectory estimates are robust to the choice of correlation structure. Among local or regional stage cancer patients, the estimated residual variance for patients who died prior to τ is comparable with that for LTS (2.45 vs 2.43), with moderate estimated intrasubject correlation (0.43 vs 0.41). Among patients with distant stage cancer, the estimated residual variance for patients who died prior to τ is slightly lower compared with that for LTS (2.16 vs 2.35), with moderate estimated intrasubject correlation (0.42 vs 0.47).

This result supports separate modeling for local/regional stage patients *versus* distant stage patients, and separate parameterization for the LTS group and the rest of the cohort. Further supportive evidence for these modeling and parameterization choices can be seen in the results presented below. To obtain appropriate penalty parameters, we used 10-fold cross-validation (Web Appendix C), and chose the penalty parameters for the local or regional stage cancer model to be $(\lambda_\beta, \lambda_\xi) = (.002, .01)$, and the penalty parameters for the distant stage cancer model to be $(\lambda_\beta, \lambda_\xi) = (.02, .1)$.

Figure 2.1 summarizes the mean cost trajectories, on the log transformed scale, corresponding to 2, 4, and 8 year survival and the LTS group. Pointwise 95% Wald-type confidence intervals were obtained from bootstrap variance estimator; 1000 bootstrap samples were used. Figure 2.2 visualizes the same trajectories but on the original scale. Since the cost data are skewed on the original scale, we plotted the pointwise 25%, 50% and 75% quantiles over time. These quantiles were calculated based on back transfor-

mation and the multivariate normal model. While Figures 2.1 and 2.2 only present cost trajectories corresponding to selected survival times, Figure 2.3 shows all cost trajectories, which form a surface in the 3D space and form a contour plot/heatmap within a triangular region in a 2D space. These panoramic views, particularly the heatmap, clearly show that distant stage cancer patients have higher overall costs than the local or regional stage patients. The heatmap further shows that distant stage cancer patients had longer periods of elevated costs during the initial and terminal care periods. Interestingly, the period of elevated costs during the terminal care phase shortened with increasing survival among distant stage cancer patients, but this pattern is much less evident among the local or regional cancer patients.

The solid curves in Figure 2.1 follow a U-shape from diagnosis to death, which agrees with the motivation behind NCI's POC approach that costs in the initial and terminal care phases are higher than the continuing care phase. In contrast, the costs of the LTS group follow an L-shape, because the terminal care phase may be years away from τ . Figure 2.1 also suggests that the cost difference between local or regional stage and distant stage became less pronounced as the survival time increased. For example, among local or regional stage patients, the estimated mean costs at quarter 12, 24, and 48 for those with 2, 4 and 8 year survival were \$3,836 (95%CI = \$3,667 to \$4,126), \$2,194 (95%CI = \$2,027 to \$2,394) and \$459 (95%CI = \$315 to \$541) respectively. The corresponding costs for the distant stage patients were \$6,640 (95%CI = \$6,049 to \$7,085), \$4,059 (95%CI = \$3,277 to \$4,546) and \$442 (95%CI = \$151 to \$805). A plausible explanation is that distant stage cancer patients who died earlier underwent more aggressive and/or more costly cancer treatments, whereas local or regional stage patients often died from other causes. This result supports separate modeling for the local or regional stage cancer and distant stage cancer, and separate parameterization for the

LTS group and the rest of the cohort. Figure 2.2 shows substantial diversity of medical costs among patients in the cohort. For example, within the first quarter after cancer diagnosis, the 75% quantile medical costs were as high as \$20,000, while the 25% quantile medical costs were around \$1,000. This was expected because during the first two quarters after cancer diagnosis, medical costs highly depend on the type of treatments prescribed to the patients to control cancer progression, such as surgical treatment, radiation therapy and immunotherapy. Figure 2.2 also suggests an interesting association between costs and survival, most notably with the initial and terminal phases. For example, as survival increases from 1 to 8 years, the 50% quantile of the costs in the last quarter increases from \$6,000 to \$8,000 for local or regional stage cancer, and from \$9,000 to \$11,000 for distant stage cancer. The heterogeneity in the longitudinal cost pattern, as discussed above, cannot be obtained by traditional methods that analyze cumulative cost, total cost, or average cost without stratification by survival. This demonstrates the new insight that can be provided by our proposed method.

Next, we report the sampling variability of the estimated cost trajectories. Figure 2.1 demonstrates that the descriptive method produces very wiggly estimates, and the variation sometimes blurs the underlying pattern. In contrast, the proposed one-stage and two-stage estimates are much smoother and easier to interpret. We propose to quantify the overall variability of the estimated cost trajectory by the mean relative efficiency (MRE). The MRE is the pointwise standard error of the estimated two-stage (or one-stage) trajectory, averaged over time, divided by the same averaged pointwise standard error of the estimated trajectory from the descriptive method. In other words, it measures the variability of the proposed method, relative to the variability of the descriptive method. Not surprisingly, both one-stage and two-stage estimators have $MRE < 1$. Further, the two-stage estimator always has lower MRE than the one-stage esti-

mator. This efficiency gain of the two-stage estimator is because it uses cost data from censored patients. This result is justified by our asymptotic theory.

2.5. Simulations

2.5.1. Simulation Design and Data Generation

We investigated the performance of the proposed estimators through extensive simulations. We simulated data that resemble the renal cell cancer cost data in Section 2.4. The simulated data has a substantial proportion of right censoring, and demonstrates L-shaped cost trajectories for LTS and U-shaped cost trajectories for patients whose follow-up ends prior to $\tau = 32$ quarters. We first generated \tilde{T}_i from exponential distribution with rate 0.05 and C_i from an independent exponential distribution with rate 0.025. They were then both rounded to years. The survival time $T_i = \min(C_i, \tilde{T}_i, \tau)$. This numerical setting produced approximately 30% censoring in the data, with 25% censoring prior to τ and 5% censoring at τ . We simulated “bathtub-shaped” mean cost trajectories for subjects who die before τ and L-shaped mean cost trajectory for LTS (trajectories shown in Web Appendix B). Normally distributed noises $\epsilon_i \sim N(\mathbf{0}, \sigma^2 \mathbf{\Sigma}_i)$ were added to the “bathtub-shaped” mean trajectories, where $\sigma = 2$, and $\mathbf{\Sigma}_i$ is a $n_i \times n_i$ exchangeable correlation matrix with correlation $\rho = 0.5$. Noises $\epsilon'_i \sim N(\mathbf{0}, \sigma'^2 \mathbf{\Sigma}'_i)$ were added to the L-shaped mean trajectory, where $\sigma' = 3$ and $\mathbf{\Sigma}'_i$ is a $n_i \times n_i$ exchangeable correlation matrix with correlation $\rho' = 0.2$. These parameter values ensure that the longitudinal costs of LTS have larger variation and lower intrasubject correlation, which reflects more heterogeneity in that group and resembles the renal cell cancer data.

The sample sizes were $n = 500$ or $n = 1000$. We also adjusted the rate of the exponential censoring distribution to 0.06 to produce an alternative scenario with 50% censoring overall. We used 1000 Monte Carlo repetitions in each simulation setting.

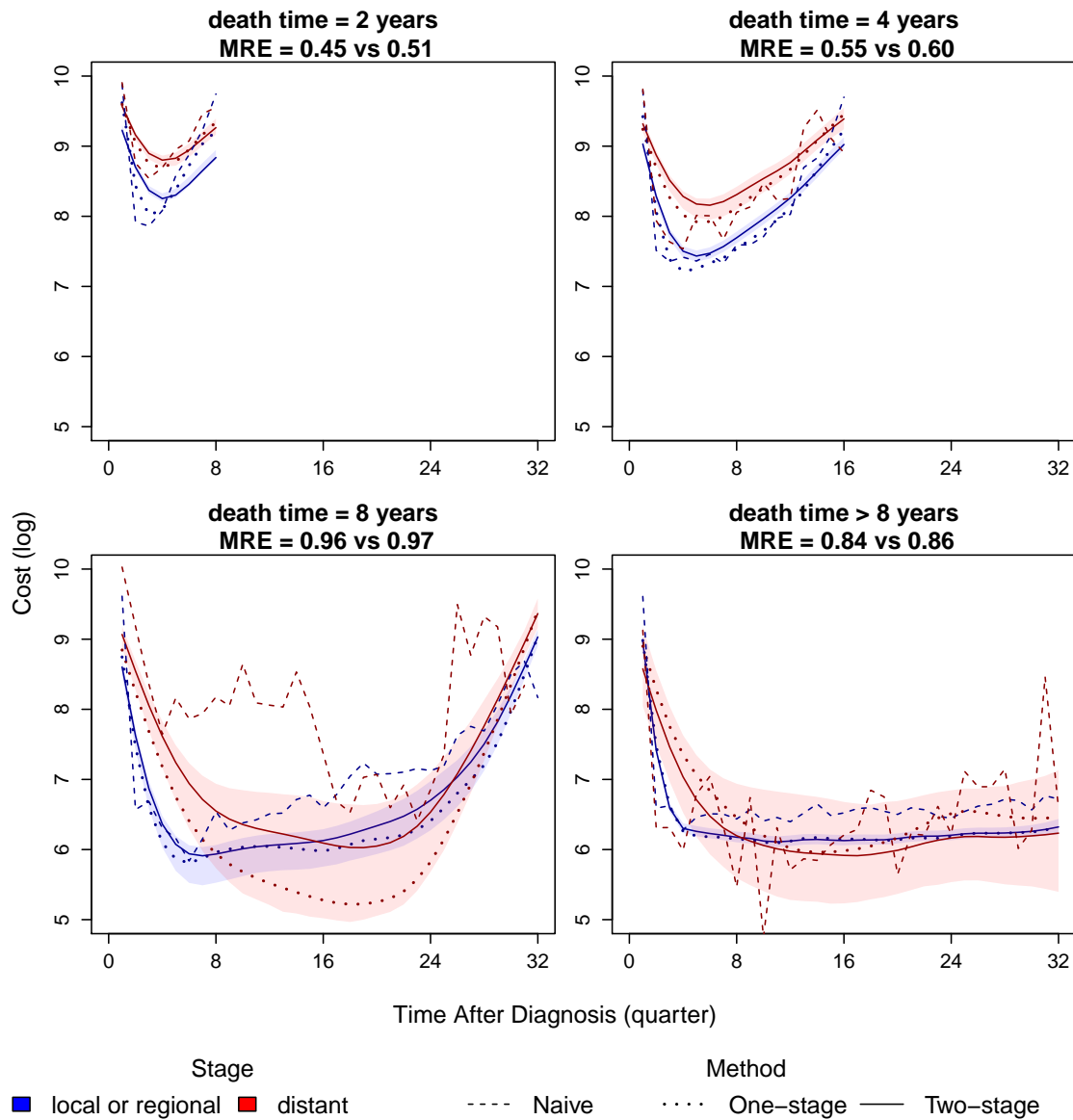


Figure 2.1: SEER data application results for the log transformed cost trajectory estimates. Estimates for patients in local or regional stage (blue) are plotted against estimates for patients in distant stage (red). The survival times are 2, 4, 8, and > 8 years. The dashed curves are descriptive estimates from aggregating quarterly costs among patients who died in the same year. The dotted curves are one-stage estimates. The solid curves are two-stage estimates. The shaded areas are 95% pointwise confidence intervals for the two-stage estimates.

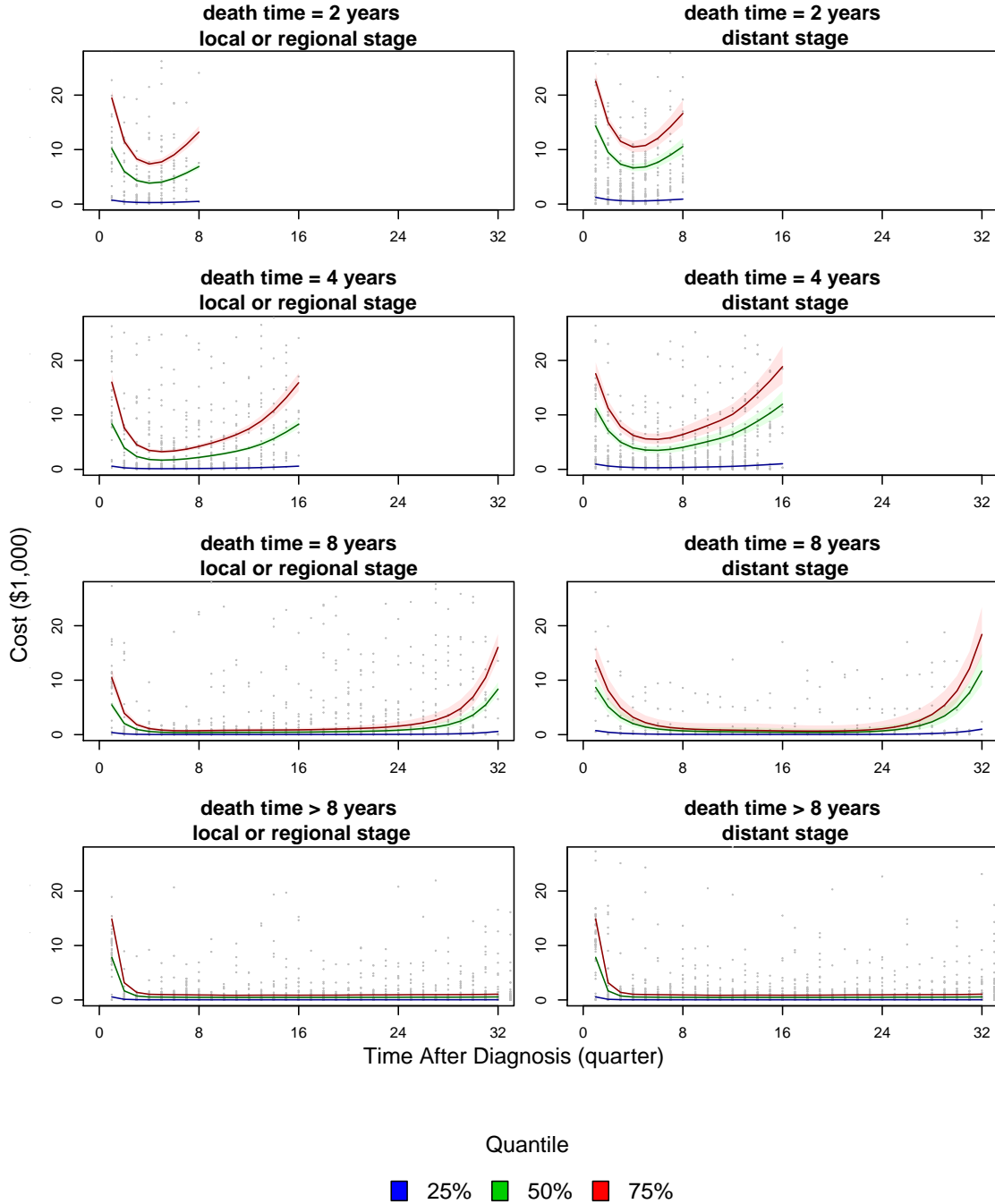


Figure 2.2: SEER data application results for the 25th (blue), 50th (green) and 75th (red) estimated quantiles of quarterly costs (per \$1,000) for patients whose death times are 2, 4, 8, > 8 years using two-stage estimates. Shaded areas present the 95% pointwise confidence intervals. A random sample of 50 uncensored patients was selected from each subgroup and their individual quarterly cost data are shown as gray dots in the background.

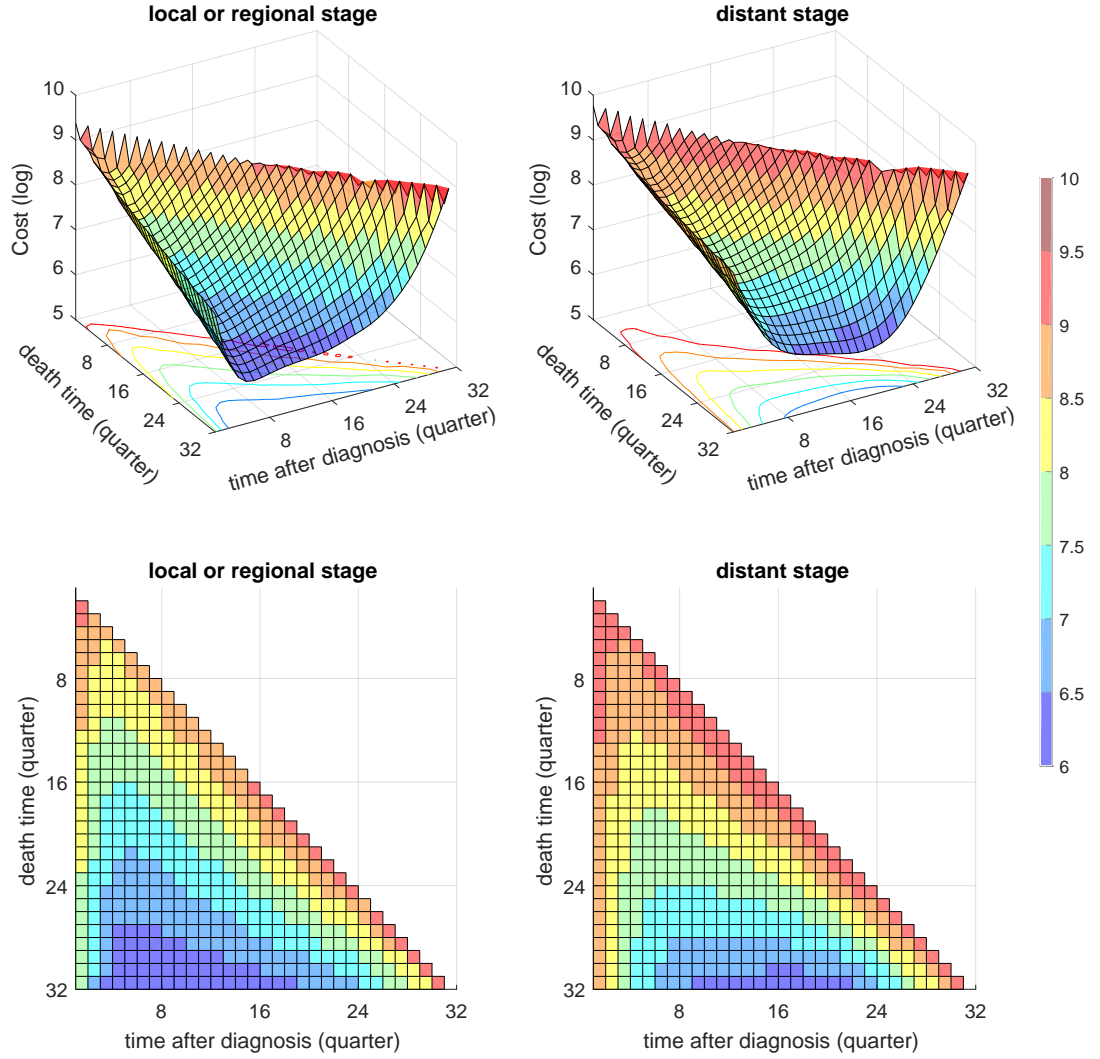


Figure 2.3: SEER data results for the estimated bivariate grid surface of mean cost trajectory conditional on survival. The two-stage estimator is presented in 3D visualization and 2D heatmap.

The point estimators for the variance-covariance parameters were assessed using bias and empirical standard deviation (emp SD). We propose a few aggregated measures to assess the overall quality of the estimated cost trajectory curves. They include the aggregated sample absolute bias (A-AB), aggregated empirical standard deviations (A-SD) and aggregated sample mean squared error (A-MSE). These summary statistics are pointwise absolute bias, empirical standard deviation and mean squared error, averaged across all discrete t given \tilde{T} : $(t, \tilde{T}), t \leq \tilde{T}$.

We fit the model with a penalty on the second order difference in two dimensions, as described in Section 3.2. Using 10-fold cross validation (details in Web Appendix C), we chose $(\lambda_{\beta}, \lambda_{\xi}) = (10^{-4}, 10^{-2})$. The convergence criterion is that for each single parameter, the absolute difference between the current estimate and previous estimate should be less than 10^{-4} . All studies were run on the same computer with an Intel Core i5-6500 3.20 GHz CPU and 16 GB memory.

2.5.2. Simulation Results

Simulation results for the global performance of cost trajectory estimation are presented in Table 2.1. Both one-stage and two-stage estimators exhibit a small bias, with A-AB less than 0.07 relative to the magnitude of the true mean trajectories (between 5 and 10). The bias is higher than that of the descriptive method, probably because of the smoothness penalty. The purpose of using the penalty is to strike a balance between bias and variance; it introduces a small amount of bias in exchange for improvement in mean squared error (MSE) and computational stability. This is shown in the notably improved A-MSEs in the one-stage and two-stage methods. The two-stage estimator outperformed both the descriptive and one-stage estimators with smaller variation (A-SD) and smaller A-MSE. The computational time is longer with the two-stage method, but still very quick. Overall, Table 2.1 demonstrates that the two-stage estimation has

the best performance. The performance of all methods improved with larger sample size and lower censoring percentage. However, when the censoring percentage increases, the A-MSE increases much more rapidly in the one-stage method than in the two-stage method. This is an advantage of the latter, due to its use of longitudinal cost data from censored subjects. In contrast to Table 2.1, which shows the global MSE aggregated over all time points, Figure 2.4 shows them at a more granular level at each time point; the conclusion is the same. First, the two-stage method has the lowest MSE and this result holds at all time points. Second, the two-stage method is more resistant to data loss due to censoring, because it uses the longitudinal cost data from censored subjects. As more subjects are censored when the survival is longer, this advantage of the two-stage method is better demonstrated at year 8 than at year 4. Results in Table 2.2 suggest that the two-stage estimators of the variance-covariance parameters $(\sigma, \sigma', \rho, \rho')$ generally work well. These parameters help to access the variation of the mean cost estimation on the original scale, which is shown in Figure 2.4. The efficiency of these estimators improves with larger sample size and lower censoring rate, as expected. When the censoring rate increases to 50%, the percent bias from estimators in Table 2.2 can be up to 10% while the percent bias from estimators Table 2.1 remains very small. Censoring seems to have a higher impact on the variance-covariance parameters than on the mean cost trajectories. In other words, the increased bias in the estimated variance-covariance parameters is not carried over to the mean cost trajectory estimation. When the censoring rate increases, there are generally less repeated measures per subject at different time points, especially at later time points. This results in loss of information for estimating the variance-covariance parameters. However, estimation of the mean function of model (2.1) is not affected. We speculate that this phenomenon is analogous to the property of generalized estimating equation (Liang and Zeger, 1986a) with identity link function. To investigate this issue systematically, we ran additional simulations to investigate the

impact of covariance matrix misspecification and violation of normality assumption, such as bi-modal and heavy-tailed residuals. The results, provided in Web Appendix B, shows empirically that the two-stage method is robust against moderate misspecification under the aforementioned scenarios.

We also explored how the penalty matrix affects the model performance (Web Appendix C). Besides the interesting finding that penalty on two directions results in lower A-MSEs than penalty on one direction, we found that the model using second order difference penalty outperforms the model using first order difference penalty in terms of A-MSEs, and therefore second order difference penalty was used for the renal cell cancer data application.

Furthermore, we compared our two-stage method to a modified version of the EM algorithm proposed in Li et al. (2018) replacing the truncated quadratic bases by our flexible discretized bases to eliminate the impact of model misspecification caused by knots selection. We observed that the EM algorithm has substantial biases ($A-AB=5.696$) when the LTS group are ignored. When the LTS group is properly modeled, the $A-AB$ of EM algorithm reduced to 0.074 with $A-MSE$ equals 0.084, while the averaged runtime is 232.3 seconds which is notably more than our proposed two-stage method. This result suggests that accounting for the LTS group is helpful to reduce bias, and the proposed two-stage estimation achieves substantial computational advantage with small loss of efficiency compared to the EM algorithm proposed in Li et al. (2018).

Lastly, we ran additional simulation studies with survival time rounded by quarter instead of by year. As expected, there was little bias using second order difference penalty, but the estimated trajectories exhibited larger variation with longer model fitting runtime. We speculate that this is due to the reduced sample size at the tail of the

Table 2.1: Simulation results to compare the descriptive, one-stage (1-stg) and two-stage (2-stg) methods under different sample size and censoring rate. n : sample size; C%: censoring rate (%); A-AB: aggregated sample absolute bias; A-SD: aggregated empirical standard deviations; A-MSE: aggregated sample mean squared error. Results are averaged over 1000 Monte Carlo samples. Runtime is in second.

n	C%	A-AB	A-SD	A-MSE	Runtime (seconds)
		(desc, 1-stg, 2-stg)	(desc, 1-stg, 2-stg)	(desc, 1-stg, 2-stg)	(desc, 1-stg, 2-stg)
500	30	(0.011, 0.050, 0.069)	(0.453, 0.311, 0.251)	(0.221, 0.109, 0.076)	(0.1, 7.9, 36.3)
1000	30	(0.008, 0.049, 0.066)	(0.316, 0.221, 0.180)	(0.106, 0.057, 0.043)	(0.1, 16.0, 79.3)
500	50	(0.019, 0.056, 0.089)	(0.694, 0.441, 0.293)	(0.551, 0.222, 0.111)	(0.1, 5.8, 49.3)
1000	50	(0.014, 0.064, 0.092)	(0.476, 0.310, 0.208)	(0.258, 0.117, 0.065)	(0.1, 11.6, 110.4)

Table 2.2: Simulation results for the two-stage estimates for variance coefficients under different sample size and censoring rate. n : sample size; C%: censoring rate; Bias: empirical bias; emp SD: empirical standard deviations. Results are from 1000 mean point estimates of simulated samples.

n	C%	Bias				Emp SD			
		$\sigma = 2$	$\sigma' = 3$	$\rho = .5$	$\rho' = .2$	$\sigma = 2$	$\sigma' = 3$	$\rho = .5$	$\rho' = .2$
500	30	0.093	-0.137	-0.019	0.045	0.062	0.061	0.030	0.028
1000	30	0.098	-0.131	-0.020	0.044	0.045	0.040	0.021	0.020
500	50	0.140	-0.288	-0.030	0.075	0.058	0.064	0.029	0.031
1000	50	0.144	-0.280	-0.033	0.074	0.042	0.044	0.020	0.022

survival distribution and an increase in the number of parameters. We ran simulations deleting relative high costs (> 8) at certain rate (10%), and found the A-AB and A-MSE remains small. This result justified the model robustness when the costs are missing not at random.

2.6. Discussion

In this article, we proposed a two-stage semiparametric likelihood-based approach to estimate the conditional distribution of longitudinal medical cost trajectory given the time of terminal event; consistency and asymptotic normality were proved. Our data application and simulation results demonstrated favorable performance of the proposed method over existing methods. The proposed methodology offers insight into the cost trends, associations between cost and survival, and intrinsic heterogeneity over time and

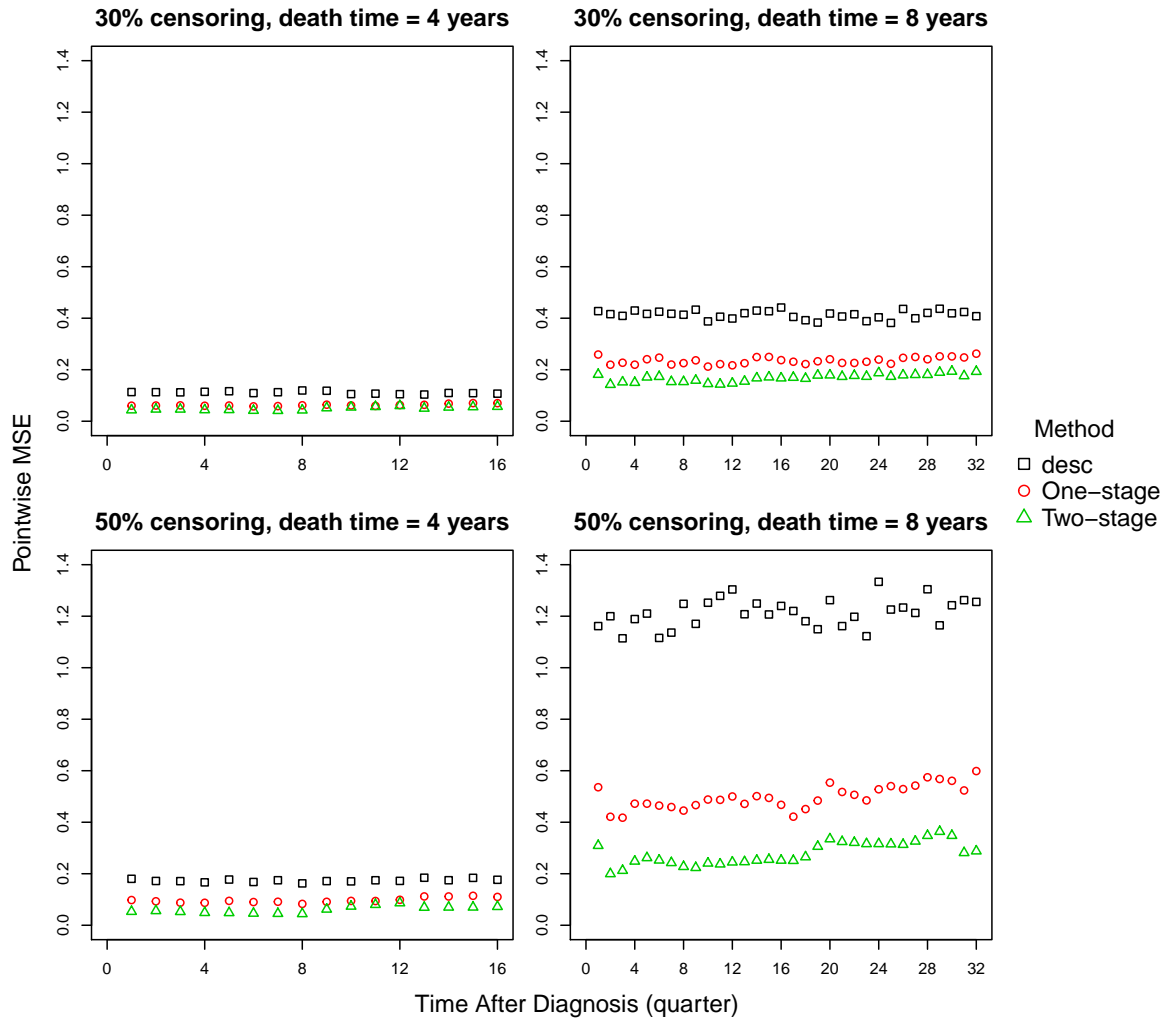


Figure 2.4: Simulation results for pointwise mean squared error (MSE) of the trajectory estimates yielding 30% and 50% censoring rate for patients whose death time are 4 and 8 years. Sample size $n = 500$. Pointwise MSE for descriptive, one- and two-stage methods, averaged over 1000 replications, are plotted in rectangle, circled, and triangle symbols, respectively.

by subgroups. Such information is not easy to obtain using traditional analyses that target cumulative cost, total cost, or costs with phase of care.

An important feature of the proposed methodology is that it uses cost data from both censored and uncensored subjects. This is achieved by using a joint model of the longitudinal cost data and survival. This joint distributional assumption is necessary to incorporate cost data of censored patients into the trajectory estimation. As demonstrated by our comparison between the one- and two-stage estimates, adequate use of cost data from censored patients helps to improve the statistical efficiency, particularly when the censoring rate is high. Analogous to existing literature, we needed a distributional assumption on the cost data and survival to properly use cost data from censored patients. It is therefore important to perform model diagnosis to make sure that the distributional assumption holds; this can be done by examining the survival groups of uncensored subjects. Descriptive goodness of fit procedures could be performed, and more flexible distributional assumptions (than normal distribution) are feasible. Due to the flexibility in the proposed basis functions, the mean cost trajectory is free from mean structure misspecification. Our simulation shows that the estimation is fairly robust against moderate deviation from multivariate normality or misspecification of the variance-covariance matrix.

Our model applies to situations where the proportion of zero costs is small. An effective way to reduce zero incident costs is to aggregate cost data into longer time intervals, such as grouping monthly costs into quarterly costs. When a substantial proportion of incident costs are zero, alternative models that can accommodate zero-inflation and skewness should be used; our 3 will explore such models.

We have not considered covariates in this chapter, since modeling the covariate

effect on cost trajectories involves strong assumptions when such trajectories are nonlinear. The purpose of this research was to estimate the cost trajectories of a well-defined patient population, and the results will be informative for future cost projections on the same population. In Section 2.4, we showed that if researchers are interested in how baseline covariates such as cancer stage at initial diagnosis affect cost trajectories, they can perform cost estimation on each stratum defined by the covariates, and compare results across strata visually. This is often feasible in population cost estimation, as we used a nationally representative database with a large sample size. In Chapter 4 we will extend the current methodology to handle baseline covariates.

CHAPTER 3

AN EXTENSION OF ESTIMATING EQUATIONS TO MODEL LONGITUDINAL MEDICAL COST TRAJECTORY WITH MEDICARE CLAIMS DATA LINKED TO SEER CANCER REGISTRY*

3.1. Introduction

Insurance claims data include reimbursed medical costs associated with every clinical encounter for each individual throughout their period of enrollment in the insurance plan. An important use of these cost data in health services research is to study the population averaged longitudinal pattern of medical costs from disease onset or diagnosis to a terminal event such as death. The overall goal of this study is to quantify and interpret the population mean cost trajectories given their survival times, which produces population medical cost projection when coupled with disease incidence and prevalence data in the population. This information helps health policy makers better understand the financial burden of a disease throughout its life course, such knowledge is critically important in future planning and allocation of healthcare resources.

The motivating dataset of this chapter consists of a national cohort of 184,491 men who were diagnosed with prostate cancer between 2003 and 2015. The data were extracted from the SEER-Medicare database and include medical costs aggregated to monthly intervals from diagnosis to the end of follow-up or death. The descriptive statistics depicted in Figure 3.1 reveal several statistical challenges that are not adequately addressed by the POC approach. Figure 3.1(a) shows that the monthly cost data are highly variable and the trajectory of average monthly costs appears to depend on sur-

*A version of Chapter 3 has been under revision in the Annals of Applied Statistics.

vival. Trajectories generally follow a U-shaped pattern among individuals who died at months 24, 48, 72, and 96 after the initial diagnosis, but an L-shaped curve among those who survived beyond 120 months. While the cost trajectory escalates in the months toward the end of life, the increase is continuous without an obvious uptick right around 12 months before death, suggesting a lack of objective justification for the terminal care phase definition. Figure 3.1(a) also supports the notion that the cost trajectory should be estimated as a nonlinear function conditional on survival time, with minimum assumption on the shape of the curve. Both Figures 3.1(a) and 3.1(b) demonstrate that the cost data are skewed to the right, and have a substantial proportion of patients with zero monthly costs (31.4%), a phenomenon often termed zero-inflation. Figure 3.1(c) shows the widely recognized cost data challenge of the variation increases with the mean. This heteroscedasticity poses a difficulty when making statistical inferences on cost trajectory. Figure 3.1(d) shows that a substantial proportion of individuals could be censored by the date of data extraction, resulting in missing information on cost and survival beyond the maximum follow-up time. If the cost trajectory and survival are correlated, then this identifiability issue needs to be properly addressed in the model formulation and estimation.

The discussion above motivates the development of a flexible, data-adaptive method to address two challenges in the estimation of cost trajectory conditional on survival. First, the method needs to include, instead of excluding, the cost data from censored individuals, and it must properly account for a lack of identifiability in survival time when the association between survival and cost is explored. Nonetheless, we allow a flexible model for estimating the nonlinear cost trajectory and its dependence on survival, which is not always observed due to censoring.

The second challenge addressed by this chapter is that the method needs to

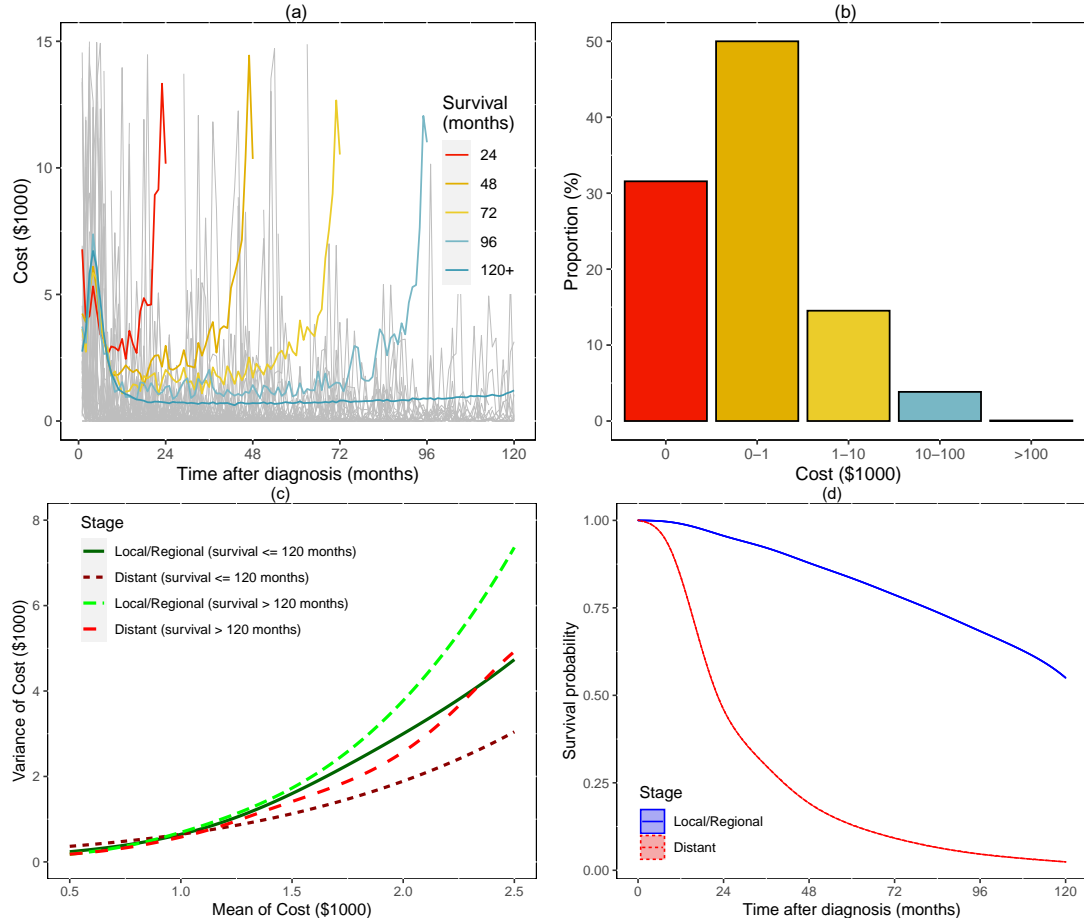


Figure 3.1: Descriptive analysis of the monthly medical costs of prostate cancer patients in SEER-Medicare data. (a) The trajectory of average monthly costs among subjects who died at selected months. A random sample of 50 uncensored subjects was selected and their individual monthly cost data are plotted in the background as gray lines. (b) The histogram of monthly costs per 1,000 US dollars. (c) The relationship between the mean and variance of the cost data, stratified by survival and by cancer stages determined at the time of diagnosis. (d) The estimated survival distributions by cancer stages. A substantial proportion of individuals with local regional stage are censored by the end of follow-up.

properly handle non-normality in the cost data, including skewness, zero-inflation, and heteroscedasticity. In other words, by estimating “cost trajectory”, we specifically refer to estimating the mean incident costs over time among individuals with the same survival. Monotone transformation (e.g., logarithm) of the cost data reduces skewness and heteroscedasticity, but not zero-inflation. We propose to model the mean cost via a novel extension of the generalized estimating equation (GEE) methodology. We considered the first-order (GEE1; Liang and Zeger (1986b)) and second-order (GEE2; Zhao and Prentice (1990)) implementations. Both approaches can be used to model the mean cost directly without fully specifying the distribution of the cost data or explicitly dealing with zero-inflation. The GEE2 further improves the efficiency of estimation by modeling the mean-variance heteroscedasticity. In this chapter, the mean incident cost depends on two time variables: the time of the incident cost (e.g., month since diagnosis) and the survival time. The relationship between cost trajectory and survival time is nonlinear in a bivariate surface and semiparametrically modeled. The survival time is subject to censoring. In the GEE literature, there is extensive discussion on semiparametric modeling of the mean structure with smoothing methods (for a comparison see Welsh et al. (2002)). However, none of the existing methods studied the situation when a covariate is subject to censoring. We developed a novel extension for the GEE algorithm to deal with this challenge.

The complicated dependence between nonlinear cost trajectory and survival, as visualized in Figure 3.1(a), can in theory be studied by joint modeling of longitudinal costs and survival data. However, these methods usually require distributional assumption on the data in order to construct the joint likelihood and/or use shared random effects to account for the dependence between longitudinal costs and survival data. In the context of our research question, it is computationally intensive or even prohibitive to

fit a joint model with a large number of fixed and random effect parameters to model the nonlinear population mean cost trajectories given survival time with splines in the likelihood function. Our proposed approach in this chapter avoids these problems with the modeling assumptions and computation.

The rest of the chapter is organized as follows: in Section 3.2, we define the generalized estimating equation for modeling the longitudinal cost trajectories. Section 3.3 elaborates on the estimation and inference procedure. A real data application is given in Section 3.4. Section 3.5 presents simulations to investigate the finite sample performance of the proposed method. Further discussions and conclusions are provided in Section 3.6. Additional theoretical proofs and numerical results are provided in the Appendix B.

3.2. Notation and Model

We adopt the notation in Chapter 2. Suppose we have i.i.d. data from n subjects, $\{\mathbf{Y}_i, T_i = \min(\tilde{T}_i, C_i), \delta_i = 1\{\tilde{T}_i \leq C_i\}; i = 1, \dots, n\}$ where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ denotes a vector of medical costs corresponding to fixed-length time intervals (e.g. monthly, quarterly or annually) up to observed survival time T_i , $j = 1, \dots, n_i$. \tilde{T}_i denotes the time of death; C_i denotes the time to censoring; δ_i denotes the event indicator for which $\delta_i = 1$ if death occurs prior to censoring, and 0 otherwise. τ denotes the maximum follow-up time by study design, which can be set empirically at $\max\{T_i; i = 1, \dots, n\}$, or alternatively a prespecified cutoff time of interest. Let $n_i = \lfloor T_i \rfloor$ be the number of cost measurements where $\lfloor T_i \rfloor$ is the largest integer smaller or equal to T_i . $\mathbf{t}_i = (t_{i1} = 1, t_{i2} = 2, \dots, t_{in_i} = \lfloor T_i \rfloor)^T$ denotes the vector of measurement times. We assume that C_i and $\{\tilde{T}_i, \mathbf{Y}_i\}$ are independent. We have used $\|\cdot\|_2$ to denote the L_2 norm, $|\cdot|$ to denote the sum of all matrix entries, and \otimes to denote outer product.

In this chapter, the goal is to estimate the mean cost trajectory given survival time at population level. Since the shape of the cost trajectory is typically nonlinear and the shape depends on the length of the trajectory, i.e., the survival duration, we formulate the estimand as a bivariate function:

$$E\{Y(t)|\tilde{T} = s\} = \mu(t, s), \quad 0 < t \leq s \leq \tau. \quad (3.1)$$

This is the conditional expectation of the cost at month t from diagnosis, $Y(t)$, given that the survival time is s . Here t is on a discrete scale, by the definition of longitudinal incident cost. But for the ease of statistical modeling, we treat both t and s as continuous variables and model $\mu(t, s)$ as a bivariate smooth surface. Two-dimensional splines basis functions such as the outer product of B-splines (De Boor, 1978) can be used to model this nonlinear surface: $\mu(t, s) = \mathbf{B}_{12}(t, s)\boldsymbol{\theta}_{12}$. We use B-spline basis throughout this chapter because of its numerical stability and theoretical properties. Since medical costs are recorded prior to time to death, the domain of the bivariate function $\mu(t, s)$ forms an upper triangular area $0 < t \leq s \leq \tau$. Thus, we re-express the bivariate basis functions through a shrinkage-expansion transformation $\mathbf{B}_{12}(t, s) = \mathbf{B}_1(t \cdot \tau/s) \otimes \mathbf{B}_2(s)$, where $\mathbf{B}_1(t \cdot \tau/s)$ is the spline basis for rescaled t , $t \cdot \tau/s$, with order p_1 and K_1 knots, $\mathbf{B}_2(s)$ is the spline basis for s with order p_2 and K_2 knots.

In insurance claims data, it is common to have a substantial proportion of subjects being censored at τ , which is due to the relatively short follow-up in comparison to the life expectancy of the study population. We call subjects who survived beyond τ long-term survivors (LTS). For example, in our motivating dataset, 17.6% of the prostate cancer patients were alive and censored at year 10 after diagnosis. If a subject was censored prior to τ , proper statistical estimation requires accounting for the probability of them being in the LTS group. We define the mean cost trajectory among all subjects in the

LTS group as

$$E\{Y(t)|\tilde{T} > \tau\} = E\{Y(t)|\tilde{T} = \tau+\} = \mu(t, \tau+), \quad 0 < t \leq \tau. \quad (3.2)$$

For notation convenience, we let $s = \tau+$ be a generic notation of LTS survival time. The univariate function $\mu(t, \tau+) = \mathbf{B}_3(t)\boldsymbol{\theta}_3$ is modeled using one-dimensional spline basis $\mathbf{B}_3(t)$ with order p_3 and K_3 knots. Models (3.1) and (3.2) constitute our proposed model for the cost trajectory estimation problem. We can express these two models together by concatenating their basis functions $\mathbf{B}_\mu(t, s) = [I(s \leq \tau)\mathbf{B}_{12}(t, s), I(s > \tau)\mathbf{B}_3(t)]$ and mean parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_{12}^T, \boldsymbol{\theta}_3^T]^T$ respectively. $I(\cdot)$ denotes the indicator function here.

3.3. Estimation

We first present a simple cost trajectory estimator by using data from subjects who died prior to τ and subjects who were censored at τ . τ is the maximum follow-up time $\max\{T_i\}, i = 1, \dots, n$, or a prespecified cut-off time such as 10 years in our data example. Then we propose a novel induced generalized estimating equation (GEE) to adjust for a covariate (\tilde{T}) that is subject to right censoring prior to τ . Next, we incorporate variance-covariance modeling to the GEE to account for heteroscedasticity in the data. Lastly, we make recommendations on knots and smoothing parameter selections.

3.3.1. Estimation without using data from subjects censored prior to τ

We first use data from uncensored subjects to estimate $E\{Y(t)|\tilde{T} = s\} = \mu(t, s), 0 < t \leq s \leq \tau$ in equation (1) and data from those censored at τ to estimate $E\{Y(t)|\tilde{T} > \tau\} = \mu(t, \tau+), 0 < t \leq \tau$ in equation (2). These estimators are consistent under the independent censoring assumption. Denote $\Delta_i = I(\{T_i \leq \tau, \delta_i = 1\} \text{ or } \{T_i = \tau, \delta_i = 0\})$ as an indicator for being either an uncensored patient or censored at τ . Thus, $\Delta_i I(\delta_i = 1)$ indicates that the data is from uncensored subjects, and $\Delta_i I(\delta_i = 0, T_i = \tau)$ indicates

that the data is from those censored at τ .

Designed analogously to the generalized estimating equation, it is the solution to the following penalized estimating equation:

$$\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{D}_i \boldsymbol{\theta}) - J_\lambda(\boldsymbol{\theta}) \boldsymbol{\theta} = 0. \quad (3.3)$$

where $\mathbf{D}_i = (D_{i1}^T, \dots, D_{in_i}^T)^T$ is the corresponding joint design matrix, with the j th column $D_{ij} = [\Delta_i I(\delta_i = 1) \mathbf{B}_\mu(t_{ij}, T_i), \Delta_i I(\delta_i = 0, T_i = \tau) \mathbf{B}_\mu(t_{ij}, \tau+)]$. \mathbf{V}_i is a working covariance matrix that should approximate $\text{cov}\{\mathbf{Y}_i | T_i, \delta_i\}$ but is allowed to be misspecified without affecting the consistency of estimators. We include a roughness penalty term in the equation to protect against over-parameterization by the spline basis terms. The penalty function $J_\lambda(\boldsymbol{\theta})$ is chosen as the commonly used difference-based penalty matrices described in Eilers and Marx (1996) with penalty parameter λ . Given \mathbf{V}_i and the penalty function, the solution to (3.3) has a closed form expression:

$$\tilde{\boldsymbol{\theta}} = \left\{ \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i + J_\lambda(\boldsymbol{\theta}) \right\}^{-1} \left\{ \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{Y}_i \right\}. \quad (3.4)$$

This expression and an estimator of \mathbf{V}_i can be obtained through iterative steps of a GEE procedure. We can also obtain a sandwich variance estimator of $\tilde{\boldsymbol{\theta}}$ as $\text{cov}(\tilde{\boldsymbol{\theta}}) = \mathbf{H}_{n,\lambda}^{-1} \mathbf{M}_n \mathbf{H}_{n,\lambda}^{-1}$, where $\mathbf{M}_n = \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{D}_i \tilde{\boldsymbol{\theta}}) (\mathbf{Y}_i - \mathbf{D}_i \tilde{\boldsymbol{\theta}})^T \mathbf{V}_i^{-1} \mathbf{D}_i$ denotes the variance of the estimating equations and $\mathbf{H}_{n,\lambda} = \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i + J_\lambda(\boldsymbol{\theta})$ denotes the Hessian matrix. Maximum efficiency is achieved if the covariance matrix is correctly specified, as shown in the Appendix B. A sandwich variance estimator for the estimated cost trajectory $\tilde{\mu}(t, s) = \mathbf{B}_\mu(t, s) \tilde{\boldsymbol{\theta}}$ can also be obtained as $\text{cov}(\tilde{\mu}(t, s)) = \mathbf{B}_\mu(t, s) \text{cov}(\tilde{\boldsymbol{\theta}}) \mathbf{B}_\mu(t, s)^T$. The robustness against the misspecification of the covariance matrix (and hence the distribution) is particularly important for our cost trajectory estimation problem. Due to

the use of flexible spline models, the chance of misspecifying the mean structure is greatly reduced, but the misspecification of the variance of the data remains highly likely because cost data are usually non-normal with skewness, zero-inflation, and heteroscedasticity.

3.3.2. Estimation with data from censored subjects

The estimator from Equation (3.4) is not efficient because it does not use data from subjects censored prior to τ . We now consider using all available data to estimate the mean and variance of the incident costs of a generic subject i with the following mean function induced by the adjustment for censoring:

$$\begin{aligned} E\{Y_{ij}|T_i, \delta_i\} = & I(T_i \leq \tau, \delta_i = 1)E\{Y_{ij}|\tilde{T}_i = T_i\} + I(T_i = \tau, \delta_i = 0)E\{Y_{ij}|\tilde{T}_i > \tau\} \\ & + I(T_i < \tau, \delta_i = 0)E\{Y_{ij}|\tilde{T}_i > T_i\}. \end{aligned} \quad (3.5)$$

This decomposition motivates accounting for subjects censored prior to τ , conditional on the observed survival time and censoring status:

$$E\{Y_{ij}|T_i < \tau, \delta_i = 0\} = E\{Y_{ij}|\tilde{T}_i > T_i\} = \frac{\int_{T_i}^{\infty} \mu(t_{ij}, s) dF_{\tilde{T}}(s)}{1 - F_{\tilde{T}}(T_i)}, \quad \text{cov}\{\mathbf{Y}_i|\tilde{T}_i > T_i\} = \mathbf{V}_i, \quad (3.6)$$

where $F_{\tilde{T}}(s)$ is the cumulative density function of survival time \tilde{T} . The integration can be decomposed into two parts $\int_{T_i}^{\tau} \mu(t_{ij}, s) dF_{\tilde{T}}(s) + \mu(t_{ij}, \tau+)[1 - F_{\tilde{T}}(\tau)]$. This induced mean function connects the conditional mean of the cost given the observed survival status to models (3.1) and (3.2), the conditional means given the possibly unobserved true survival status. The models (3.1) and (3.2) have a covariate \tilde{T} that is subject to censoring. As a result, those mean structures cannot be used directly in GEE if one wants to incorporate data from censored subjects into the estimation. The induced mean function is based on the observed quantities, hence it can be used to construct the GEE, as shown below. To the best of our knowledge, this is the first chapter on a GEE model with a censored

covariate.

We propose the following penalized generalized estimating equation that utilizes all available data:

$$\sum_{i=1}^n \mathcal{D}_i^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathcal{D}_i \boldsymbol{\theta}) - J_\lambda(\boldsymbol{\theta}) \boldsymbol{\theta} = \mathbf{0}. \quad (3.7)$$

The induced design matrix can be decomposed into two parts $\mathcal{D}_i = \Delta_i \mathbf{D}_i + (1 - \Delta_i) E(\mathbf{D}_i)$. Without the second term, the solution reduces to that of Equation (3.4). $E(\mathbf{D}_i) = (E(D_{i1}^T), \dots, E(D_{in_i}^T))^T$ denotes the expected design matrix with respect to the true survival time for subjects who were censored before τ , where

$$E(D_{ij}) = \frac{1}{1 - F_{\bar{T}}(T_i)} \left\{ \int_{T_i}^{\tau} \mathbf{B}_\mu(t_{ij}, s) dF_{\bar{T}}(s) + \mathbf{B}_\mu(t_{ij}, \tau+) [1 - F_{\bar{T}}(\tau)] \right\}.$$

Note that this quantity depends on the survival distribution (see Section 4.3 for estimation). Given the variance function \mathbf{V}_i , the solution of Equation (3.7) is

$$\hat{\boldsymbol{\theta}} = \left\{ \sum_{i=1}^n \mathcal{D}_i^T \mathbf{V}_i^{-1} \mathcal{D}_i + J_\lambda(\boldsymbol{\theta}) \right\}^{-1} \left\{ \sum_{i=1}^n \mathcal{D}_i^T \mathbf{V}_i^{-1} \mathbf{Y}_i \right\}. \quad (3.8)$$

The specification and estimation of the variance function in the presence of censored data are discussed in Section 3.3.4. The iterative estimation of the mean and variance functions in the GEE steps are discussed in Section 3.3.6.

3.3.3. Estimating the survival distribution

The survival cumulative density function $F_{\bar{T}}(s)$ is an indispensable part of the GEE estimation with induced mean function. We use a flexible penalized spline to estimate $F_{\bar{T}}(s)$ with a finite number of nuisance parameters (Kauermann, 2005; Cai et al., 2002). Let $\log h(s) = \mathbf{B}_h(s) \boldsymbol{\beta}$ be a spline approximation for the log-hazard function, where

$\mathbf{B}_h(s)$ is a vector of basis functions with order p and K knots, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{p+K})^T$. The estimator $\hat{\boldsymbol{\beta}}$ can be obtained by fitting a Poisson model of pseudo-observations (Cai et al., 2002; Li and Greene, 2008). To derive the Poisson model, we first approximate numerical integration through the trapezoidal rule. Let $0 = d_0 < d_1 < \dots < d_M = \tau$ be a grid of points that span the range of the study period. Define index M_i through $d_{M_{i-1}} < T_i \leq d_{M_i}$. $d_{M+1} = \infty$, and $M = 100$ is considered a sufficiently large slice. The approximation in survival estimating equations implies a strong relationship to penalized quasi-likelihood estimation in a generalized linear mixed model (Breslow and Clayton, 1993; Shun and McCullagh, 1995). Thus, the selection of λ' can use the Poisson approximation to construct a GCV criteria

$$\text{GCV}(\lambda'; \boldsymbol{\beta}) = \frac{n^{-1} \sum_{i=1}^n \sum_{m=1}^{M_i} \text{Dev}(\tilde{Y}_{\text{im}}, \Lambda_{\text{im}}; \boldsymbol{\beta})}{\{1 - p(\lambda'; \boldsymbol{\beta})/n\}^2},$$

where the numerator is the Poisson deviance, and the definition of $p(\lambda'; \boldsymbol{\beta}) = p_{\boldsymbol{\beta}} s_{\boldsymbol{\beta}}$ is the effective number of parameters in the penalized model, where $p_{\boldsymbol{\beta}} = p + K$ is the number of covariates, $s_{\boldsymbol{\beta}} = \|\boldsymbol{\beta}(\lambda')\|_2 / \|\boldsymbol{\beta}(0)\|_2$ is the shrinkage rate of the penalized GEE estimator over the non-penalized GEE estimator.

3.3.4. Variance-covariance model

The medical costs may have zero-inflation, skewness to the right, and heteroscedasticity in the sense that the variability of the cost increases with the average. While the proposed GEE algorithm produces a consistent cost trajectory estimator without the need for explicitly modeling the distribution of the cost data, incorporating some of these features into the estimation could further improve efficiency. For that purpose, we propose to apply GEE2 (Liang et al., 1992) to our algorithm by modeling the heteroscedasticity.

The variance function is decomposed as $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$. In this expression,

$\mathbf{R}_i(\boldsymbol{\alpha})$ is a symmetric working correlation matrix (e.g., compound symmetry or AR(1)) of \mathbf{Y}_i with parameter $\boldsymbol{\alpha}$; $\mathbf{A}_i = \text{diag}\{\text{var}(Y_{i1}), \dots, \text{var}(Y_{in_i})\}$ with $\text{var}(Y_{ij}) = \mathcal{V}(\mu_{ij})$, where $\mu_{ij} = \mu(t_{ij}, \tilde{T}_i)$ and $\mathcal{V}(\cdot)$ is a smooth function that quantifies the relationship between variance and mean. We will estimate this function nonparametrically. Let the variance model for subjects who die prior to τ be $\epsilon_{ij}^2 = (Y_{ij} - \mu_{ij})^2 = \exp\{\phi(\mu_{ij}) + \xi_{ij}\}$ where ξ_{ij} is an independent error term with mean zero. The exponential function is used here to ensure that the variance is positive. The unknown smooth $\phi(\cdot)$ can be estimated by minimizing the penalized least square among the uncensored subjects

$$\sum_{i=1}^n \delta_i \{\log \hat{\epsilon}_{ij}^2 - \phi(\hat{\mu}_{ij})\}^2 - \mathcal{J}(\phi), \quad (3.9)$$

where $\mathcal{J}(\phi)$ is appropriate penalty term with respect to the spline approximation of $\phi(\cdot)$. The correlation parameter $\boldsymbol{\alpha}$ can be estimated by solving estimating equations based on distinct pairwise products of residuals denoted as $\mathbf{U}_i(\boldsymbol{\theta}, \boldsymbol{\beta}) = (\epsilon_{i1}\epsilon_{i2}, \epsilon_{i1}\epsilon_{i3}, \dots, \epsilon_{i,n_i-1}\epsilon_{i,n_i})^T$. Let $\rho_i(\boldsymbol{\alpha})$ be the corresponding expectations given the prespecified correlation structure. The estimator for $\boldsymbol{\alpha}$ can be solved by the second moment estimating equations (Liang and Zeger, 1986b; Fitzmaurice et al., 2008)

$$\sum_{i=1}^n \delta_i \{\partial \rho_i(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}\}^T \{\mathbf{U}_i(\boldsymbol{\theta}, \boldsymbol{\beta}) - \rho_i(\boldsymbol{\alpha})\} = 0. \quad (3.10)$$

In our preliminary analysis, we observed that the intra-subject correlation of incidence costs among those censored at τ is relatively smaller than the rest of the subjects, presumably because this group include subjects with greater variability of \tilde{T} . Therefore, a separate model of the variance function and correlation parameters for LTS should be specified; the estimation follows Equations (3.9) and (3.10), with the indicator replaced by $I(\delta_i = 0, T_i = \tau)$

3.3.5. Variance estimation with the sandwich method

In the estimation with Equation (3.7), the survival distribution needs to be estimated at an estimator from Section 3.3.3. While this plug-in approach does not affect the consistency of the point estimator, rigorous variance estimation requires properly accounting for the sampling variability from the estimated survival distribution. For this purpose, we use the following sandwich variance formula for both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$:

$$\text{cov}([\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\beta}}^T]^T) = \mathbf{H}_{n,\lambda}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})^{-1} \mathbf{M}_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) \mathbf{H}_{n,\lambda}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})^{-1}, \quad (3.11)$$

where $\mathbf{M}_n(\boldsymbol{\theta}, \boldsymbol{\beta})$ is the sum of squares of the joint penalized spline estimating equation matrix in equation (3.7) and the survival estimating equation; the corresponding Hessian matrix is approximated by

$$\mathbf{H}_{n,\lambda}(\boldsymbol{\theta}, \boldsymbol{\beta}) \approx - \sum_{i=1}^n \begin{bmatrix} \mathcal{D}_i^T \mathbf{V}_i^{-1} \mathcal{D}_i & \mathbf{0} \\ \mathbf{C}_i & \sum_{m=1}^{M_i} \Lambda_{im} \mathbf{B}_h^T(t) \mathbf{B}_h(t) \end{bmatrix} - J_\lambda(\boldsymbol{\theta}, \boldsymbol{\beta}),$$

where the bottom left block $\mathbf{C}_i = -\partial\{\mathcal{D}_i^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathcal{D}_i \boldsymbol{\theta})\}/\partial \boldsymbol{\beta}$; $J_\lambda(\boldsymbol{\theta}, \boldsymbol{\beta})$ is the combined penalty matrix of $J_\lambda(\boldsymbol{\theta})$ and $\mathcal{J}_{\lambda'}(\boldsymbol{\beta})$. The sandwich variance estimator can properly account for the variability of estimated survival distribution.

We studied the asymptotic properties of the proposed penalized spline estimator in the Appendix B. We showed that, similar to the univariate penalized splines (Chen and Wang, 2011; Chen et al., 2013), the asymptotic properties for the bivariate smoothing in the presence of right-censoring resembles the small knots scenario in Zhu et al. (2008). The rate of convergence of the asymptotic bias, variance and mean squared error (MSE) are shown in the Appendix B. We show that the asymptotic variance is minimized when the true covariance is used (similar to Welsh et al. (2002)), when

the penalty parameter converges to infinity at a certain rate. The asymptotic normality of the proposed estimator is also proved.

3.3.6. Computation Algorithm

An algorithm for implementing the proposed method through an iterative alternating optimization procedure is described in Algorithm 1. The two-step algorithm consists

Algorithm 1 Estimate the mean medical cost trajectory

- 1: **Inputs:**
Claims data $\{\mathbf{Y}_i, T_i, \delta_i; i = 1, \dots, n\}$
 - 2: **Initialize:**
 $\tilde{\boldsymbol{\theta}}^{(0)}$ by Equation (3.4) where \mathbf{V}_i is identity matrix, and set design matrix as \mathbf{D}_i
 - 3: **while** convergence criteria for $\tilde{\boldsymbol{\theta}}$ is not satisfied **do**
 - 4: Update $\phi(\cdot)$ and $\boldsymbol{\alpha}$ following the steps given in Section 3.3.4
 - 5: Estimate \mathbf{V}_i and update $\tilde{\boldsymbol{\theta}}$ by Equation (3.4)
 - 6: **end while**
 - 7: **Reinitialize:**
 $\hat{\boldsymbol{\theta}}^{(0)} = \tilde{\boldsymbol{\theta}}$, and estimate $\boldsymbol{\beta}$ to set the new design matrix as \mathcal{D}_i
 - 8: **while** converge criteria for $\hat{\boldsymbol{\theta}}$ is not satisfied **do**
 - 9: Update $\phi(\cdot)$ and $\boldsymbol{\alpha}$ following the steps given in Section 3.3.4
 - 10: Estimate \mathbf{V}_i , and update $\hat{\boldsymbol{\theta}}$ by Equation (3.8)
 - 11: **end while**
-

of estimation for $\tilde{\boldsymbol{\theta}}$ and for $\hat{\boldsymbol{\theta}}$ sequentially. In line 2 we start with a consistent but inefficient initial guess $\tilde{\boldsymbol{\theta}}^{(0)}$ assuming constant variance; the estimation is based on data from uncensored subjects and those censored at τ . Cycling between lines 4 and 5 until convergence leads to $\tilde{\boldsymbol{\theta}}$. In line 7 the estimation of the survival parameter $\boldsymbol{\beta}$ leads to a design matrix that includes subjects who were censored prior to τ . Iterating through lines 9 and 10 until convergence will lead to the final estimator $\hat{\boldsymbol{\theta}}$.

3.3.7. Knots and smoothing parameter selection

In a penalized spline approach, the trade-off between efficiency and bias is controlled by a choice of lavish number of knots and corresponding penalty parameters. We

follow the guideline in Eilers and Marx (2010) to use $k = \min\{\kappa, \max\{n_i\}/4\}$ equally spaced knots on each dimension of the bivariate surface. For example, if $k = 5$, there are 30 knots involved in cost trajectory estimation. Our experience shows that an excessively large number of knots beyond this guideline does not guarantee improvement on the quality of estimation, but substantially increases the computation burden. This coincides with the previous theoretical work (Li and Ruppert, 2008). As pointed out by Eilers and Marx (2002), the degree of the polynomial to construct the B-spline basis function is less important in the case of penalized spline models, where a relatively large number of knots are used. Hence, we use quadratic penalized splines throughout this chapter.

The penalty parameter $\boldsymbol{\lambda}_\theta$ can be selected via a grid search to minimize the QGCV criteria (Fu, 2003):

$$\text{QGCV}(\boldsymbol{\lambda}_\theta; \boldsymbol{\theta}) = \frac{n^{-1} \sum_{i=1}^n \mathbf{r}_i^T \mathbf{R}_i^{-1} \mathbf{r}_i}{\{1 - p(\boldsymbol{\lambda}_\theta; \boldsymbol{\theta})/N\}^2}. \quad (3.12)$$

In this expression, the numerator is the average of the weighted square of the deviance residuals \mathbf{r}_i corresponding to the n_i observations, with the weights being the inverse of the working correlation matrix \mathbf{R}_i . In the denominator, $p(\boldsymbol{\lambda}_\theta; \boldsymbol{\theta}) = p_\theta s_\theta$ is the effective number of parameters in the penalized model, where $p_\theta = (p_1 + K_1)(p_2 + K_2) + (p_3 + K_3)$ is the number of covariates, $s_\theta = \|\boldsymbol{\theta}(\boldsymbol{\lambda}_\theta)\|_2 / \|\boldsymbol{\theta}(\mathbf{0})\|_2$ is the shrinkage rate of the penalized GEE estimator over the non-penalized GEE estimator, and $N = \sum_{i=1}^n n_i^2 / |\mathbf{R}_i|$ is the effective degree of freedom of correlated observations.

3.4. Analyses of Prostate Cancer Cost Data from SEER-Medicare

We applied the proposed method to model the medical cost data of a cohort of prostate cancer patients from the SEER-Medicare linked database as described earlier. The goal was to estimate the population-averaged medical cost trajectory, stratified by

the cancer stage at the time of diagnosis (local/regional vs. distant). The study cohort consisted of 184,491 patients over age 65 who were diagnosed with prostate cancer between January 1, 2003 and December 31, 2015. The medical costs included Medicare payments for inpatient and outpatient services covered by Medicare Parts A and B, from the date of diagnosis to death or the end of follow-up (i.e. December 31, 2016). To ensure the completeness of cost data within the first 12 months of cancer diagnosis, we further restricted the study cohort to patients with continuous Parts A and B enrollment for this duration. The median survival time was 63 months. All patients who survived beyond τ were administratively censored at $\tau = 10$ years. Only 16.3% subjects died between τ and 180 months (the actual maximum follow-up time). Thus, the sparse survival data beyond τ may lead to an identifiability issue if a longer τ is selected. Due to loss of follow-up prior to τ , 65.1% subjects are right-censored. There were 172,166 (93.3%) subjects with local or regional stage cancer at initial diagnosis, and 12,325 (6.7%) with distant stage cancer; the overall censoring rates were 78.0% and 25.8%, respectively. A descriptive analysis of the cost data is visualized in Figure 3.1, which suggests an association between the shape of the cost trajectories and survival, while illustrating several data features including skewness, zero-inflation, heteroscedasticity, censoring, and the presence of a LTS group.

We specified a quadratic B-spline basis for the mean cost surface $\mu(t, s)$ with ten equally spaced knots for survival time s and measurement time t , which resulted in a total of 156 basis functions. The intra-subject correlation of the monthly costs is modeled using a working correlation matrix of compound symmetry structure. Motivated by the heteroscedasticity observed in Figure 3.1, we used a flexible nonparametric function to model the mean-variance relationship in the GEE. As explained previously, this relationship is specified for LTS and non-LTS patients separately.

Figure 3.2 illustrates the estimated monthly cost trajectories corresponding to 2,

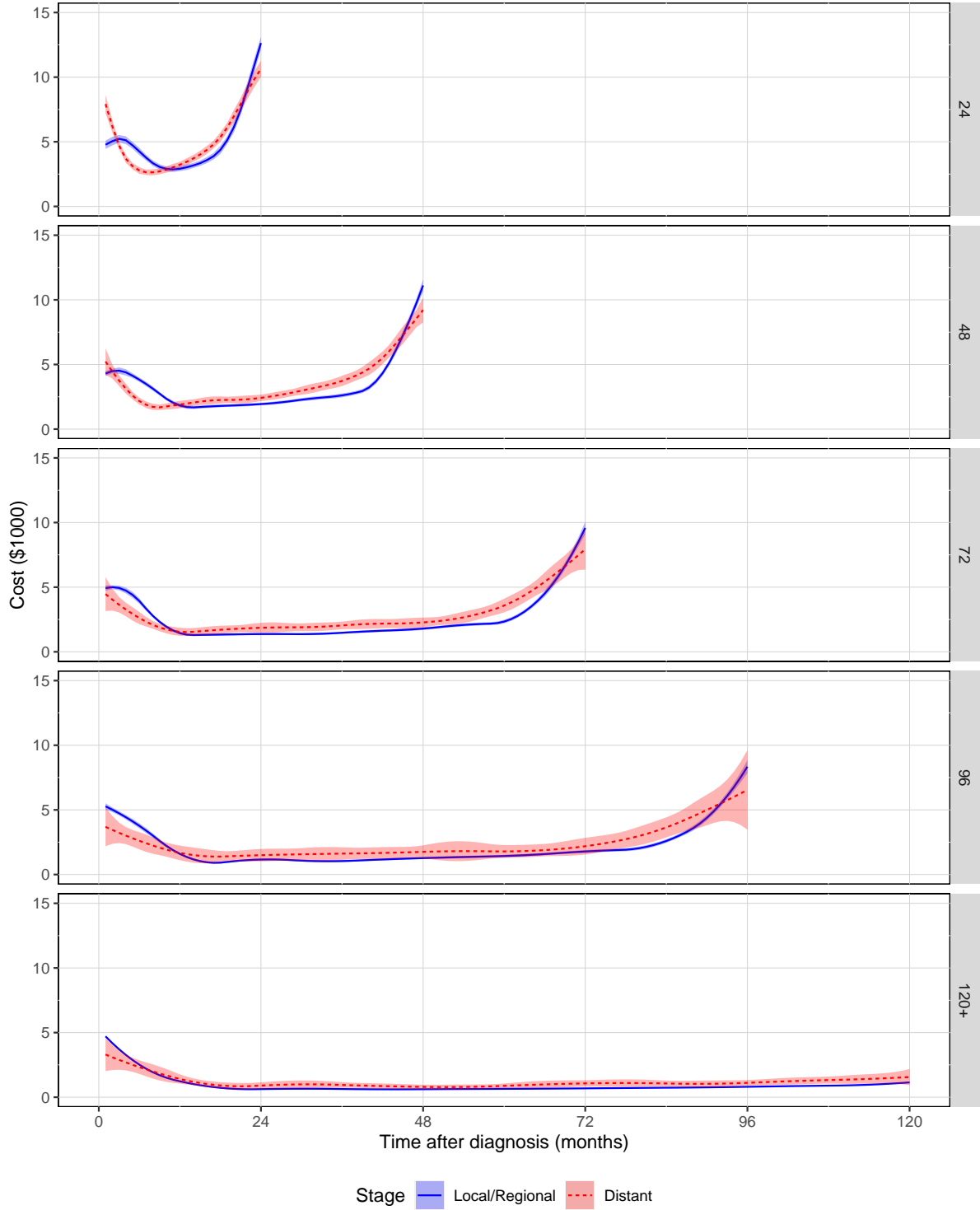


Figure 3.2: Prostate cancer care cost analyses for the estimated monthly cost trajectories $\hat{\mu}(t, s)$ when the survival time (month) equals $s = 24, 48, 72, 96$ and $120+$ (LTS). The results are compared between local/regional stage (blue) and distant stage (red). The shaded areas are 95% confidence intervals.

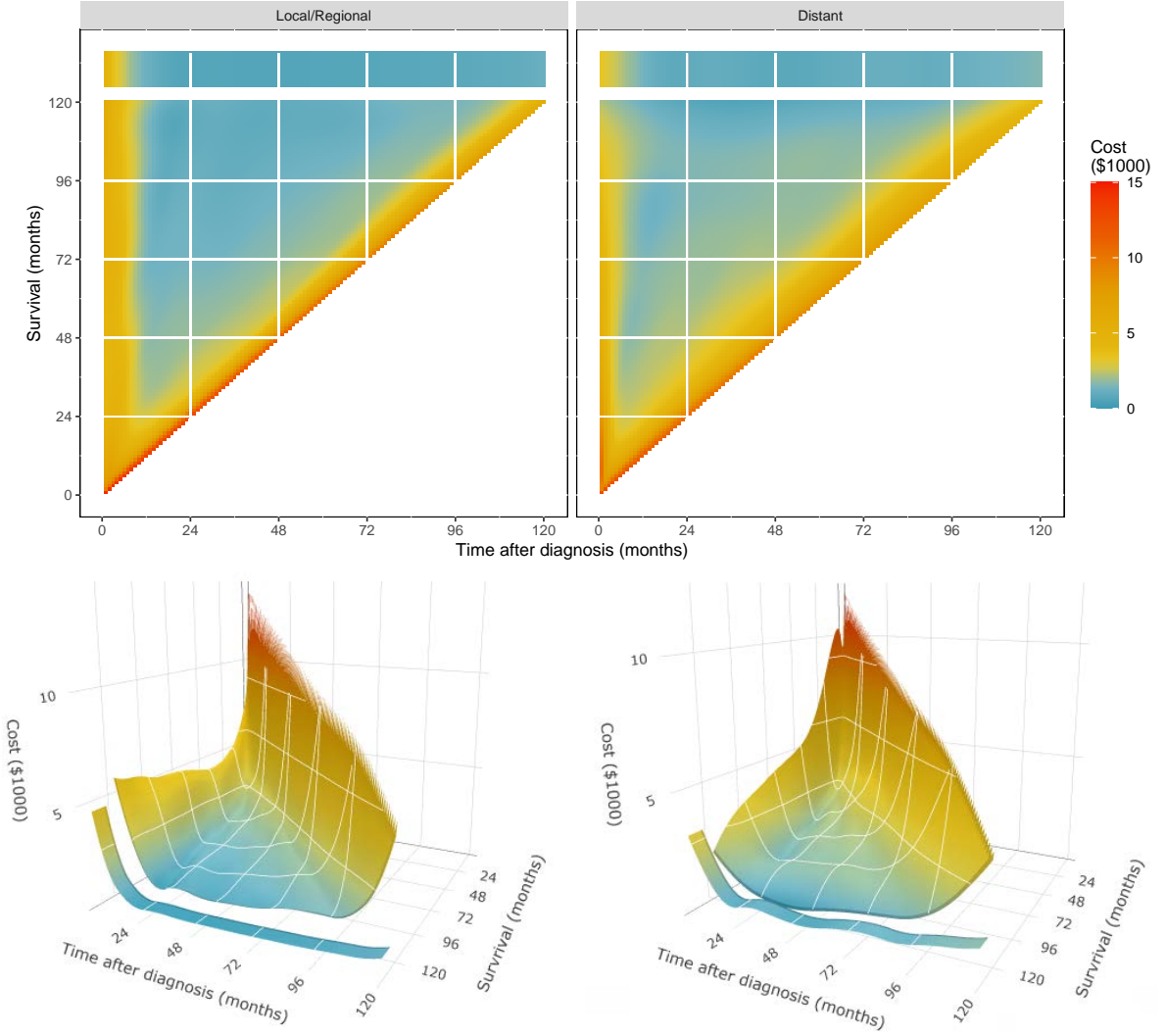


Figure 3.3: Prostate cancer care cost analyses for the 2D heatmaps and 3D surfaces. $\hat{\mu}(t, s)$ in 2D heatmaps and 3D surfaces. The results are compared between local/regional stage (left) and distant stage (right). The estimates of the LTS group are displayed separately in each plot. White contour curves are shown on the 3D plot at \$3000, \$6000 and \$9000 elevation.

4, 6 and 8 years of survival and the LTS group (> 10 years of survival). Pointwise 95% confidence intervals were obtained by the proposed sandwich variance estimator. The cost trajectories of all survival groups are U-shaped, except for LTS (L-shaped), which agrees with the NCI Phase-of-Care approach. This is because patients tended to receive active treatments right after the initial cancer diagnosis or various intensive care near the end of life. However, there are some important data features that are not captured by the Phase-of-Care approach. First, it is not straightforward to define or justify 12 months prior to death as the cut-off point that separates the terminal care phase from the continuing care phase. This is most clearly shown among the distant stage patients, where the monthly costs increase steadily beyond 12 months. Second, the monthly costs for local/regional stage patients elevate steeply around one year before death, but the elevation is less obvious among distant stage patients. For example, the end-of-life medical costs increased from \$3,000 up to \$12,000 per month for local/regional stage patients who survived less than four years as they approached end-of-life care; the increment is smaller for the distant stage patients, from \$3,000 up to \$10,000. For patients who died within a relatively short time period after the diagnosis, the cost trajectories mostly reflected intensive care and treatment in the initial and terminal care phases, with some possible overlap. However, the difference in the end-of-life cost between the two cancer stages reduces as survival time increases. In the continuing care phase the monthly medical costs are between \$1000 and \$3000, and the distant stage patients on average had higher care cost than local/regional cancer patients.

The cost trajectory is characterized by a bivariate surface $\mu(t, s)$, with s being the survival time and t being the month of the cost. Since $t \leq s \leq \tau$, this surface is defined on a triangular area. The cost trajectory of the LTS group is characterized as a separate curve. The estimated surfaces from the local/regional stage cancer and the

distant stage cancer are presented in Figure 3 in both 2D heatmaps and 3D surfaces. The surfaces are clearly highly nonlinear, and they demonstrate complicated interaction between the cost trajectory and survival. There also is a notable difference in cost by cancer stages. The contour curves at costs equal to \$3000, \$6000, and \$9000 are depicted on the surface plots. Interestingly, we found that as survival time increases, the costs at the first month after distant stage cancer diagnosis drops, while subjects in earlier stages who live longer tend to have higher medical costs at the first month. The LTS group does not have a visible elevated cost at the end, because survival time and cost data are both censored at τ (i.e. their cost trajectory is L-shaped). These cost trajectory details cannot be obtained from previous methodologies, such as the NCI Phase-of-Care approach. We report a sensitivity analysis in the B, which compares results from $\tilde{\mu}(t, s)$ and $\hat{\mu}(t, s)$ at $s = 2, 4, 6, 8$ years and with stratification by cancer stage. The results are generally close to each other, as they are both consistent estimators under independent censoring.

3.5. Simulations

In this section we study the finite sample numerical performance of the proposed estimator $\hat{\mu}$ denoted by GEE2-Y (assuming a flexible variance function using all available data), including bias, efficiency, confidence interval coverage, and convergence rate in small to moderate sample sizes. While our motivating dataset has a large sample size, smaller studies are common when inclusion/exclusion criteria are applied to select the target population, or when the research interest focuses on certain subgroups or strata. We compare the proposed method to the estimator from Section 3.3.1: GEE1-Y (assuming a constant variance using all available data) and GEE1-N (assuming a constant variance, but not using data from subjects censored prior to τ); in terms of efficiency. We compare the proposed method to a previously published cost trajectory estimation

method (Li et al., 2018), which is based on a normality distribution assumption for the transformed cost data, to demonstrate robustness against skewness, zero-inflation, heteroscedasticity and misspecification of the data distribution. To study the benefit of modeling the variance function, we evaluate the proposed method with and without variance modeling.

3.5.1. Data generation and analysis

Time to death \tilde{T} was generated from a Weibull distribution with shape = 2 and scale = 1. Censoring time C is generated independently from Uniform (0, 1.1), yielding a 65% censoring rate and a 5% censoring at $\tau = 1$. As shown in Figure 3.4, the mean incidence costs are generated at time points 0.01, 0.02, ... to mimic the U-shaped trajectories for the non-LTS subjects and an L-shaped trajectory for LTS. For subjects with $\tilde{T} < \tau$, the incident cost at time t follows a gamma distribution with chosen shape and scale parameters such that the variance function is $\text{var}(Y) = \mu^2$. In addition, we simulate a compound symmetry correlation structure with $\alpha = 0.2$ by using the R package `simstudy` (Goldfeld and Wujciak-Jens, 2020). The monthly cost for LTS follows a gamma distribution of less variation (i.e., $\text{var}(Y) = 0.8\mu^2$) and intra-subject correlation is from the compound symmetry correlation structure with $\alpha = 0.05$. For zero-inflated gamma ($z\Gamma$) setting, 50% of responses are randomly set to 0.

For model fitting, we used a quadratic B-spline basis with five equally spaced knots in the mean and variance functions of costs and the hazard function for survival. The convergence criteria for the GEE iteration is that for all the parameters, the absolute difference between the current estimate and previous estimate should be less than 10^{-4} . Since the initial values of the parameters are consistent estimators, the algorithm usually converged within 20 iterations. According to the proposed GCV-based criteria, we selected the penalty parameter λ to be 10^{-5} through a grid search on five simulated

datasets. The sample sizes were $n = 500$ or 1000 , and 1000 Monte Carlo repetitions were used in each simulation setting. All studies were run on a computer with an Intel Core i5-6500 3.20 GHz CPU and 16 GB memory.

3.5.2. Simulation Results

Tables 3.1 and 3.2 show the pointwise absolute error (AE), MSE, and the coverage probability (CP) of a 95% confidence interval at selected t and s values on the estimated cost trajectory surface. The results are compared among GEE2-Y, the proposed estimator, and GEE1-Y and GEE1-N, the reduced estimators from Section 3.3.1. All estimators show little bias, decreased MSE with increased sample size, and coverage probabilities close to 95%. The efficiency gain from the use of censored data GEE1-Y compared to GEE1-N for the proposed estimator is clearly seen in all scenarios. For example, under the Gamma distribution setting with $n=500$, when $(t, s)=(0.3, 0.6)$ the MSE for GEE1-Y reduces by 0.045 compared to the MSE for GEE1-N; when $(t, s)=(0.4, 0.8)$ the reduction increases towards 0.152. This suggests that using censored cost data is particularly beneficial for improving efficiency on mean cost trajectory estimation when survival time is large. This is because fewer observed incidence costs are available due to loss of follow-up. This result demonstrates that the efficiency can be improved substantially by properly using cost data from subjects censored prior to τ . The simulation was done when the cost data were skewed (Gamma) in Table 3.1 or skewed with zero-inflation in Table 3.2. Slightly larger MSEs of proposed estimation for zero-inflated gamma setting may be due to the increased variation. Consistent estimation of the cost trajectory is achieved without explicitly modeling the full distribution of the cost data. This makes the proposed method convenient to use.

Figure 3.4 is a graphical illustration of the estimation result of Table 3.1 in the form of 2D cost trajectories at selected survival times ($\tilde{T} = 0.4, 0.6, 0.8$). To quantify the

Table 3.1: Simulation results for the estimation of cost trajectory $\mu(t, s)$ at selected t and s . The cost data follow Gamma distributions. The sample size $n = 500$ or 1000 . Pointwise absolute error (AE), mean squared error (MSE), and coverage probability (CP) are based on aggregated results from 1000 Monte Carlo repetitions. GEE1-N the estimator assuming constant variance ignoring subjects censored prior to τ ; GEE1-Y the estimator assuming constant variance using all available data; and GEE2-Y the proposed estimator assuming flexible variance using all available data from Section 3.3.1.

		GEE1-N			GEE1-Y			GEE2-Y		
t	s	AE	MSE	CP	AE	MSE	CP	AE	MSE	CP
n=500										
0.10	0.4	0.042	0.321	0.954	0.076	0.285	0.970	0.037	0.277	0.964
0.20	0.4	0.225	0.281	0.944	0.252	0.266	0.951	0.151	0.220	0.950
0.30	0.4	0.034	0.315	0.960	0.060	0.285	0.970	0.001	0.283	0.959
0.15	0.6	0.027	0.193	0.945	0.052	0.154	0.945	0.012	0.159	0.953
0.30	0.6	0.150	0.159	0.922	0.166	0.143	0.919	0.128	0.131	0.933
0.45	0.6	0.035	0.189	0.953	0.059	0.162	0.947	0.044	0.179	0.949
0.20	0.8	0.078	0.254	0.957	0.029	0.156	0.941	0.086	0.182	0.945
0.40	0.8	0.098	0.232	0.964	0.059	0.148	0.949	0.045	0.181	0.941
0.60	0.8	0.075	0.254	0.945	0.058	0.173	0.973	0.143	0.227	0.934
n=1000										
0.10	0.4	0.026	0.155	0.962	0.063	0.143	0.965	0.026	0.139	0.964
0.20	0.4	0.210	0.159	0.932	0.240	0.160	0.932	0.138	0.122	0.942
0.30	0.4	0.029	0.156	0.959	0.060	0.144	0.956	0.007	0.141	0.956
0.15	0.6	0.027	0.107	0.940	0.052	0.087	0.944	0.012	0.084	0.950
0.30	0.6	0.142	0.095	0.918	0.164	0.086	0.976	0.123	0.076	0.922
0.45	0.6	0.025	0.101	0.948	0.049	0.082	0.948	0.037	0.083	0.964
0.20	0.8	0.116	0.141	0.950	0.000	0.076	0.971	0.080	0.084	0.957
0.40	0.8	0.121	0.123	0.969	0.074	0.078	0.936	0.057	0.084	0.963
0.60	0.8	0.115	0.141	0.962	0.031	0.083	0.976	0.146	0.116	0.932

Table 3.2: Simulation results for the estimation of cost trajectory $\mu(t, s)$ at selected t and s . The cost data follow zero-inflated Gamma distribution. The sample size $n = 500$ or 1000. Pointwise absolute error (AE), mean squared error (MSE), and coverage probability (CP) are based on aggregated results from 1000 Monte Carlo repetitions. GEE1-N the estimator assuming constant variance ignoring subjects censored prior to τ ; GEE1-Y the estimator assuming constant variance using all available data; and GEE2-Y the proposed estimator assuming flexible variance using all available data from Section 3.3.1.

		GEE1-N			GEE1-Y			GEE2-Y		
t	s	AE	MSE	CP	AE	MSE	CP	AE	MSE	CP
n=500										
0.10	0.4	0.019	0.620	0.959	0.050	0.521	0.977	0.018	0.511	0.961
0.20	0.4	0.257	0.538	0.961	0.294	0.479	0.966	0.200	0.433	0.954
0.30	0.4	0.027	0.614	0.961	0.048	0.547	0.969	0.004	0.522	0.966
0.15	0.6	0.019	0.335	0.945	0.053	0.263	0.959	0.031	0.256	0.955
0.30	0.6	0.160	0.278	0.931	0.175	0.233	0.945	0.150	0.223	0.943
0.45	0.6	0.027	0.316	0.954	0.054	0.271	0.961	0.065	0.266	0.957
0.20	0.8	0.082	0.401	0.945	0.024	0.252	0.917	0.064	0.266	0.952
0.40	0.8	0.104	0.399	0.949	0.059	0.247	0.882	0.024	0.265	0.947
0.60	0.8	0.073	0.402	0.960	0.041	0.274	0.912	0.115	0.305	0.947
n=1000										
0.10	0.4	0.037	0.283	0.974	0.063	0.246	0.976	0.037	0.234	0.975
0.20	0.4	0.245	0.292	0.957	0.279	0.278	0.956	0.185	0.235	0.962
0.30	0.4	0.036	0.309	0.974	0.059	0.277	0.977	0.015	0.264	0.974
0.15	0.6	0.014	0.166	0.955	0.050	0.134	0.962	0.022	0.128	0.960
0.30	0.6	0.173	0.162	0.918	0.193	0.141	0.919	0.156	0.127	0.929
0.45	0.6	0.011	0.160	0.961	0.038	0.133	0.968	0.039	0.126	0.967
0.20	0.8	0.114	0.213	0.960	0.006	0.122	0.941	0.044	0.124	0.956
0.40	0.8	0.118	0.192	0.967	0.067	0.115	0.944	0.029	0.118	0.969
0.60	0.8	0.114	0.220	0.951	0.014	0.138	0.936	0.100	0.148	0.953

global performance of our proposed estimator, AE and MSE are further averaged over all the time points as shown in the legend of the top panel in Figure 3.4. We find that the proposed estimator fits the mean cost surfaces well under all scenarios. The EM method (Li et al., 2018) shows remarkable biases especially when survival time is large, possibly due to misspecifying the distribution assumption and failure to properly account for LTS. Global performance in the legend shows that even without using censored cost data, our estimator GEE1-N outperforms the EM method for both accuracy and efficiency. The proposed GEE2-Y estimate further reduces the mean average MSE by around 50% compared to the EM method for all scenarios, suggesting its superior efficiency and robust finite sample properties. Under a larger sample size the improvement becomes more significant. The above observations likely all arise from misspecifying the distributional assumption. Therefore, these results highlight the usefulness of the proposed method, which flexibly models the trajectory (the mean function) without distribution assumptions on the data.

The bottom panel in Figure 3.4 depicts that the averaged estimated survival probability (red dashed curve) is close to the truth (black solid curve) with a relatively small sample size. In the second and third panel we simulate time to death from exponential and Gamma distributions respectively. The estimated survival probability shows little bias. Additional simulations that evaluate the proposed method under various settings with different censoring rates, proportions of zeros, and levels of skewness are presented in Appendix B.

3.6. Discussion

In this chapter, we propose a novel marginal model and inference procedure for the mean medical cost trajectory in the presence of right-censoring. It uses a flexible semi-parametric function to model the nonlinear cost trajectory, and the estimation does not

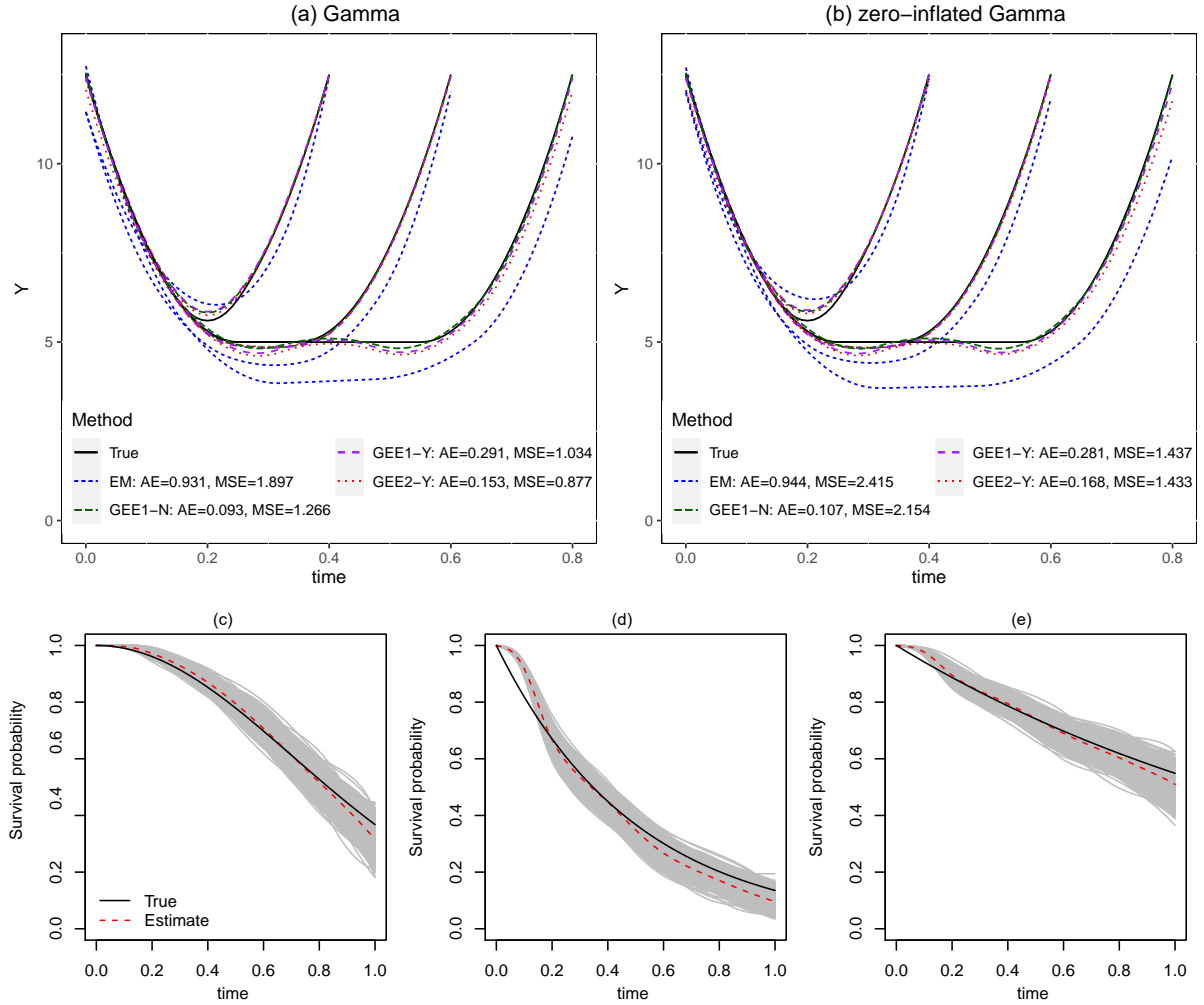


Figure 3.4: Simulation results for mean cost trajectories given survival time $\mu(t, s)$ at $s = 0.4, 0.6, 0.8$ based on (a) Gamma and (b) zero-inflated Gamma cost data; and (c, d, e) the penalized spline estimation of the survival function. Sample size $n=500$. The true trajectories are plotted in black curves; on the top panel the mean estimated trajectories are plotted in blue dotted (EM), green dashed (GEE1-N), purple long dashed (GEE1-Y) and red lined (GEE2-Y) curves. Mean average absolute error (AE) and mean squared error (MSE) for $\mu(t, s)$ are averaged across all observations within each dataset, and then over 1000 simulation replicates. On the bottom panel the solid black curve is the true survival function. The red dashed curve is the estimated survival function, averaged overall all 1000 Monte Carlo repetitions. The gray curves are all estimators.

rely on a distribution assumption of the cost data. This feature is particularly attractive because the cost data are often skewed with zero-inflation, heteroscedasticity, and intra-subject correlation, which makes them difficult to model with parametric distributions. While a consistent estimator can be obtained by discarding the cost data from censored subjects, the proposed method can effectively use cost data from censored subjects to improve efficiency. The proposed method can also accommodate an LTS group with distinct mean and variance parameters.

To our knowledge, this is the first time that GEE-based methods have been extended to deal with a censored covariate. This extension is motivated by our data application, where the mean function for the cost trajectory depends on survival, which is subject to censoring. The sandwich variance estimator is robust against misspecification of the covariance matrix. This is because the induced mean function is unbiased for the outcome regardless of the distribution of cost data. Our study on the asymptotic properties (Online Appendix) shows that our method has similar properties as the regression spline, and the estimated mean function is efficient when the working covariance matrix is correctly specified. Our simulation studies show that the proposed method estimates the cost trajectories well under moderate sample size and censoring rate, regardless of the data distribution. Efficiency improvement is clearly gained from GEE2 compared to GEE1 for mean cost trajectory estimation. Therefore, we recommend modeling the variance function and adjusting for heteroscedasticity.

This chapter focuses on estimating the mean cost because it is directly related to the total population expenditure, which is of policy interest. Future work will be considered to extend the methodology to modeling median cost or other quantiles. This chapter does not study the effect of baseline covariates, i.e., how the baseline covariates or their linear combination change the shape of the cost trajectory, along with survival

\tilde{T} . This is because our goal is to estimate the average costs in a defined population. When the sample size permits, the effect of categorical baseline covariates can be studied by analyzing the cost trajectories within each subgroup defined by the categories, as is done in our analysis of local/regional vs. distant stage cancer. A general methodology to model the relationship between multiple baseline covariates and their effects on both the varying length and nonlinear shape of the cost trajectories is beyond the scope of this chapter and is explored in Chapter 4.

CHAPTER 4

LONGITUDINAL VARYING COEFFICIENT SINGLE-INDEX MODEL WITH CENSORED COVARIATES*

4.1. Introduction

In recent years, the longitudinal medical cost data have received extensive attention in health service research and health policy making (Mariotto et al., 2020; Yamin, 2020). Its development originates from the fact that the advancing data collection and management techniques allow medical cost data to be routinely recorded by hospitals and insurance companies. In the reports from National Cancer Institute (NCI) for the estimation and projection of cancer care costs, the time from cancer diagnosis to death is divided into three phases of care, including initial, continuing and terminal phases (Mariotto et al., 2011). They reported summary statistics by averaging the total costs in different phases, regardless of the duration of the medical cost trajectory.

The longitudinal medical cost trajectory, defined as the averaged incident medical cost (by month, quarter or year) from diagnosis to a terminal event such as death (Li et al., 2018), provides more holistic picture on time-dependent healthcare consumption than the phase-based estimators. It can adjust for the time-span of care phases with respect to survival, which is subject to informative dropout (Little and Rubin, 2019). The growing cancer patients population and emerging new treatments are triggering fast evolution of both demand and supply in the health care field. Patient characteristics, such as race, age and initial treatment, may have different effect on the longitudinal medical cost at different time during patient's lifespan. For example, in the motivating

*A version of Chapter 4 has been submitted to the Journal of the American Statistical Association.

case study of this chapter, the data includes the quarterly medical costs of a cohort of patients with prostate cancer, from their initial diagnosis until death. Figure 4.1 shows that among patients who died at 16 quarters after the diagnosis of localized stage prostate cancer, Black had higher costs than white.

The goal of this work is to develop a parsimonious, flexible, and interpretable regression model to study how covariates or subgroups affect the longitudinal medical cost trajectory. It is of central interest to policy makers because it not only reveals the association between the patients' baseline characteristics and medical costs across phases of care, but also provides the building block of cancer care cost projection, which is of direct relevance to healthcare decision-making. To the best of our knowledge, we are among the first to study the relationship between baseline covariates and the longitudinal medical cost trajectory given survival time using a regression model.

Recent work on the estimation of longitudinal medical cost trajectory conditional on survival (Li et al., 2018; Wang et al., 2020) does not model the covariate effect on cost trajectory. For example, we may want to compare medical cost among races with adjustment to age group at baseline; or to compare longitudinal cost differences among patients receiving different first-line treatments.

In this chapter, there are four novel aspects in the regression model that estimates the longitudinal medical cost trajectory flexibly, along with other baseline characteristics parsimoniously, while properly accounting for the censored terminal event time. First, the longitudinal medical cost data is modelled without distribution assumption, and potentially non-normality in the cost data is properly handled, including skewness, zero-inflation, and heteroscedasticity. Second, patient characteristics are collapsed into a single-index to represent the propensity of healthcare utilization, which graphically

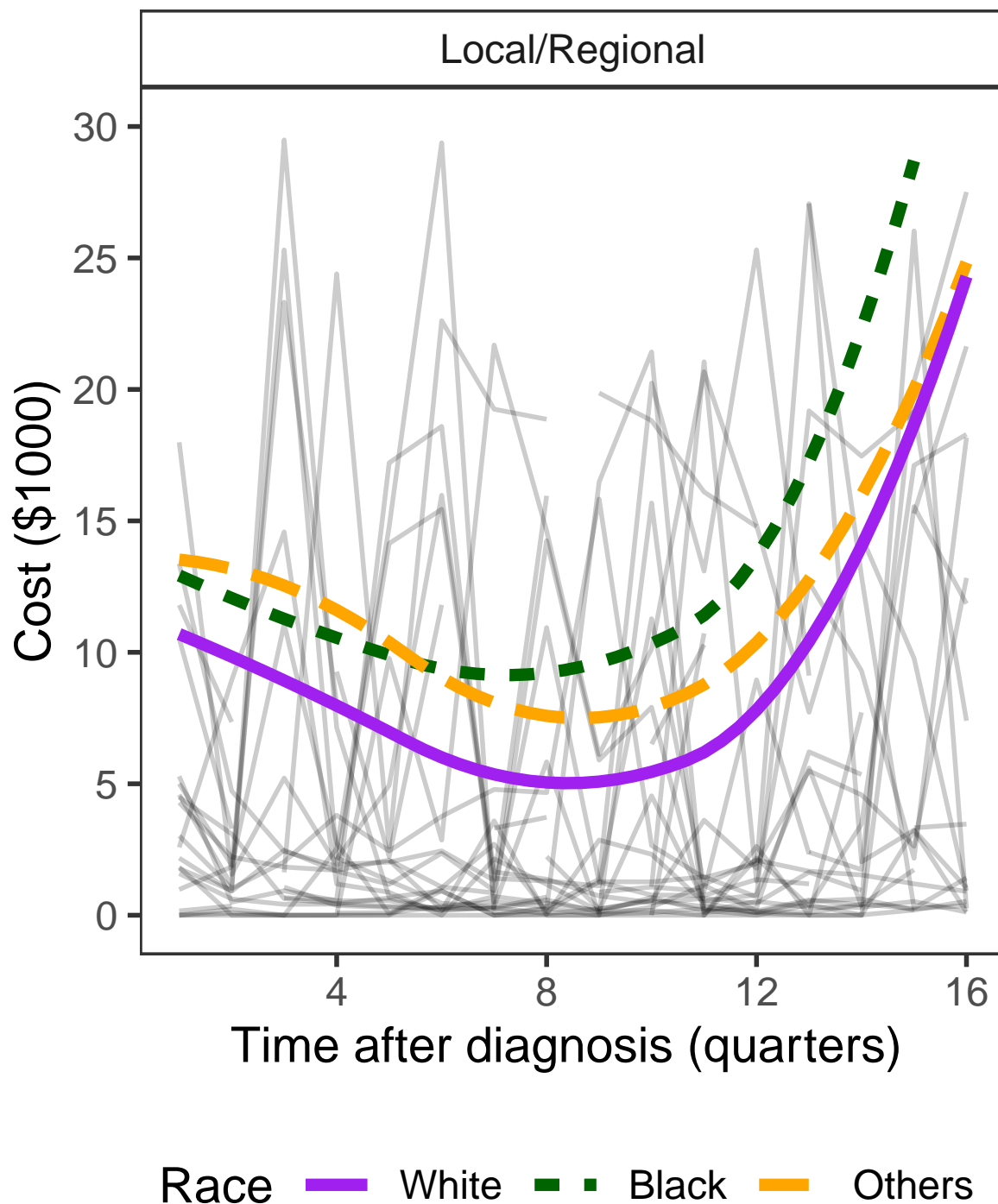


Figure 4.1: Plots of the estimated mean curves of quarterly medical cost by quadratic splines with five equally spaced knots for the SEER-Medicare prostate cancer cohort by three racial groups: white (thick purple line), Black (dotted green line), and others (dashed orange line). A random sample of 50 uncensored subjects was selected and their individual quarterly cost data are plotted in the background as gray lines. Death time is 16 quarters after cancer diagnosis.

quantifies the heterogeneous effects given various levels of patient characteristics. The uncertainty associated with covariates and cost trajectories are properly modeled. Third, the complex relationship between cost trajectory and survival time is modelled by a flexible bivariate varying coefficient surface. Fourth, we appositely cope with the censored covariate (i.e., survival time) to ensure the identifiability of the conditional mean of cost given survival.

To achieve these goals, we propose a semiparametric marginal approach for estimating the longitudinal medical cost trajectory given survival, and conducting inference for the baseline covariates. The potential heterogeneous subgroup effect on the outcome cost trajectory through an unknown function of covariates can be assessed. The proposed method enjoys flexibility while achieving efficiency by utilizing cost data from both uncensored and censored patients through a novel extended generalized estimating equation (GEE) approach. It is motivated by the varying-coefficient model considered in Wu et al. (2019) for risk assessment with multiple covariates, which can handle large amounts of temporal data and baseline covariates. To utilize all the covariates effectively, our single-index term focus on a weighted linear combination of all covariates as a summary predictor. We model the cost trajectory curve as a function of the this summary predictor, the measurement time and survival with possible interaction terms. The coefficients (or weights) in the single-index term measure how important the covariate is for predicting the outcomes.

An iterative estimation procedure is developed for the proposed models. The theoretical results for the asymptotic properties allow the construction of pointwise confidence intervals for the cost trajectory curves. Based on the inference procedure, we can identify the important characteristics and the associated subgroup of patients that has different mean costs in different time period, and the global relationships between the

costs and baseline covariates.

The rest of this chapter is organized in five sections. We introduce the model notation in Section 4.2. Estimation and inference procedures are given in Section 4.3. Numerical results on a data example and simulation data are presented in Section 4.4 and 4.5 respectively. Proofs of main results and additional simulations are presented in Appendix C.

4.2. Model

We observe independent and identically distributed data from n subjects, $\{\mathbf{Y}_i, \mathbf{X}_i, T_i = \min(\tilde{T}_i, C_i), \delta_i = 1\{\tilde{T}_i \leq C_i\}; i = 1, \dots, n\}$. $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ is a vector of longitudinal outcome up to observed survival time T_i , $j = 1, \dots, n_i$. \mathbf{X}_i is the patient's baseline variables of length p . \tilde{T}_i is the time to terminal event of interest. C_i is the time to censoring. δ_i is the event indicator: $\delta_i = 1$ if terminal event occurs prior to censoring. τ can be the maximum observed event time (i.e., $\max\{T_i | \delta_i = 1\}, i = 1, \dots, n$), or alternatively a prespecified cutoff time of interest, such as ten years in the motivating case study. $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{in_i})^T$ is the measurement time and n_i is the number of longitudinal measurements. For convenience, we write $Y_{ij} = Y_i(t_{ij})$. We further assume that C_i and (\tilde{T}_i, Y_{ij}) are independent given covariates \mathbf{X}_i .

An intuitive approach to extending the cost trajectory estimation methods (Li et al., 2018; Wang et al., 2020) by incorporating covariates is to assume that the covariates have linearly additive effects:

$$g(E\{Y_i(t_{ij}) | \tilde{T}_i, \mathbf{X}_i\}) = \mu_{01}(t_{ij}, \tilde{T}_i) + \mathbf{X}_i^T \boldsymbol{\theta}_1, \quad 0 \leq t_{ij} \leq \tilde{T}_i \leq \tau. \quad (4.1)$$

In this partially linear model, $\mu_{01}(t, \tilde{T})$ is a bivariate function of baseline trajectory surface

(e.g., the surface corresponding to zero covariate values). Each of the p covariates has a constant coefficient, denoted by $\boldsymbol{\theta}_1 = (\theta_{11}, \dots, \theta_{1p})^T$. $g(\cdot)$ is the link function. Although this model has a simple form and straightforward interpretation, it assumes that the covariate increment is associated with a constant shift on the entire cost trajectory, regardless of the time t and trajectory length \tilde{T} . This is a very strong assumption and it often does not hold, as shown by the example in Figure 4.1.

To have a model with flexible and interpretable covariate effects on cost trajectory, we propose the following longitudinal varying coefficient single-index model:

$$g(E\{Y_i(t_{ij})|\tilde{T}_i, \mathbf{X}_i\}) = \mu_{01}(t_{ij}, \tilde{T}_i) + \mathbf{X}_i^T \boldsymbol{\theta}_1 \mu_{11}(t_{ij}, \tilde{T}_i), \quad 0 \leq t_{ij} \leq \tilde{T}_i \leq \tau. \quad (4.2)$$

In this model, we rescale the covariate effect by $\mu_{11}(t, \tilde{T})$, a bivariate varying coefficient function of measurement time and survival. We refer to $\mathbf{X}_i^T \boldsymbol{\theta}_1$ as the index variable, which is a linear combination of all baseline covariates.

The interpretation of the model deserves a short explanation. Here are two points:

1. The covariate effect $\mu_{11}(t, \tilde{T})$ of the index variable can be viewed as the bivariate nonlinear “slope” of the index $\mathbf{X}_i^T \boldsymbol{\theta}_1$ on the outcome, i.e., incident cost at time t given death at \tilde{T} . It depicts how the index propensity bend and twist the trajectory shape. Accordingly, the baseline bivariate function $\mu_{01}(t, \tilde{T})$ can be considered as the “intercept”, which represents the mean baseline cost trajectory.
2. To make the single-index parameter $\boldsymbol{\theta}_1$ comparable, we consider two standardizations. First, we normalize the binary and categorical baseline covariates by dummy variables (i.e., coded as 0 or 1). We further standardize every continuous baseline covariates with mean 0 and variance 1, so that the coefficients can be used to com-

pare the relative influence. If the covariate effect $\mu_{11}(t, \tilde{T}) \equiv 0$, the covariates do not have any effect on the outcome. If $\mu_{11}(t, \tilde{T})$ is positive, the direction of $\boldsymbol{\theta}_1$ takes on the general meaning as effect for linear additive models, and the magnitude of $\boldsymbol{\theta}_1$ measures the contribution of covariates to the outcome; if $\mu_{11}(t, \tilde{T})$ is negative, the interpretation for the direction of $\boldsymbol{\theta}_1$ is flipped.

Since the covariate \tilde{T}_i in model (4.2) is subject to censoring, it is necessary to model the conditional distribution $F_{\tilde{T}}(s|T, \mathbf{X})$ of its residual distribution. For illustration purpose, we use the following Cox proportional hazard model, though any other survival regression model with time-independent covariates can also be used:

$$h(s|\mathbf{X}) = h_0(s) + \mathbf{X}^T \boldsymbol{\beta}_1, 0 < s \leq \tau. \quad (4.3)$$

For convenience in joint estimation with longitudinal data, we model the log baseline hazard as a penalized spline (Kauermann, 2005). Denote the parameters in the baseline hazard as $\boldsymbol{\beta}_0$. Then $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$ is the parameter vector in model (4.3). If the support of \tilde{T} lies outside the support of C , \tilde{T} is non-identifiable at (τ, ∞) without further assumptions.

In practice, the limited follow-up of a study and/or longer life expectancy for some individuals could result in a significant proportion of subjects to be censored at the maximum follow up time (τ) . For convenience, we call subjects who survived beyond τ long-term survivors (LTS). Since their terminal event is non-identifiable, our proposed method involves two parts of coefficient weights for the additive unknown functions for LTS and non-LTS.

Therefore, we further specify a longitudinal varying coefficient single-index model

for subjects with $\tilde{T} > \tau$, i.e., long-term survivors (LTS):

$$g(E\{Y_i(t_{ij})|\tilde{T}_i > \tau, \mathbf{X}_i\}) = \mu_{02}(t_{ij}) + \mathbf{X}_i^T \boldsymbol{\theta}_2 \mu_{12}(t_{ij}), \quad 0 \leq t_{ij} \leq \tau. \quad (4.4)$$

This model is similar to (4.2) but the varying coefficients do not depend on \tilde{T} . The interpretation can be easily extended. Obviously, the index parameter vectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ are not identifiable without further scale restrictions. Without loss of generality, we let $\boldsymbol{\theta}$ belong to the parameter space $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k : \|\boldsymbol{\theta}_k\| = 1, \theta_{k1} > 0, \boldsymbol{\theta}_k \in R^p, k = 1, 2\}$, where $\|\cdot\|$ denotes the l_2 norm of a vector. In other words, $\boldsymbol{\theta}$ is standardized to have unit norm. Under this restriction, $\boldsymbol{\theta}$ in model (4.2) and (4.4) are identifiable if μ 's are identifiable using spline approximation as we will discuss in Section 4.3. If we eliminate the first component in $\boldsymbol{\theta}_k$ ($k = 1, 2$), the resulting parameter space is denoted by $\boldsymbol{\Theta}_{-1} = \{\boldsymbol{\theta}_{k,-1} = (\theta_{k2}, \dots, \theta_{kp})^T : \sum_{j>1} \theta_{kj}^2 < 1, k = 1, 2\}$, and $\theta_{k1} = \sqrt{1 - \sum_{j>1} \theta_{kj}^2}$.

4.3. Estimation

In this section, we discuss the estimation of the proposed two-part model (4.2) and (4.4). We first present a simple estimator by using data from subjects who experienced the terminal event prior to τ and subjects who were censored at τ . Then we propose a novel extension of GEE approach that uses data from subjects who are censored prior to τ . Next, we incorporate variance-covariance modeling to the GEE to account for heteroscedasticity in the data. Lastly, we provide remarks on technical issues in the model fitting, including selection of knots and smoothing parameters, survival distribution estimation, and hypothesis testing on the coefficients.

4.3.1. No right-censoring prior to maximum follow-up time

We start with estimation using data from subjects who were not censored prior to τ . Under independent censoring assumption, the estimator is consistent (Wang et al.,

2020). We consider the two-part estimating equations for the medical cost trajectory from patient i as

$$E\{Y_{ij}|T_i, \delta_i, \mathbf{X}_i\} = \begin{cases} E\{Y_{ij}|\tilde{T}_i = T_i, \mathbf{X}_i\}, & \delta_i = 1 \text{ and } T_i \leq \tau, \\ E\{Y_{ij}|\tilde{T}_i > \tau, \mathbf{X}_i\}, & \delta_i = 0 \text{ and } T_i = \tau, \end{cases} \quad (4.5)$$

$$\text{cov}\{\mathbf{Y}_i|T_i, \delta_i, \mathbf{X}_i\} = \mathbf{V}_i,$$

where \mathbf{V}_i is the working covariance matrix that may be misspecified. The functions $\mu_{01}(\cdot, \cdot), \mu_{11}(\cdot, \cdot), \mu_{02}(\cdot), \mu_{12}, \mu_{02}(\cdot)$ are unspecified and can be approximated using spline functions with truncated polynomial basis. Since the measurement time cannot exceed the survival time, the bivariate cost trajectory function $\mu_{01}(\cdot, \cdot), \mu_{11}(\cdot, \cdot)$ is defined on an upper triangular area ($0 < t \leq \tilde{T}$). Following Li et al. (2018), we exploit a shrinkage-expansion transformation to define $\mu_{k1}(t, s) = \bar{\mu}_{k1}(u = t(\tau/s), s), 0 < t \leq s$. The expanded surface $\bar{\mu}_{k1}(u, s)$ is defined on the rectangular area $u, s \in (0, \tau]$, where conventional polynomial splines for a bivariate surface are defined. As a result, the non-parametric functions can be approximated well by spline functions such that $\mu_{k1}(t, s) \approx \mathbf{B}_{\mu_{k1}}(t, s)\boldsymbol{\gamma}_{k1}, k = 0, 1$, where $\mathbf{B}_{\mu_{k1}}(t, s)$ is a vector of bivariate truncated polynomial basis. Each basis function consists of elements of the outer product matrix of vector $B_t(t)$ and $B_s(s)$ with prespecified internal knots on the scale of t and s respectively. Similarly, for LTS the univariate cost trajectory function $\mu_{k2}(t) \approx \mathbf{B}_{\mu_{k2}}(t)\boldsymbol{\gamma}_{k2}, k = 0, 1$, where $\mathbf{B}_{\mu_{k2}}(t)$ is a vector of truncated polynomial basis with some prespecified internal knots. Recall the unknown index parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. Let the spline coefficients $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T$ where $\boldsymbol{\gamma}_1 = (\boldsymbol{\gamma}_{01}^T, \boldsymbol{\gamma}_{11}^T)^T$ and $\boldsymbol{\gamma}_2 = (\boldsymbol{\gamma}_{02}^T, \boldsymbol{\gamma}_{12}^T)^T$. The estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are the solutions of equation (4.5) where the j th element of $E\{\mathbf{Y}_i|T_i, \delta_i, \mathbf{X}_i\} =: \boldsymbol{\eta}_i$ is

approximated by

$$\eta_{ij} \approx g^{-1} \left(\delta_i \mathbf{B}_1(t_{ij}, T_i, \mathbf{X}_i^T \boldsymbol{\theta}_1) \boldsymbol{\gamma}_1 + (1 - \delta_i) 1(T_i = \tau) \mathbf{B}_2(t_{ij}, \mathbf{X}_i^T \boldsymbol{\theta}_2) \boldsymbol{\gamma}_2 \right),$$

where $\mathbf{B}_1(t, s, z) = (\mathbf{B}_{\mu_{01}}(t, s), z \mathbf{B}_{\mu_{11}}(t, s))$ and $\mathbf{B}_2(t, z) = (\mathbf{B}_{\mu_{02}}(t), z \mathbf{B}_{\mu_{12}}(t))$. To overcome the roughness in the trajectory functions, we exploit parameter regularization and specify the penalty function by $J_\lambda(\boldsymbol{\gamma})$, where the tuning parameter λ controls the level of the roughness in $\boldsymbol{\gamma}$. An iterative algorithm is applied as follows.

- Step 1.1: Given $\boldsymbol{\theta}$ obtain $\boldsymbol{\gamma}$.

Re-expressing the estimating equation (4.5) gives

$$g(E\{Y_{ij}|T_i, \delta_i, \mathbf{X}_i\}) \approx D_{\gamma_{ij}} \boldsymbol{\gamma}$$

where $D_{\gamma_{ij}} = (\delta_i \mathbf{B}_1(t_{ij}, T_i, \mathbf{X}_i^T \boldsymbol{\theta}_1) \boldsymbol{\gamma}_1, (1 - \delta_i) 1(T_i = \tau) \mathbf{B}_2(t_{ij}, \mathbf{X}_i^T \boldsymbol{\theta}_2) \boldsymbol{\gamma}_2)$. It can be viewed as a penalized regression model, and the penalized GEE is

$$\sum_{i=1}^n \mathbf{D}_{\gamma_i}^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\eta}_i) - J_\lambda(\boldsymbol{\gamma}) \boldsymbol{\gamma} = 0, \quad (4.6)$$

where $\mathbf{D}_{\gamma_i} = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\gamma} = (\dot{g}^{-1}(D_{\gamma_{i1}} \boldsymbol{\gamma}) D_{\gamma_{i1}}^T, \dots, \dot{g}^{-1}(D_{\gamma_{in_i}} \boldsymbol{\gamma}) D_{\gamma_{in_i}}^T)^T$ is the corresponding subject-level joint design matrix of uncensored subjects and LTS for spline coefficients $\boldsymbol{\gamma}$.

- Step 1.2: Given $\boldsymbol{\gamma}$ obtain $\boldsymbol{\theta}$.

Let $\boldsymbol{\theta}_k^{\text{old}}$ and $\boldsymbol{\theta}_{k,-1}^{\text{old}}$ be the current estimates for $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k,-1}$, respectively. We approximate

$$\boldsymbol{\theta}_k \approx \boldsymbol{\theta}_k^{\text{old}} + J(\boldsymbol{\theta}_k^{\text{old}})(\boldsymbol{\theta}_{k,-1} - \boldsymbol{\theta}_{k,-1}^{\text{old}}) \quad (4.7)$$

where $J(\boldsymbol{\theta}_k) = \partial \boldsymbol{\theta}_k / \partial \boldsymbol{\theta}_{k,-1} = (-\boldsymbol{\theta}_{k,-1} / \sqrt{1 - \|\boldsymbol{\theta}_{k,-1}\|_2^2}, \mathbf{1}_{p-1})^T$ is the Jacobian matrix of size p by $p-1$. To obtain the estimates of $\boldsymbol{\theta}_k$ we carry out a regression with $[\{\delta_i B_{\mu_1}(t_{ij}, T_i) \mathbf{X}_i^T J(\boldsymbol{\theta}_1^{\text{old}})\}^T, \{(1 - \delta_i) 1(T_i = \tau) B_{\mu_2}(t_{ij}) \mathbf{X}_i^T J(\boldsymbol{\theta}_2^{\text{old}})\}^T]^T$ as the regressors with a known intercept term. This produces an updated vector $\boldsymbol{\theta}_{k,-1}^{\text{new}}$. Then we set $\boldsymbol{\theta}_k^{\text{new}} = (\sqrt{1 - \|\boldsymbol{\theta}_{k,-1}^{\text{new}}\|_2^2}, (\boldsymbol{\theta}_{k,-1}^{\text{new}})^T)^T$ for $k = 1, 2$. The GEE for $\boldsymbol{\theta}_{-1} = (\boldsymbol{\theta}_{1,-1}^T, \boldsymbol{\theta}_{2,-1}^T)^T$ is

$$\sum_{i=1}^n \mathbf{D}_{\boldsymbol{\theta}_i}^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\eta}_i) = 0, \quad (4.8)$$

where $\mathbf{D}_{\boldsymbol{\theta}_i} = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\theta}_{-1}$ is the subject-level joint design matrix of uncensored subjects and LTS for $p-1$ index parameters $\boldsymbol{\theta}_{-1}$.

Repeatedly solving equations (4.6) and (4.8) until convergence leads to the estimator from “partial data”, denoted by $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\theta}}$. The corresponding sandwich covariance estimator for $\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\theta}}_{-1}$ is $\tilde{\boldsymbol{\Sigma}}_{-1} = \mathbf{H}_{n,\lambda}^{-1} \mathbf{M}_n \mathbf{H}_{n,\lambda}^{-1}$ where $\mathbf{M}_n = \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\eta}_i\} \{\mathbf{Y}_i - \boldsymbol{\eta}_i\}^T \mathbf{V}_i^{-1} \mathbf{D}_i$ denotes the variance of the estimating equations; the Hessian matrix is $\mathbf{H}_{n,\lambda} = \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i + J_\lambda(\boldsymbol{\theta})$; and \mathbf{D}_i is the corresponding design matrix combining $\mathbf{D}_{\boldsymbol{\theta}_i}$ and $\mathbf{D}_{\boldsymbol{\gamma}_i}$. From delta method, the sandwich covariance estimator for $\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\gamma}}$ is $\tilde{\boldsymbol{\Sigma}} = \Delta \tilde{\boldsymbol{\Sigma}}_{-1} \Delta^T$ where $\Delta = \text{diag}\{J(\boldsymbol{\theta}_1), J(\boldsymbol{\theta}_2), \mathbf{I}_\gamma\}$ and \mathbf{I}_x is an identity matrix whose dimension is the length of x .

4.3.2. Adjustment for right-censoring prior to maximum follow-up time

In the presence of right censoring prior to τ , equation (4.5) may lose efficiency because it does not use cost data from censored subjects. To use such data, we consider the following induced mean function:

$$E\{Y_{ij} | T_i, \delta_i, \mathbf{X}_i\} = E\{Y_{ij} | \tilde{T}_i > T_i, \mathbf{X}_i\}, \quad \delta_i = 0 \text{ and } T_i < \tau. \quad (4.9)$$

For subjects who were censored prior to τ , the mean and covariance functions are

$$\begin{aligned} E\{Y_{ij}|\tilde{T}_i > T_i, \mathbf{X}_i\} &= \int_{T_i}^{\tau} E\{Y_{ij}|\tilde{T}_i = s, \mathbf{X}_i\} dF_{\tilde{T}}(s|T_i, \mathbf{X}_i) \\ &\quad + E\{Y_{ij}|\tilde{T}_i > \tau, \mathbf{X}_i\}[1 - F_{\tilde{T}}(\tau|T_i, \mathbf{X}_i)], \\ \text{cov}\{\mathbf{Y}_i|\tilde{T}_i > T_i, \mathbf{X}_i\} &= \mathbf{V}_i. \end{aligned} \tag{4.10}$$

We re-denote $E\{\mathbf{Y}_i\} =: \boldsymbol{\eta}'_i$ for this subsection. The iterative algorithm in Section 3.1 can be modified as follows.

- Step 2.1: Given $\boldsymbol{\theta}$ obtain $\boldsymbol{\gamma}$.

The spline approximated model (4.9) gives

$$g(E\{Y_{ij}|T_i, \delta_i, \mathbf{X}_i\}) \approx \mathcal{D}_{\gamma ij} \boldsymbol{\gamma},$$

where $\mathcal{D}_{\gamma ij} = D_{\gamma ij} + (1 - \delta_i)1(T_i < \tau) \left[\int_{T_i}^{\tau} \mathbf{B}_1(t_{ij}, s, \mathbf{X}_i^T \boldsymbol{\theta}_1) dF_{\tilde{T}}(s|T_i, \mathbf{X}_i), \mathbf{B}_2(t_{ij}, \mathbf{X}_i^T \boldsymbol{\theta}_2) [1 - F_{\tilde{T}}(\tau|T_i, \mathbf{X}_i)] \right]$. The penalized induced GEE is

$$\sum_{i=1}^n \mathcal{D}_{\gamma i}^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\eta}_i) - J_{\lambda}(\boldsymbol{\gamma}) \boldsymbol{\gamma} = 0, \tag{4.11}$$

where $\mathcal{D}_{\gamma i} = \partial \boldsymbol{\eta}'_i / \partial \boldsymbol{\gamma} = (\dot{g}^{-1}(\mathcal{D}_{\gamma i1} \boldsymbol{\gamma}) \mathcal{D}_{\gamma i1}^T, \dots, \dot{g}^{-1}(\mathcal{D}_{\gamma in_i} \boldsymbol{\gamma}) \mathcal{D}_{\gamma in_i}^T)^T$ is the corresponding subject-level induced joint design matrix of all available subjects for spline coefficients $\boldsymbol{\gamma}$. Here we provide some details of the solution. If $g(\cdot)$ is identity, the close form solution is

$$\hat{\boldsymbol{\gamma}} = \left\{ \sum_{i=1}^n \mathcal{D}_{\gamma i}^T \mathbf{V}_i^{-1} \mathcal{D}_{\gamma i} + J_{\lambda}(\boldsymbol{\gamma}) \right\}^{-1} \left\{ \sum_{i=1}^n \mathcal{D}_{\gamma i}^T \mathbf{V}_i^{-1} \mathbf{Y}_i \right\}. \tag{4.12}$$

If $g(\cdot)$ is non-identity, we adopt the Newton-Raphson method with Fisher scoring,

$$\hat{\gamma}^{\text{new}} = \hat{\gamma}^{\text{old}} + \left\{ \sum_{i=1}^n \mathcal{D}_{\gamma i}^{\text{old},T} \mathbf{V}_i^{-1} \mathcal{D}_{\gamma i}^{\text{old}} + J_{\lambda}(\gamma) \right\}^{-1} \left\{ \sum_{i=1}^n \mathcal{D}_{\gamma i}^{\text{old}} \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \hat{\boldsymbol{\eta}}_i^{\text{old}}\} - J_{\lambda}(\gamma) \hat{\gamma}^{\text{old}} \right\}, \quad (4.13)$$

where $\mathcal{D}_{\gamma i}^{\text{old}}$ and $\hat{\boldsymbol{\eta}}_i^{\text{old}}$ are the updated design matrix and outcome estimate at current iteration.

- Step 2.2: Given γ obtain $\boldsymbol{\theta}$.

Similar to the approximation in equation (4.7), the GEE for $\boldsymbol{\theta}$ is

$$\sum_{i=1}^n \mathcal{D}_{\boldsymbol{\theta} i}^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\eta}_i) = 0, \quad (4.14)$$

where $\mathcal{D}_{\boldsymbol{\theta} i} = \partial \boldsymbol{\eta}'_i / \partial \boldsymbol{\theta}_{-1} = (\dot{g}^{-1}(\mathcal{D}_{\gamma i 1} \boldsymbol{\theta}) \mathcal{D}_{\boldsymbol{\theta} i 1}^T, \dots, \dot{g}^{-1}(\mathcal{D}_{\gamma i n_i} \boldsymbol{\theta}) \mathcal{D}_{\boldsymbol{\theta} i n_i}^T)^T$ is the corresponding subject-level induced joint design matrix of all available subjects for index parameters $\boldsymbol{\theta}$. $\mathcal{D}_{\boldsymbol{\theta} i j} = \left[\{ \mathbf{X}_i^T \boldsymbol{\theta}_1 \mathcal{B}_{ij,1} \gamma_1 \mathbf{X}_i^T J(\boldsymbol{\theta}_1) \}^T, \mathbf{X}_i^T \boldsymbol{\theta}_2 \mathcal{B}_{ij,2} \gamma_2 \mathbf{X}_i^T J(\boldsymbol{\theta}_2) \}^T \right]^T$, where

$$\mathcal{B}_{ij,1} = \delta_i B_{\mu_{11}}(t_{ij}, T_i) + (1 - \delta_i) 1(T_i < \tau) \int_{T_i}^{\tau} B_{\mu_{11}}(t_{ij}, s) dF_{\tilde{T}}(s | T_i, \mathbf{X}_i),$$

$$\mathcal{B}_{ij,2} = (1 - \delta_i) 1(T_i = \tau) B_{\mu_{12}}(t_{ij}) + (1 - \delta_i) 1(T_i < \tau) B_{\mu_{12}}(t_{ij}) [1 - F_{\tilde{T}}(\tau | T_i, \mathbf{X}_i)].$$

If $g(\cdot)$ is non-identity, denote the intercept term as $\boldsymbol{\mathcal{W}}_i = (\mathcal{W}_{i1}^T, \dots, \mathcal{W}_{i n_i}^T)^T$, where

$$\mathcal{W}_{ij} = \mathbf{X}_i^T \boldsymbol{\theta}_1 \mathcal{B}_{ij,1} \gamma_1 - \mathbf{X}_i^T \boldsymbol{\theta}_1 \mathcal{B}_{ij,1} \gamma_1 \mathbf{X}_i^T J(\boldsymbol{\theta}_1) \boldsymbol{\theta}_{1,-1} + \mathbf{X}_i^T \boldsymbol{\theta}_2 \mathcal{B}_{ij,2} \gamma_2 - \mathbf{X}_i^T \boldsymbol{\theta}_2 \mathcal{B}_{ij,2} \gamma_2 \mathbf{X}_i^T J(\boldsymbol{\theta}_2) \boldsymbol{\theta}_{2,-1}.$$

$\boldsymbol{\theta}$ can be updated by

$$\hat{\boldsymbol{\theta}} = \left\{ \sum_{i=1}^n \mathcal{D}_{\boldsymbol{\theta} i}^T \mathbf{V}_i^{-1} \mathcal{D}_{\boldsymbol{\theta} i} \right\}^{-1} \left\{ \sum_{i=1}^n \mathcal{D}_{\boldsymbol{\theta} i}^T \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mathcal{W}}_i) \right\}. \quad (4.15)$$

If $g(\cdot)$ is non-identity, we adopt the Newton-Raphson method with Fisher scoring,

$$\hat{\boldsymbol{\theta}}^{\text{new}} = \hat{\boldsymbol{\theta}}^{\text{old}} + \left\{ \sum_{i=1}^n \mathcal{D}_{\boldsymbol{\theta}i}^{\text{old},T} \mathbf{V}_i^{-1} \mathcal{D}_{\boldsymbol{\theta}i}^{\text{old}} \right\}^{-1} \left\{ \sum_{i=1}^n \mathcal{D}_{\boldsymbol{\theta}i}^{\text{old}} \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\eta}_i^{\text{old}}\} \right\}. \quad (4.16)$$

Repeatedly solving equations (4.11) and (4.14) until convergence leads to the estimator from “all data”, denoted by $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$. To adjust for the variation introduced by the survival estimation, the joint sandwich covariance formula for $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}_{-1}$ is $\mathcal{H}_{n,\lambda}^{-1} \mathcal{M}_n \mathcal{H}_{n,\lambda}^{-1}$ where \mathcal{M}_n denotes the corresponding variance of the joint estimating equations, and $\mathcal{H}_{n,\lambda}$ denotes the joint Hessian matrix. Specifically,

$$\mathcal{M}_n = \sum_{i=1}^n \begin{bmatrix} u_i(\boldsymbol{\beta}) \\ \mathcal{D}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\eta}_i\} \end{bmatrix} [u_i(\boldsymbol{\beta})^T, (\mathcal{D}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\eta}_i\})^T]^T$$

where $u_i(\boldsymbol{\beta})$ is the score function for the survival model parameter $\boldsymbol{\beta}$, and \mathcal{D}_i is the corresponding design matrix combining $\mathcal{D}_{\boldsymbol{\theta}i}$ and $\mathcal{D}_{\boldsymbol{\gamma}i}$.

$$\mathcal{H}_{n,\lambda} = \begin{bmatrix} \sum_{i=1}^n \nabla u_i(\boldsymbol{\beta}) & 0 \\ -\mathcal{D}_i^T \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}_i^T} & \mathbf{H}_{n,\lambda} \end{bmatrix}$$

From delta method, the sandwich covariance estimator for $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}$ is $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Delta} \hat{\boldsymbol{\Sigma}}_{-1} \boldsymbol{\Delta}^T$ where $\boldsymbol{\Delta} = \text{diag}\{\mathbf{I}_{\boldsymbol{\beta}}, \boldsymbol{\Delta}\}$. The sandwich covariance estimators for baseline $\hat{\mu}_{01}(\cdot, \cdot), \hat{\mu}_{02}(\cdot)$ are thus given by

$$\begin{aligned} \text{cov}(\hat{\mu}_{01}(t, s)) &= \mathbf{B}_{\mu_{01}}(t, s) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}_{01}} \mathbf{B}_{\mu_{01}}(t, s)^T \\ \text{cov}(\hat{\mu}_{02}(t)) &= \mathbf{B}_{\mu_{02}}(t) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}_{02}} \mathbf{B}_{\mu_{02}}(t)^T, \end{aligned} \quad (4.17)$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}_{0l}}$ is the subset of $\hat{\boldsymbol{\Sigma}}$ corresponding to $\boldsymbol{\gamma}_{0l}, l = 1, 2$. The variance estimator shares

the robustness property with parametric sandwich estimator, i.e., the consistency holds even if the covariance matrix is misspecified. The maximum efficiency will be achieved if the covariance matrix is correctly specified. The sandwich variance estimator can properly account for the variability of estimated survival distribution. When the sample size is relatively small, ignoring the uncertainty from estimated survival distribution underestimate variance estimate. An algorithm for implementing the proposed method through an iterative alternating optimization procedure is described in Algorithm 2.

4.3.3. Estimating the covariance structure

The covariance structure is modeled by following the idea for GEE (Liang et al., 1992): $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, where $\mathbf{R}_i(\boldsymbol{\alpha})$ is a symmetric working correlation matrix of \mathbf{Y}_i with parameter $\boldsymbol{\alpha}$. Common choices of the working correlation include compound symmetry or AR(1). $\mathbf{A}_i = \text{diag}\{\text{var}(Y_{i1}), \dots, \text{var}(Y_{in_i})\}$. Although the medical costs may be zero-inflated and skewed to the right, indicating clear evidence of heteroscedasticity, modeling such features can lead to complicated model formulation, numerical instability, and convergence issues. Therefore, we prefer using a simpler variance model for practical consideration. Let $\rho_i(\boldsymbol{\alpha})$ be the corresponding expectations given the prespecified correlation structure. The estimator for $\boldsymbol{\alpha}$ can be solved by the second-moment estimating equations (Liang and Zeger, 1986b; Fitzmaurice et al., 2008)

$$\sum_{i=1}^n \delta_i \{\partial \rho_i(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}\}^T \{\mathbf{U}_i(\boldsymbol{\theta}, \boldsymbol{\beta}) - \rho_i(\boldsymbol{\alpha})\} = 0, \quad (4.18)$$

where $\mathbf{U}_i(\boldsymbol{\theta}, \boldsymbol{\beta}) = (\epsilon_{i1}\epsilon_{i2}, \epsilon_{i1}\epsilon_{i3}, \dots, \epsilon_{i,n_i-1}\epsilon_{i,n_i})^T$ is the estimating equations based on distinct pairwise products of residuals. In preliminary analysis, we observed that the intra-subject correlation of incidence cost for LTS is relatively smaller than the rest of the subjects, therefore, a separate model of the variance function and correlation parameters for LTS should be specified. Replacing the indicator δ_i in equation (4.18) by $I(\delta_i = 0, T_i =$

τ), such a model can be estimated by the same calculation procedure described above. Note that the covariance \mathbf{V}_i for subjects censored prior to τ does not contribute to the estimation for variance-covariance model, as their true survival time is unobserved. However, their conditional survival distribution covers the interval $(T_i, \tau]$ as well as the mass (τ, ∞) , therefore the covariance structure \mathbf{V}_i is derived based on variance-covariance models of both uncensored subjects and LTS.

Algorithm 2 Fit a varying coefficient single-index model with censored covariates

```

1: Inputs:
   Data  $\{\mathbf{Y}_i, \mathbf{X}_i, T_i, \delta_i; i = 1, \dots, n\}$ 
2: Initialize:
    $\tilde{\gamma}^{(0)} = 0$ ;  $\boldsymbol{\theta}^{(0)}$  is estimated through fitting a generalized regression
   model assuming  $\mu_{12}(t, s)$  and  $\mu_{22}(t)$  to be constant
3: while convergence criteria is not satisfied do
4:   Update  $\phi(\cdot)$  and  $\boldsymbol{\alpha}$  for  $\mathbf{V}_i$  following the steps given in Section 4.3.3
5:   Update  $\tilde{\gamma}$  and  $\hat{\boldsymbol{\theta}}$  sequentially by Step 1.1 and Step 1.2
6: end while
7: Reinitialize:
    $\hat{\boldsymbol{\theta}}^{(0)} = \tilde{\boldsymbol{\theta}}$ , and estimate survival model for  $\boldsymbol{\beta}$  to set the new design
   matrix as  $\mathcal{D}_i$ 
8: while converge criteria is not satisfied do
9:   Update  $\phi(\cdot)$  and  $\boldsymbol{\alpha}$  for  $\mathbf{V}_i$  following the steps given in Section 4.3.3
10:  Update  $\hat{\gamma}$  and  $\hat{\boldsymbol{\theta}}$  sequentially by Step 2.1 and Step 2.2
11: end while

```

4.3.4. Selecting Knots and smoothing parameter

Similar to penalized spline regression models, the trade-off between efficiency and bias is controlled by the knot choice and smoothing parameter. Following the guideline in Eilers and Marx (2010), we use 5 equally spaced knots on each dimension of the bivariate surface. As pointed out by Eilers and Marx (2002), the degree of the polynomial to construct the spline basis function is less important in the case of penalized spline models, which use a relatively large number of knots. We use quadratic penalized splines throughout this chapter. The penalty parameter $\boldsymbol{\lambda}_{\boldsymbol{\theta}}$ can be selected through a grid search to minimize the quasi-Generalized Cross-Validation (QGCV) criteria originated

in Fu (2003):

$$\text{QGCV}(\boldsymbol{\lambda}_{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \frac{n^{-1} \sum_{i=1}^n \mathbf{r}_i^T \mathbf{R}_i^{-1} \mathbf{r}_i}{\{1 - uv/N\}^2}. \quad (4.19)$$

The numerator is the average of the weighted square of the deviance residuals \mathbf{r}_i corresponding to the n_i observations, with the weights being the inverse of the working correlation matrix \mathbf{R}_i . In the denominator, uv is the effective number of parameters in the penalized model, where u is the number of covariates, v is the shrinkage rate (sum of squares of parameters) of the penalized GEE estimator over the non-penalized GEE estimator, and $N = \sum_{i=1}^n n_i^2 / |\mathbf{R}_i|$ is the effective degree of freedom of correlated observations.

4.3.5. Estimating the survival distribution

An important step to accommodate the censored covariate \tilde{T} is to estimate the conditional distribution $F_{\tilde{T}}(s|T, \mathbf{X})$. The baseline hazard can be estimated by the Breslow's method, but we do not choose this approach for two reasons. First, the variance of the cost trajectory estimator is more difficult to obtain because the Breslow's method introduces in theory infinitely many nuisance parameters in the GEE estimation. Second, the Breslow's method does not directly lead to an estimated density function of survival with proper smoothness control. Using a spline approximation for the log-hazard function, the estimator of survival parameter $\boldsymbol{\beta}$ can be obtained by fitting a Poisson regression model (Kauermann, 2005).

The integration in equation (4.10) is computationally intensive. Hence we use coarsening of survival time to approximate the integration of mean cost trajectories with respect to time to death. Suppose patient i is censored at time $T_i < \tau$, we assume that the true time to death \tilde{T}_i falls in one of G_i prespecified time intervals, denoted as $L_{ig} (g = 1, \dots, G_i)$. $\cup_{g=1}^{G_i-1} L_{ig} = (T_i, \tau]$ and $L_{iG_i} = (\tau, \infty)$. Let l_{ig} be a representative point

of corresponding interval; we specify l_{ig} as the middle point of L_{ig} , $g = 1, \dots, G_{i-1}$, and l_{iG_i} indicates (τ, ∞) throughout this chapter. To balance between approximation accuracy and computational feasibility, we suggest the intervals of each censored subject to have equal length of at least one unit of the measurement time, and a cap at $G_i \leq 10$ to be the maximum number of intervals that a censored patient could have. We approximate the expectation of mean cost trajectories for censored patients by a weighted average of possible mean cost trajectories,

$$E\{Y_i(t)|\tilde{T} > T_i\} \approx \sum_{g=1}^{G_i} E(Y(t)|\tilde{T} = l_{ig})P_{ig}(\boldsymbol{\beta}),$$

where $P_{ig}(\boldsymbol{\beta}) = P(\tilde{T} \in L_{ig}|\tilde{T} > T_i)$ denotes the conditional probability weight for the actual time to death of subject i that falls within interval L_{ig} .

4.3.6. Hypothesis tests

We study the asymptotic properties of the proposed estimator in the supplementary material. Applying the estimation procedure described in previous section, we provide two hypothesis tests for the trajectory curves and covariate index parameters. In practice, parsimonious models are always desirable to reduce computation burden. To test whether the function $\mu_{11}(\cdot, \cdot)$ has a specific parametric form, we set up the hypothesis testing as $H_0 : \mu_{11}(\cdot, \cdot) = \mu_{\boldsymbol{\gamma}}(\cdot, \cdot)$ versus $H_a : \mu_{11}(\cdot, \cdot) \neq \mu_{\boldsymbol{\gamma}}(\cdot, \cdot)$ where $\mu_{\boldsymbol{\gamma}}(\cdot, \cdot)$ is a certain given parametric function with parameter vector $\boldsymbol{\gamma}$. For example, setting $\mu_{\boldsymbol{\gamma}}(t, s) = \gamma_0$ for some constant γ_0 , we aim to test whether there exists non-constant trajectory effects on the covariate index; while setting $\mu_{\boldsymbol{\gamma}}(t, s) = \gamma_0 + \gamma_1 t$ (a linear function), we attempt to test whether there exists linear interaction effects between t and covariate(s). We introduce a Wald statistic to test whether the complex trajectory part on the covariate index can be replaced by a simpler nested parametric function. The Wald test is defined

by testing whether multiple spline parameters equal to zero,

$$T_n = (R\hat{\gamma}_k)^T [R\hat{V}_{\gamma_k} R^T]^{-1} (R\hat{\gamma}_k),$$

where R is the index matrix corresponds to the spline parameters. From Appendix C, T_n asymptotically follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the null and alternative.

Given that the trajectory effect for the covariate index is not zero, $\mu_{11}(\cdot, \cdot) \neq 0$, we can also set up a Wald test procedure to test whether the individual covariate effect θ_{kl} ($k = 1, 2; l = 2, \dots, p$) equals to zero. This test is based on the asymptotic distribution $\hat{\theta}_{kl}/\sigma_{\theta_{kl}} \rightarrow_d N(0, 1)$ under H_0 .

4.4. Analyses of Prostate Cancer Cost Data from SEER-Medicare

Prostate cancer is the most common cancer and the second leading cause of cancer death among men in the United States (Cronin et al., 2018), as 3.1 million new cases of prostate cancer diagnosed between 2003 and 2017 (Siegel et al., 2020). It is a slow growing cancer mostly for elder men, which largely affect patients' quality of life with various symptoms including pain, fatigue or difficult urinating. It is also one of the most expensive health care problems in America, and the economic burden for prostate cancer in 2020 is around 20 billion U.S. dollars (Mariotto et al., 2020). Both the survival and economic burden of prostate cancer continues to increase, and the differences by patient characteristics could inform public health planning related to cancer survivor care.

We apply the proposed method to the medical cost data of patients with local/regional prostate cancer from the SEER-Medicare linked database. The study cohort consisted of 161,630 patients over age 65 who were diagnosed with advanced prostate cancer between January 1, 2003 and December 31, 2015. The medical costs included

Medicare payments for inpatient and outpatient services covered by Medicare Parts A and B, from the month of diagnosis to death or the end of follow-up (i.e. December 31, 2016). To ensure the completeness of cost data within the first year of cancer diagnosis, we further restricted the study cohort to patients with continuous Parts A and B enrollment, and were not enrolled in HMOs in this duration. The goal was to analyze the effect of patient baseline characteristics as well as whether the patient received definitive treatment (i.e., surgical or radiation) within the first year of diagnosis on the population-averaged medical cost trajectory.

In our analysis, the response Y is a right-skewed and zero-inflated continuous outcome on original dollar scale. We specified a compound symmetry structure to accommodate the within-subject correlation. The explanatory variables included comorbidity scores, age at diagnosis, race and ethnicity, and receipt of definitive treatment within 12 months after initial diagnosis (radiotherapy (yes/no) or surgery (yes/no)). We dichotomized comorbidity score (0-1 vs. >1) and age at diagnosis (65-74 vs. ≥ 75), respectively. Race/ethnicity was classified into three categories: non-Hispanic white, non-Hispanic Black, and others. For model fitting, we used truncated quadratic spline basis with five equally spaced knots in both the mean functions of costs and the hazard function of survival. For model selection purpose, we used the proposed Wald statistic and found that at 1% significance level, the varying coefficient effect for covariate index was not zero, and thus the varying coefficients $\mu_{11}(t, s)$ and $\mu_{12}(t, s)$ were non-linear. The optimal penalty parameter was selected via the QGCV criterion. The convergence criterion was that the maximum absolute difference between the current parameter estimate and previous parameter estimate was less than 10^{-4} . The corresponding pointwise confidence intervals for mean baseline cost trajectory were calculated.

Table 4.1 summarizes the patient characteristics, point estimates of θ and the

corresponding 95% confidence intervals (CI). Figure 4.4 (a) shows the estimated reference medical cost trajectory and their 95% pointwise CI on selected survival time at 16, 24, 32, and over 40 quarters (LTS). The 3D reference cost trajectory surfaces are depicted in Figure 4.4 (b-c). To aid in the interpretation for the varying coefficient function of single-index $\mu_{11}(t, s)$ and $\mu_{12}(t, s)$, we depicted the estimated mean cost trajectory at selected index values and survival time in Figure 4.3. Figure 4.2 shows the nomogram to visually translate the covariate effect towards total cost.

Cohort Characteristics.

For patients whose survived below and over 40 quarters, the mean age was 73.2 and 71.8 years respectively. Table 4.1 shows that for patients whose survived over 40 quarters (N=27,467), a high proportion of patients have 0-1 comorbidity score (80.7%). In comparison, 67.2% of patients who survived below 40 quarters (N=134,163) have 0-1 comorbidity score. Racial distributions were similar in both survival groups, although patients who survived over 40 quarters (LTS) had a higher proportion of non-Hispanic White and a lower proportion of “old-old” ages (e.g. ≥ 75 , Schilling (2005)). These LTS had a disproportionately high rate of initial treatment within 12 months after cancer diagnosis (e.g. radiotherapy or surgery).

Nomogram on total cost summary

We present a nomogram in Figure 4.2 to summarize the covariate effect through index points on total cost given survival time. Nomogram is widely used as a visualization tool to show the relationship between covariates and outcomes for complex statistical models such as cox proportional hazard model and logistic regression model. In the past two decades, nomograms are widely used in clinical decision making for prostate cancer, such as the prediction of time-to-PSA level elevation recurrence (Kattan, 2003), and the *Partin Table* for pathological cancer stage (Partin et al., 1993). For example, from Figure

4.2 we read that a patient in the reference group has 0 total points, and if he died at 16, 24 or 32 quarters after cancer diagnosis, he costed \$144K, \$167K or \$182K on average, respectively. Given other conditions unchanged, a similar non-Hispanic Black patient has 0.266 total points, and the mean total cost increase \$16K, \$19K or \$21K, respectively. If the same patient received radiotherapy in initial treatment and had over 1 comorbidity scores, the total points becomes 0.895, and the average total costs reach \$199K, \$230K or \$254K, respectively. Comparatively, the mean total cost for LTS in reference group is \$106K. Fixing other conditions, the total points for a similar non-Hispanic Black is -0.055, and thus he costed \$2K less on average for LTS. Receiving radiotherapy in initial treatment barely change the total cost for LTS, while high comorbidity score is associated to \$43K increase in total cost. The proposed nomogram can translate the complicated regression model into informative graphical representations which can play a critical role for cost estimation and evaluation.

Varying coefficient and Linear index associated with cost.

Figure 4.3 clearly shows that the index values were positively associated with cost trajectory, and the relations were not linear. This evidence aligns with the hypothesis test result ($H_0 : \mu_{1k}(t, s)$ is constant, $k=1, 2$; P-value <0.0001) suggesting that a simple model such as (4.1) is inadequate to capture the complicated relationship between covariates and cost trajectory. Figure 4.3 suggests a dramatic elevation of the cost trajectory right after local/regional cancer diagnosis, especially for subjects who received radiotherapy during initial treatment. We can see that among patients who died at 16 quarters after cancer diagnosis, the first quarter costs were highest for non-Hispanic Black (\$9,292; \$8,985 to \$9,599), lower for non-Hispanic White (\$7,446; \$7,117 to \$7,775) and intermediate for others (\$8,383; \$8,074 to \$8,692). After one year of diagnosis, the costs goes down gradually, and the trajectories for different index seem to overlap on the surveil-

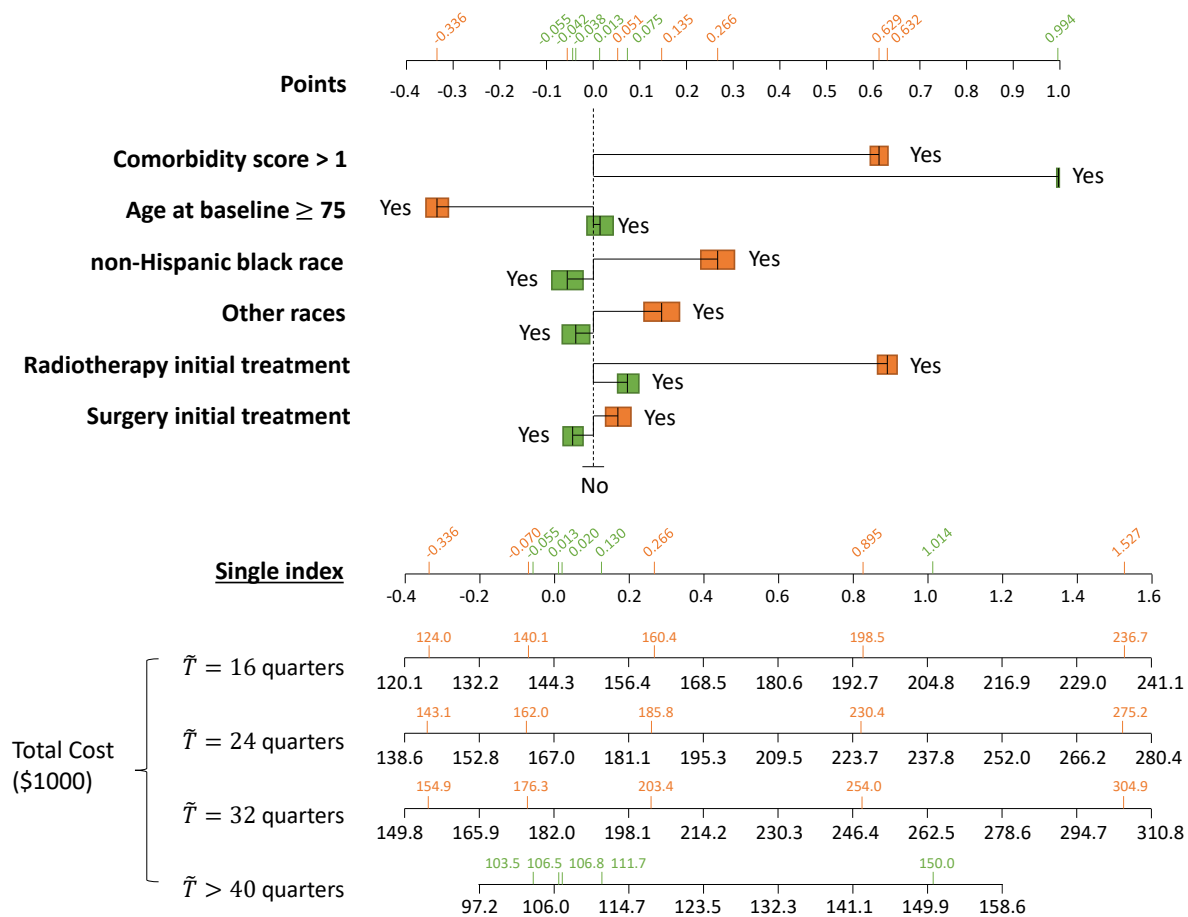


Figure 4.2: SEER-Medicare data application results of the cost Nomogram. Instructions for policy maker are as follows: first, locate the patient's baseline covaraites on the top axes. Draw a vertical line straight upwards to the points axis to determine how many points are contributed; second, sum the points achieved for each predictor and locate this sum on the single index axis below. Third, draw a vertical line straight down to find the patient's total cost (\$1000) assuming the time to death is 16, 24, 32 or over 40 quarters.

lance phase. Within one year before death, the cost trajectory quickly increases, and the rates of increment for higher index values are generally higher. The area under the cost trajectory is the mean total cost reported in the nomogram; for example, for non-Hispanic Black aged 65-74 who had over 1 comorbidity scores and received radiotherapy as initial treatment, the single index was 1.014 and the mean total cost was \$150K.

In Table 4.1, among patients whose survived below 40 quarters, we observe higher mean cost trajectory when subjects, had higher comorbidity score, were “young-old” (e.g. from age 65 to 74) and non-Hispanic Black, received radiotherapy or surgery as initial treatment. This result is consistent with previously published findings in Schmid et al. (2016); Trogon et al. (2019). Since all variables are binary, and the linear coefficients are on a unit circle, we are able to compare the relative effects of the covariates. For example, in Table 4.1, the mean effect of being non-Hispanic Black (0.266; 0.231 to 0.302) is twice as large as being in “other” race categories (0.135; 0.097 to 0.173). Table 4.1 shows some interesting findings for patients who survived over 40 quarters (LTS). We observe slightly lower cost for patients who were non-Hispanic Black or other races, received surgery within the first 12 months of cancer diagnosis. It suggests the policy makers to explore deeper the latent influential factors for racial disparities in the society and the Medicare system. This demonstrates the importance of our research goal and the significant contribution of the proposed methodology in health economics as well as health policy research (Zheng et al., 2019).

Reference cancer care cost.

The reference group represents non-Hispanic White patients aged 65-74 who had comorbidity score 0-1, and did not receive radiotherapy and surgery as initial treatment within 12 months after cancer diagnosis. As shown in Figure 4.4(a), the reference medical cost trajectory has a nonlinear dependence on the time after diagnosis, and the corresponding

relationships for different cancer stages are highly heterogeneous for different survival time. The trajectories are “U-shaped”, which means that patients are likely to receive active care right after the initial cancer diagnosis as well as intensive hospital or palliative care close to the end-of-life (EOL).

The treatment guideline for prostate cancer from American Cancer Society states that the initial treatment are determined by physicians based on the stage, the risk of cancer recurrence after the initial treatment and on the man’s life expectancy (ACS, Accessed Apr 13, 2022). For the reference group who are healthier, observation or active surveillance are often recommended, which is less expensive. If there are signs of disease progression (e.g. elevated PSA level and Gleason score), some definitive treatment such as systemic therapy may be added, and result in the elevated costs and billing delay. In Figure 4.4(a), we first look “forward” at the trajectory after diagnosis, the costs for patients who died within 16 quarters have an increasing trend within the first 4 quarters after diagnosis, indicating that such patients underwent observation or active surveillance right after diagnosis, and received active care later. This pattern is supported by the fact that the rate of systemic therapy as initial treatment for those who survived within or beyond 16 quarters after diagnosis were 64.2% vs. 44.4% respectively.

Looking “backward” at the trajectory pattern from the time to death, there is a change point around 4 quarters before death, which aligns with NCI’s definition on EOL-care phase (Mariotto et al., 2020). in Figure 4.4(a). Higher EOL cost is a common trend for the elderly due to all services: cancer-related inpatient stays, hospice care, and outpatient services (i.e., office or emergency room visits, hospital procedures, and others) (Duncan et al., 2019). However, the trajectory is “L-shaped” on average among LTS, possibly because the observed costs are not all from their EOL care.

Besides the cost right after diagnosis and before death, in the middle of life-span we observe that patients with longer-term survival have lower average quarterly cost. For example, the cost at 8, 12, 16 and 20 quarters for patients who died at 16, 24, 32 and over 40 quarters follows a decreasing trend: \$5,888, \$4,007, \$3,864 and \$ 3,787, respectively. However, the cumulative cost in the middle of patients' life-span may be higher for patients who survive longer, and thus, a simple summation used in existing literature (Mariotto et al., 2020) may be inappropriate for accurate cost estimation and projection of cancer care consumption.

Putting together all the reference trajectory at different survival quarters $s = 1, 2, \dots$, we can depict a visually “smooth” bivariate surface as shown in Figure 4.4 (b-c). Since the measurement time $t = 1, 2, \dots, s$, the surface is on a triangular region. The complex nonlinear interdependencies are clearly seen between time after diagnosis, time to death, and costs for both stages at diagnosis. The estimated compound symmetry correlation for patients who survived within 40 quarters after diagnosis is 0.07, which suggests low positive within-subject correlation in costs. The corresponding correlations are slightly lower for LTS (0.05) because the LTS include subjects with different survival times and the variability of costs is thus larger.

4.5. Simulations

In this section, we study the finite sample performance of the proposed method by two simulation examples:

Example 1. The outcome Y is sampled from a Normal distribution with baseline mean trajectory $\mu_{01}(t, s) = \cos(2\pi t/s) + 1$ to mimic U-shaped trajectories for the non-LTS subjects ($0 < s \leq \tau$), and $\mu_{02}(t) = \cos(2\pi t) \times 1(t \leq 0.5) - 1(t > 0.5) + 1$ to mimic the L-shaped trajectories for LTS ($s > \tau$). The variance is 0.25 with compound symmetry

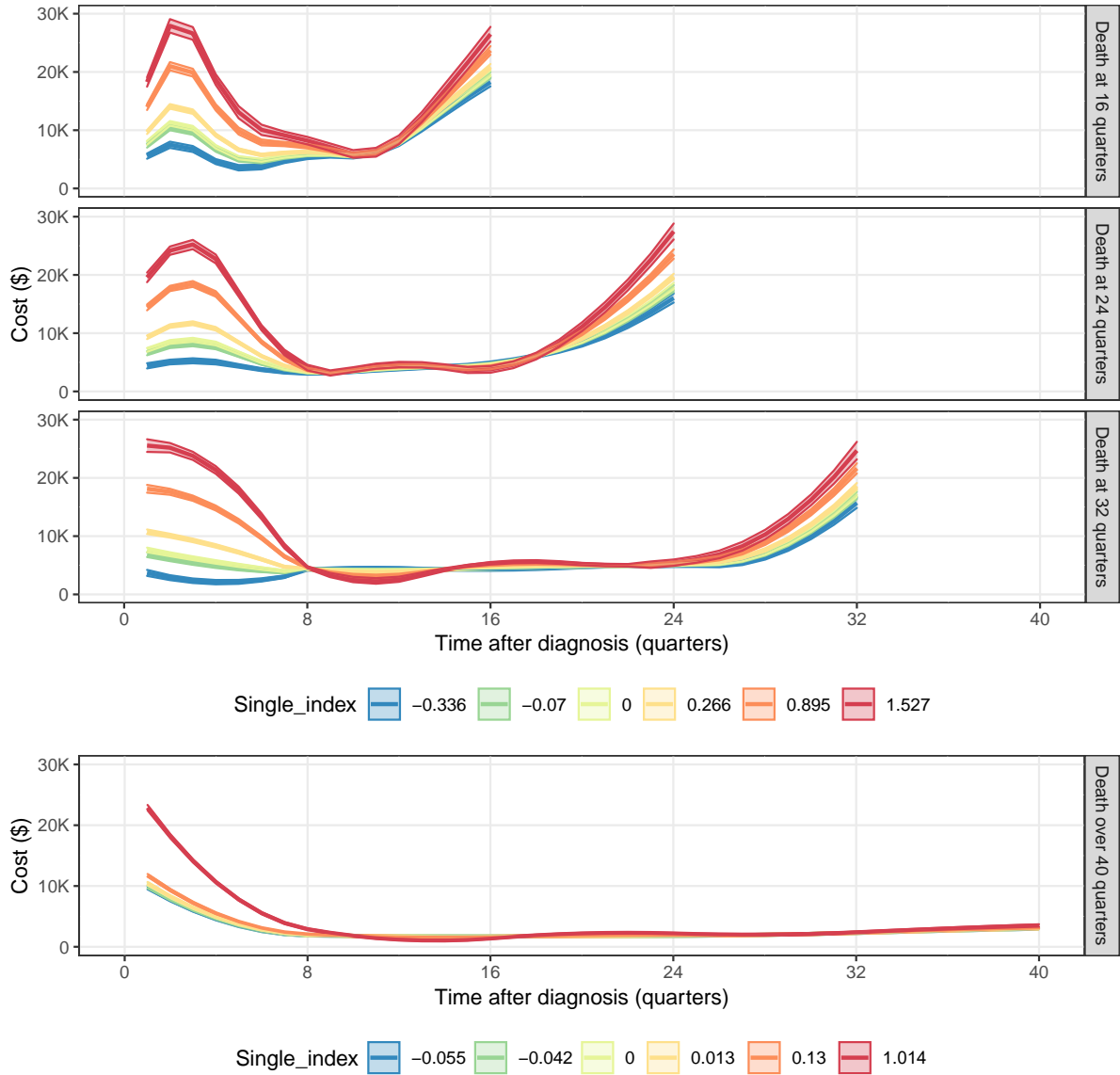


Figure 4.3: SEER-Medicare data application results of the cost trajectories of different index when survival time is 16, 24, 32 or over 40 quarters. Single index equals 0 for reference group. For patients who died within 40 quarters after cancer diagnosis (or for LTS), single index is -0.336 (0.013) for “old-old”; 0.266 (-0.055) for non-Hispanic Black; -0.07 (-0.042) for “old-old” non-Hispanic Black; 0.895 (0.13) for non-Hispanic Black who received radiotherapy as initial treatment; 1.527 (1.014) for non-Hispanic Black who have over 1 comorbidity score and received radiotherapy as initial treatment. Unmentioned conditions are the same as the reference group.

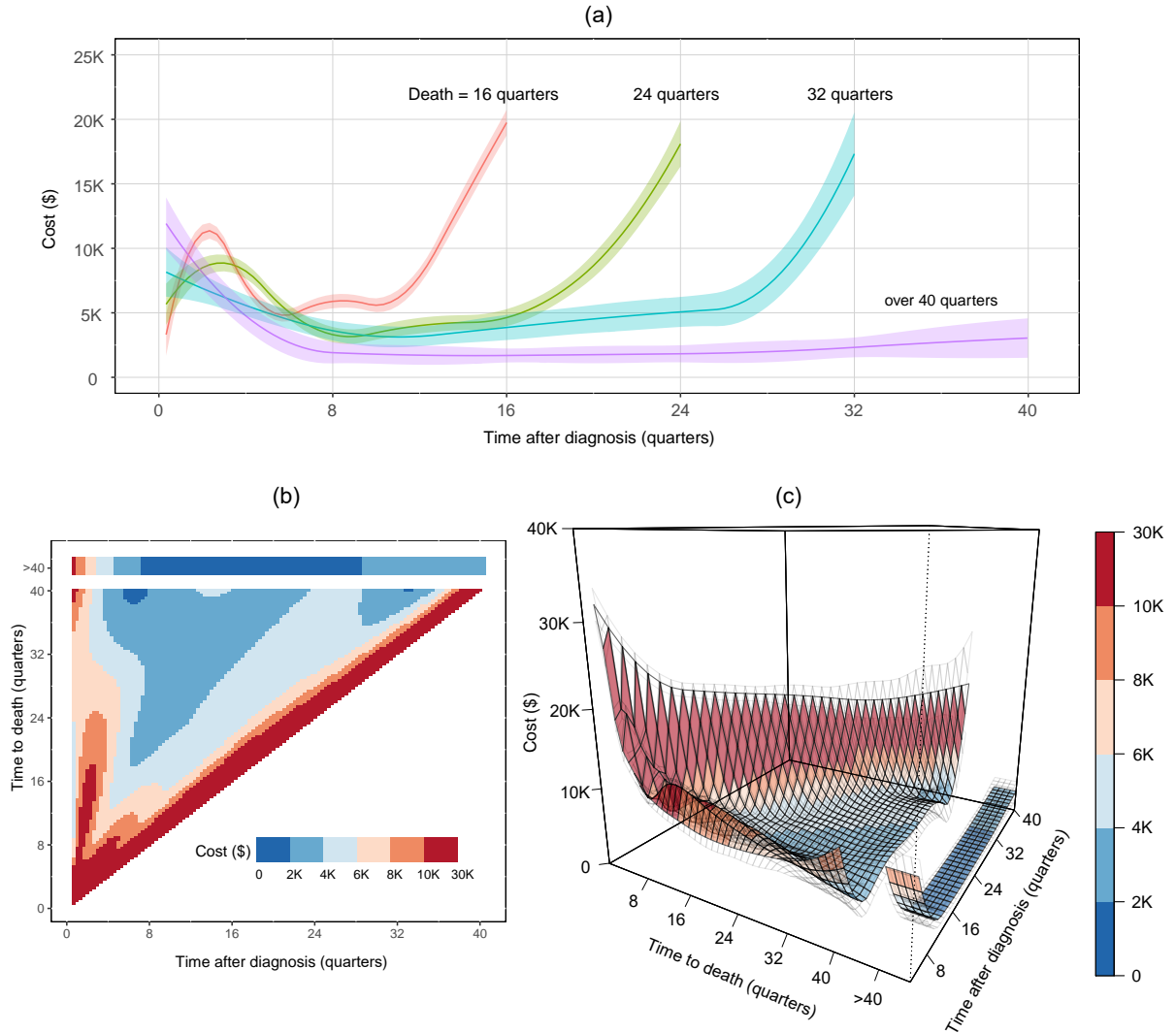


Figure 4.4: SEER-Medicare data application results for the estimated reference cost trajectories (a) when the time to death (quarter) equals $s = 16, 24, 32$, and over 40 (LTS) with 95% confidence intervals in shaded areas; (b) 2D heatmaps and (c) 3D “surface” with 95% confidence intervals with 95% upper and lower bounds in grey meshes.

correlation structure and the correlation is 0.2.

Example 2. The outcome Y is sampled from a Gamma distribution with chosen shape and scale parameters such that the baseline mean trajectory follows $\mu_{01}(t, s) = \exp(t/s)$ to mimic increasing trajectories for the non-LTS subjects ($0 < s \leq \tau$), and $\mu_{02}(t) = \exp(t)$ for LTS ($s > \tau$). The variance depends on mean as $\text{var}(Y) = \mu^2$. To simulate zero-inflation, the outcome Y is multiplied by 10/9, and then 10% outcome values are randomly set to 0.

For both examples, the covariates $X = [X_1, X_2]$ are independently generated from Uniform[0,1] and Bernoulli(0.5). The mean index functions are $(X_1 + X_2)/\sqrt{2}$, and multiplicative interact with $1 + \kappa t/s$ where κ is the non-constant strength. [Not sure what you mean.] The covariance matrix has a compound symmetry correlation structure with $\alpha = 0.2$ by using the R package `simstudy` (Goldfeld and Wujciak-Jens, 2020). The failure time of each subject is sampled such that the baseline hazard follows a Weibull distribution (scale=1/3 and shape=1), and the log of relative hazard is $2X_1 + X_2$. Independent censoring is drawn from exponential distribution with rate 2, and all survival times are administratively censored at 1. We set $n=2000$ and 4000, and repeated the simulations 1000 times for each scenario. To obtain smooth estimates of trajectory curve, we choose 5 equally spaced knots and truncated quadratic basis. The penalty parameter is chosen based on the proposed GCV-based criteria through grid search on five simulated datasets. The following metrics are evaluated for the index coefficient θ , and the nonparametric functions $\mu_{0k}, k = 1, 2$: the bias, the mean squared error (MSE), and the average coverage probability (CP) of their estimators. The results are summarized in Table 4.2, Figure 4.5 and Figure 4.6.

In Table 4.2, when the outcomes follow a multivariate Normal distribution, we see

that for index parameter estimates, the coverage probabilities are close to the nominal 95% confidence level. The bias is small, and the MSE decreases as sample size increases. Compared to the method using only uncensored or LTS (partial) data, the method using all data reduce the MSE. For example, when $n=2000$, the MSE of θ_{21} reduces by 50% when cost data from censored subjects are used. This indicates that properly accounting for censoring is very beneficial for improved efficiency when the number of observations is relatively small.

When the outcomes follow a zero-inflated Gamma distribution, the bias for index parameter estimates is small, and the MSE decreases as sample size increases. It indicates that the proposed estimation under identity link function is robust against skewness and zero-inflation because it does not model the full distribution of the cost data. The coverage probabilities for the method using partial data are slightly less than 95%, but close to 90% when the sample size is 2000. The coverage probability is closer to 95% when censored cost data is used, and the estimates of confidence interval become more accurate as sample size increases.

Figure 4.5 visualizes the simulation results of the estimated baseline trajectories for Example 1 and 2 when $n=2000$ at terminal event times $s = 0.4$ and baseline covariates $x=(0,0)$. We see that the coverage probabilities are close to the nominal 95% confidence level. The pointwise bias and MSE are both small, suggesting that the proposed estimator fit the baseline trajectory well. The MSE curve for the method using all data is lower as expected. For a few boundary points, MSE is large due to the small bias caused by limited sample size in these regions.

Lastly, we evaluate the proposed Wald test statistic in the two examples mentioned above. Figure 4.6 shows the power function curves under the given significance levels.

When the outcome is Normal, Figure 4.6 (a) and (b) shows that the power curves increase rapidly with the non-constant strength κ for $\mu_{11}(t, s)$ and $\mu_{12}(t, s)$, which are the vary coefficient functions in the proposed two-part model. The method using all data has better statistical power. When the effect is close to 0, the test sizes are all approximately at the significance level. Figure 4.6 (c) and (d) shows that the power curve for index coefficient θ_{12} and θ_{22} , which represent the second index parameter in each part of the model. As the sample size in partial data method for estimating the θ_{12} is large, its power is similar to the power of all data method. The second row of Figure 4.6 indicates similar findings for zero-inflated Gamma data. Additional simulation in appendix for multivariate gamma and zero inflated normal data shows unbiased estimation of both index coefficients and trajectory functions.

4.6. Discussion

In this chapter, we proposed a longitudinal varying coefficient single index model to detect and test for the complicated nonlinear relationships between longitudinal medical cost trajectory and baseline covariates in the presence of right-censoring. This model helps health services researchers and policy maker to understand how the baseline patient and treatment characteristics affect subsequent healthcare costs, and how that effect vary over time. Since the healthcare costs is related to both the survival time, which is subject to censoring, and the time since initial diagnosis, the model has to account for both. To our knowledge, there has been no published statistical methods for this problem. Our proposed model is flexible and interpretable, and the estimation does not rely on a distribution assumption of the cost data or correct modeling of the within-subject correlation. From a methodological perspective, this is an extension of the GEE to incorporating censored covariates.

One advantage of the proposed method is that a consistent initial estimator can be

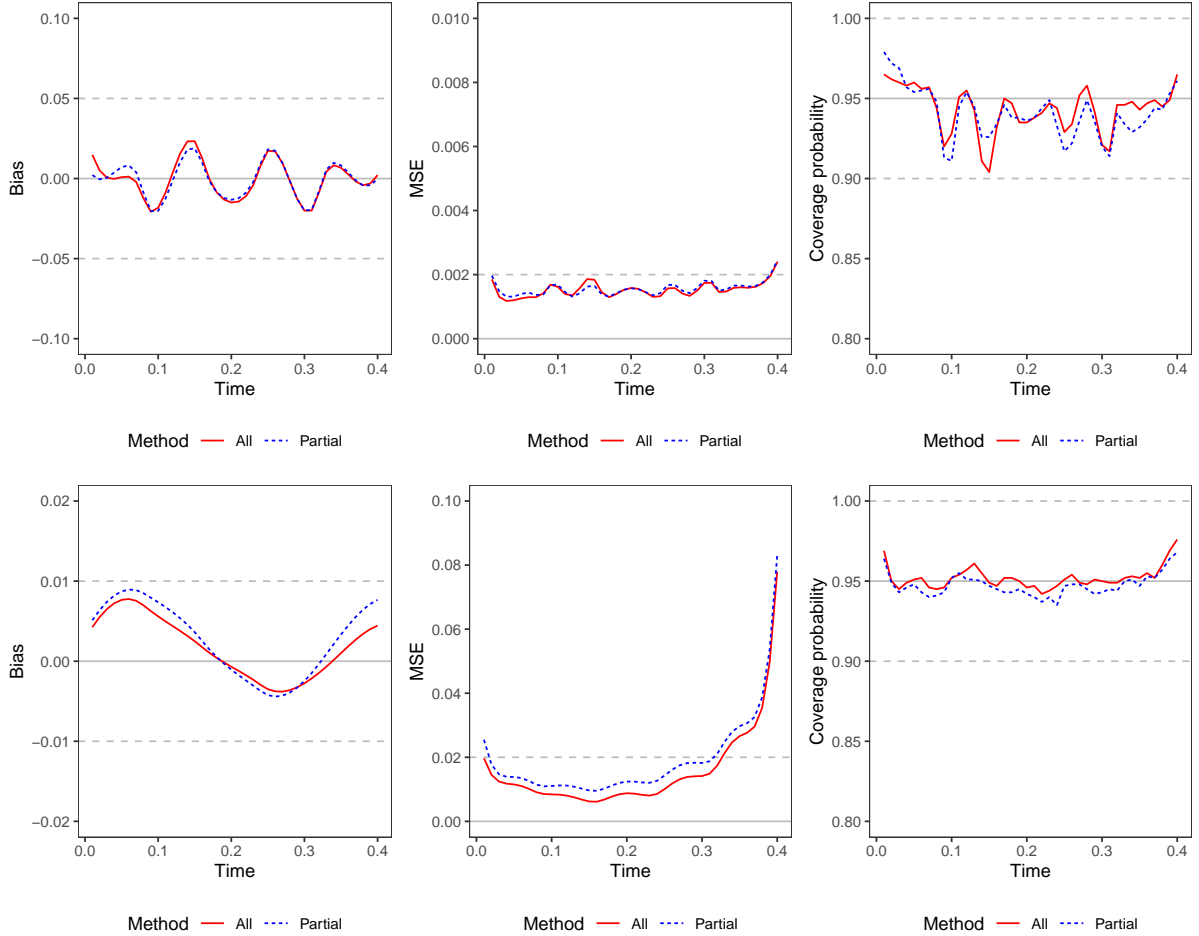


Figure 4.5: Simulation results of estimated trajectories for Normal data (top) and zero-inflated Gamma data (bottom). $n=4000$, $x=(0,0)$, $s=0.4$. Left: plot of point-wise biases; middle: plot of point-wise mean squared errors; right: plot of pointwise empirical coverage probabilities. We compare lines for methods using all data (red solid) or using partial data (blue dashed).

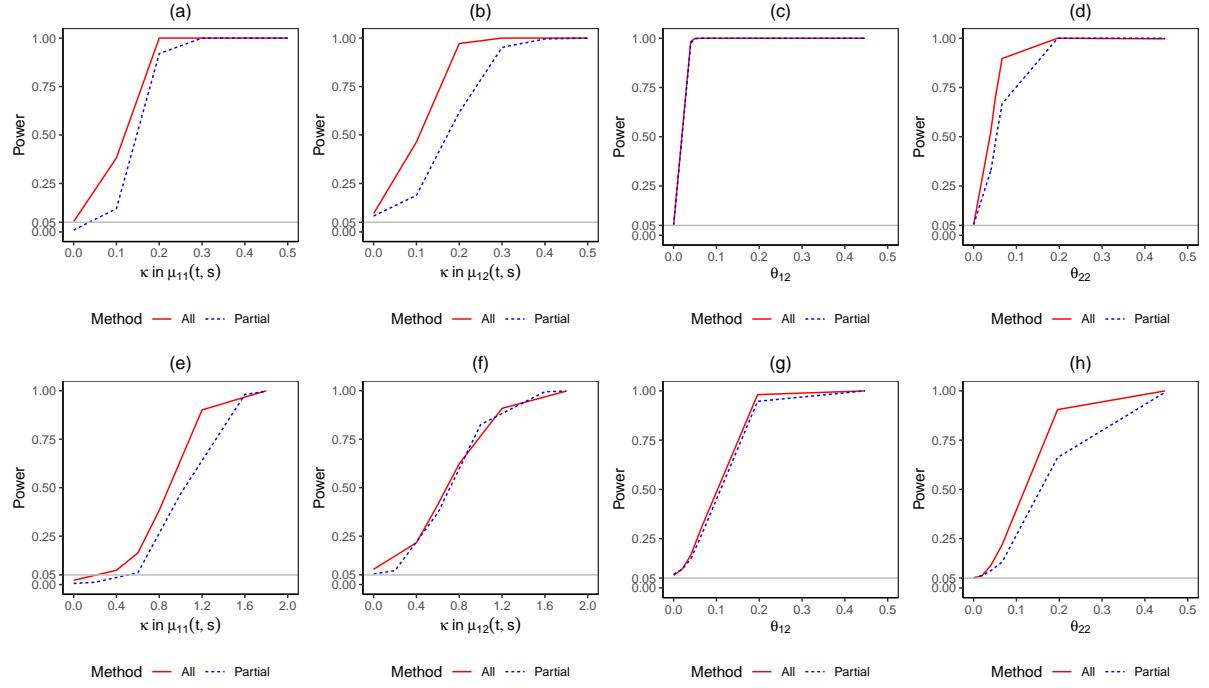


Figure 4.6: Simulation results of estimated power curve for Normal data (top) and zero-inflated Gamma data (bottom). $n=4000$, $\mathbf{x}=(0,0)$, $s=0.4$. Left: plot of point-wise biases; middle: plot of point-wise mean squared errors; right: plot of pointwise empirical coverage probabilities. We compare lines for methods using all data (red solid) or using partial data (blue dashed).

obtained by analyzing the uncensored subjects and LTS. This is helpful from a practical perspective because the data analyst can quickly explore the model formulation using standard software. The efficiency and statistical power can be significantly improved by using cost data from censored subjects, who can be a large proportion in any real world dataset. This final analysis can be completed by using the proposed method in this chapter. Motivated by the desire to provide an inference procedure for longitudinal medical cost trajectory with multiple baseline covariates. The approach can be extended to accommodate binary or count outcomes, so in principle, our method may have important policy implications for other longitudinal trajectories learned from clinical or observational data.

Table 4.1: Estimates of linear coefficients for medical cost trajectory from SEER-Medicare prostate cancer data.

	Count	Percent	Estimate	95% CI
Time to death \leq 40 quarters (N=134,163)				
Comorbidity score				
0-1	90,130	67.2	-	-
>1	44,033	32.8	0.632	[0.613, 0.651]
Age at baseline				
65-74	84,969	63.3	-	-
≥ 75	49,194	36.7	-0.336	[-0.360, -0.313]
Race				
Non-Hispanic White	101,810	75.9	-	-
Non-Hispanic Black	15,466	11.5	0.266	[0.231, 0.302]
Others	16,887	12.6	0.135	[0.097, 0.173]
Initial treatment within 12 months after cancer diagnosis				
Radiotherapy	57,474	42.8	0.629	[0.608, 0.649]
Surgery	38,964	29.0	0.051	[0.025, 0.078]
Time to death $>$ 40 quarters (N=27,467)				
Comorbidity score				
0-1	22,162	80.7	-	-
>1	5,305	19.3	0.994	[0.991, 0.997]
Age at baseline				
65-74	19,768	72.0	-	-
≥ 75	7,699	28.0	0.013	[-0.015, 0.041]
Race				
Non-Hispanic White	21,928	79.8	-	-
Non-Hispanic Black	2,400	8.7	-0.055	[-0.088, -0.022]
Others	3,139	11.5	-0.038	[-0.067, -0.009]
Initial treatment within 12 months after cancer diagnosis				
Radiotherapy	14,499	52.8	0.075	[0.053, 0.097]
Surgery	8,975	32.7	-0.042	[-0.063, -0.021]

Table 4.2: Bias, MSE and CP of linear coefficients in the simulation studies. Data generated from normal distribution and zero-inflated Gamma distribution

Example 1: Normal data							
		All data			Partial data		
n		Bias	MSE ($\times 10^{-2}$)	CP	Bias	MSE ($\times 10^{-2}$)	CP
2000	θ_{11}	-0.000	0.018	0.935	0.001	0.019	0.951
	θ_{12}	-0.000	0.018	0.935	0.001	0.019	0.951
	θ_{21}	-0.001	0.079	0.937	-0.003	0.158	0.937
	θ_{22}	0.000	0.077	0.941	0.001	0.150	0.938
4000	θ_{11}	-0.000	0.008	0.949	-0.000	0.009	0.947
	θ_{12}	0.000	0.008	0.946	0.000	0.009	0.948
	θ_{21}	-0.001	0.040	0.935	-0.001	0.074	0.947
	θ_{22}	0.000	0.040	0.932	0.000	0.073	0.944
Example 2: zero-inflated Gamma data							
		All data			Partial data		
n		Bias	MSE ($\times 10^{-2}$)	CP	Bias	MSE ($\times 10^{-2}$)	CP
2000	θ_{11}	-0.011	0.526	0.942	-0.016	0.830	0.907
	θ_{12}	0.004	0.459	0.935	0.005	0.702	0.887
	θ_{21}	-0.013	0.922	0.921	-0.029	2.032	0.901
	θ_{22}	0.001	0.808	0.916	0.003	1.567	0.888
4000	θ_{11}	-0.005	0.257	0.930	-0.008	0.376	0.913
	θ_{12}	0.002	0.241	0.933	0.003	0.335	0.926
	θ_{21}	-0.005	0.403	0.935	-0.012	0.875	0.925
	θ_{22}	-0.001	0.382	0.938	-0.000	0.781	0.925

CHAPTER 5

SUMMARY

In this dissertation, several innovative methods are proposed for modeling the longitudinal medical cost data. Right-censoring brings unique challenges to this topic, since we cannot observe complete data for all the subjects in the study, and it leads to partially observed longitudinal medical cost data. The concept of medical cost trajectory starts in (Li et al., 2018), and says we can form it by plotting the average monthly costs from diagnosis to the terminal event. Unlike the total or cumulative medical cost, which summarizes all costs of a patient throughout a follow-up period of interest, the longitudinal incident cost (by month, quarter or year) provides details on healthcare consumption across different phases of the disease continuum.

Estimating the mean cost trajectory has several statistical challenges. The shape of the trajectory is usually highly nonlinear with varying durations, depending on the diagnosis-to-death population time distribution. The terminal event may be right censored, resulting in missing subsequent costs. To address these challenges, in Chapter 2, we propose a two-stage likelihood-based approach to estimate the longitudinal cost trajectories from a joint model of longitudinal medical costs and survival. The longitudinal cost trajectories corresponding to various survival times form a bivariate surface in a triangular area. The cost trajectories are estimated using the tensor products of discretized measurement time and survival, as well as effective ridge penalties for data in two-dimensional arrays. The proposed approach balances the practical considerations of model flexibility, statistical efficiency and computational tractability.

Besides the nonlinearity in longitudinal medical cost data and right-censoring in

survival data, there are some additional challenges. Medical costs often have skewed distributions with zero-inflation and heteroscedasticity, which may not fit well with the commonly used parametric family of distributions. In Chapter 3, we propose a flexible semi-parametric model to address challenges without imposing a cost data distributional assumption. The estimation procedure is based on generalized estimating equations with censored covariates. The proposed model adopts a bivariate surface that quantifies the interrelationship between longitudinal medical costs and survival, and results in the nonlinear population mean cost trajectories given survival time. We develop a novel generalized estimating equations algorithm to accommodate covariates subject to right-censoring, without fully specifying the joint distribution of the cost and survival data.

Estimating the current cost of cancer care is important to health policy makers to understand how the patient characteristics affect the healthcare utilization. The model formulation that quantifies the association between multiple patient characteristics and nonlinear cost trajectory curves of varying lengths must take into consideration parsimony, flexibility, and interpretation. In Chapter 4, we propose a novel longitudinal varying coefficient single-index model. Patient characteristics are parsimoniously collapsed into a single-index, which represents a patient's overall propensity of healthcare utilization. The effect of the single-index on the nonlinear longitudinal cost trajectories of varying lengths is flexibly modeled by a varying coefficient, a bivariate function of longitudinal cost measurement time and survival. The model is estimated by a penalized marginal approach, with an induced mean structure that properly accounts for censoring in covariates. We established the asymptotic properties, and proposed statistical inference procedures including pointwise confidence interval of the varying coefficient and a test of covariate effect.

APPENDIX A

APPENDIX FOR CHAPTER 2

Appendix A includes three sections. In Appendix A.1, we present the proof of theoretical properties of the proposed estimator. In Appendix A.2, we study the robustness of the proposed method under misspecification of covariance structures and violation of normal assumption. In Appendix A.3, we report model checking results from analysis of the renal cell cancer cost data.

A.1. Asymptotic properties

In this section, we prove the asymptotic theory stated in the following Theorem.

Theorem A1. If the regularity assumptions A1-A8 are satisfied,

1. *(Consistency) $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are consistent estimator of $\boldsymbol{\theta}_0$.*
2. *(Asymptotic normality) $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically zero-mean normal with covariance matrix V_1 ; $n^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically zero-mean normal with covariance matrix V_2 .*
3. *(Efficiency) The two-stage estimator is more efficient than the one-stage estimator, and the efficiency gain depends on the accuracy of estimation on the right tail of $F_{\tilde{T}}(C|\mathbf{Y}; \hat{\boldsymbol{\theta}}, \hat{\eta})$ beyond the censoring time C for censored subjects.*

This theorem is developed for the one- and two- stage estimators described in Section 3.2, which corresponds to the case in Section 3.3 when penalty parameter $\lambda_{\beta} = \lambda_{\xi} = 0$.

Since the model parameters for uncensored patients ($\boldsymbol{\theta}_1$) and “long-term survivors” (LTS) ($\boldsymbol{\theta}_2$) are separately specified, it is convenient to view uncensored patients and LTS as a unified group for the parameter estimation of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Therefore, for notation simplicity, we define $\Delta_i = \delta_i 1\{\tilde{T}_i > \tau\}$ to indicate patients in this group. Denote $\hat{G}_i(\boldsymbol{\theta}; u) = F_{\tilde{T}}(u|\mathbf{Y}_i; \boldsymbol{\theta}, \hat{\eta}) = 1 - \int_u^\infty f(\mathbf{Y}_i|t; \boldsymbol{\theta}) d\hat{\eta}(t)$ to be the conditional cumulative density function of death time \tilde{T}_i given observed medical cost data \mathbf{Y}_i , the unknown parameter $\boldsymbol{\theta}$, and estimated marginal distribution of survival $\hat{\eta}(\cdot)$. Denote $G_{0i}(\boldsymbol{\theta}; u) = F_{\tilde{T}}(u|\mathbf{Y}_i; \boldsymbol{\theta}, \eta_0)$. Recall that $l(\mathbf{Y}_i, u; \boldsymbol{\theta}) = \log f(\mathbf{Y}_i|\tilde{T}_i = u; \boldsymbol{\theta}) f_{\tilde{T}}(u)$ to be the log-likelihood for subject i . Recall that one-stage and two-stage estimators are denoted as $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ respectively. Analogous to equation (2) in main text, the log-likelihood of the two-stage estimator $\tilde{\boldsymbol{\theta}}$ can be decomposed into a summation of two sub-likelihoods given one-stage estimator $\hat{\boldsymbol{\theta}}$ and Kaplan-Meier estimate of the marginal survival function $\hat{\eta}(\cdot)$:

$$\begin{aligned}
L_n(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) &:= L_n^u(\boldsymbol{\theta}) + L_n^c(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) \\
&\propto n^{-1} \sum_{i=1}^n \left\{ \Delta_i l(\mathbf{Y}_i, T_i; \boldsymbol{\theta}) \right\} + n^{-1} \sum_{i=1}^n \left\{ (1 - \Delta_i) \frac{\int_{T_i}^\infty l(\mathbf{Y}_i, u; \boldsymbol{\theta}) dF_{\tilde{T}}(u|\mathbf{Y}_i; \hat{\boldsymbol{\theta}}, \hat{\eta})}{1 - F_{\tilde{T}}(T_i|\mathbf{Y}_i; \hat{\boldsymbol{\theta}}, \hat{\eta})} \right\} \quad (\text{A.1}) \\
&= n^{-1} \sum_{i=1}^n \left\{ \Delta_i l(\mathbf{Y}_i, T_i; \boldsymbol{\theta}) \right\} + n^{-1} \sum_{i=1}^n \left\{ (1 - \Delta_i) \frac{\int_{T_i}^\infty l(\mathbf{Y}_i, u; \boldsymbol{\theta}) d\hat{G}_i(\hat{\boldsymbol{\theta}}; u)}{1 - \hat{G}_i(\hat{\boldsymbol{\theta}}; T_i)} \right\}
\end{aligned}$$

The first part is the “unified” data log-likelihood of uncensored patients and LTS, and the second part is the expectation of “censored” data log-likelihoods of patients who are censored prior to τ , where

$$\begin{aligned}
L_n^u(\boldsymbol{\theta}) &\propto n^{-1} \sum_{i=1}^n \Delta_i l(\mathbf{Y}_i, T_i; \boldsymbol{\theta}) \\
L_n^c(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) &\propto n^{-1} \sum_{i=1}^n (1 - \Delta_i) \frac{\int_{T_i}^\infty l(\mathbf{Y}_i, u; \boldsymbol{\theta}) d\hat{G}_i(\hat{\boldsymbol{\theta}}; u)}{1 - \hat{G}_i(\hat{\boldsymbol{\theta}}; T_i)} \quad (\text{A.2})
\end{aligned}$$

Denote the pseudo score equation be $U_n(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) = \partial L_n(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) / \partial \boldsymbol{\theta} = U_n^u(\boldsymbol{\theta}) + U_n^c(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta})$

with $U_n^u(\boldsymbol{\theta}) = \partial L_n^u(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and $U_n^c(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}) = \partial L_n^c(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta})/\partial \boldsymbol{\theta}$. Score equations for a subject with observation \mathbf{Y}, δ, T are denoted as $U(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta}), U^u(\boldsymbol{\theta})$ and $U^c(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta})$. We work under the following regularity conditions:

- A1. $f(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})$ are distinct with respect to $\boldsymbol{\theta}$;
- A2. The true value $\boldsymbol{\theta}_0$ is an interior of open parameter space $\boldsymbol{\Theta}$;
- A3. The set $A = \{(\mathbf{Y}, \tilde{T}) : f(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta}) > 0\}$ is independent of $\boldsymbol{\theta}$;
- A4. The derivatives $f'(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})$, $f''(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})$ and $f'''(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})$ exist for all $(\mathbf{Y}, \tilde{T}) \in A$, all elements in $f'''(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})$ are continuous, and the corresponding derivatives of the integral $\int f(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})$ exist and can be obtained by differentiating under the integral sign;
- A5. For all (\mathbf{Y}, \tilde{T}) in set A , there exist functions $M_{ijk}(\mathbf{Y}, \tilde{T})$ and a positive number c , such that $|\frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log f(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})| < M_{ijk}(\mathbf{Y}, \tilde{T})$ for all $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 < c$, where $E[M_{ijk}(\mathbf{Y}, \tilde{T})|\boldsymbol{\theta}] < \infty$ for all i, j, k .
- A6. $-E\{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\mathbf{Y}|\tilde{T}; \boldsymbol{\theta})\}$ is positive definite with all elements finite.
- A7. Censoring time C is independent to death time \tilde{T} and response Y ;
- A8. Maximum follow-up time is bounded: $\tau < \infty$.

Denote the survival function of censoring time be $\bar{F}_C(\cdot) = 1 - F_C(\cdot)$, then from condition A7, $\bar{F}_C(\cdot)$ is independent of $\dot{l}(\boldsymbol{\theta}; \mathbf{Y}, \tilde{T})$. For a random sample with observation $(\mathbf{Y}, \tilde{T} = T, \Delta)$, we have $U^u(\boldsymbol{\theta}) = \Delta \dot{l}(\boldsymbol{\theta}; \mathbf{Y}, \tilde{T}) = 0$ at $\boldsymbol{\theta}_0$. Thus from conditions A1-A4, a consistent estimator of $\boldsymbol{\theta}_0$ can be solved from the score equation $U_n^u(\boldsymbol{\theta}) = 0$. (Lehmann (2004) Theorem 7.5.1) In addition to conditions A1-A4, we add conditions A5-A6, then

from Taylor expansion, the one-stage estimator can be approximately expressed as a summation of independent and identically distributed representation,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = n^{-1} \sum_{i=1}^n \boldsymbol{\gamma}_i + o_p(n^{-1/2}), \quad (\text{A.3})$$

where $\boldsymbol{\gamma}_i = -[E\{\partial U^u(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T\} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}]^{-1} \left\{ \Delta_i \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, T_i) \right\}$. Therefore, $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with variance $V_1 = E\{\boldsymbol{\gamma}_i^{\otimes 2}\}$ (Lehmann (2004) Theorem 7.5.2). The consistency of $\hat{\boldsymbol{\theta}}$ together with the uniform consistency of Kaplan-Meier estimate $\hat{\eta}(t)$ (Fleming and Harrington (1991) Theorem 3.4.2) leads to the unbiased estimator of conditional distribution $\hat{G}(\hat{\boldsymbol{\theta}}; t) = F_{\hat{T}}(t | \mathbf{Y}, \hat{\boldsymbol{\theta}}, \hat{\eta})$. From Taylor expansion

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \mathbf{D}^{-1} U_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}, \hat{\eta}) + o_p(n^{-1/2}), \quad (\text{A.4})$$

where $\mathbf{D} = \lim_{n \rightarrow \infty} \left\{ \frac{\partial U_n(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \hat{\eta})}{\partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}$.

By the consistency of $\hat{G}(\hat{\boldsymbol{\theta}}; t)$ and $\hat{\boldsymbol{\theta}}$, $U_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}, \hat{\eta})$ converges to zero in probability. It then follows the consistency of $\tilde{\boldsymbol{\theta}}$.

Denote the first and second derivative of the $l(\mathbf{Y}_i, T_i; \boldsymbol{\theta})$ of subject i with respect to $\boldsymbol{\theta}$ to be $\dot{l}(\boldsymbol{\theta}; \mathbf{Y}_i, \tilde{T}_i)$ and $\ddot{l}(\boldsymbol{\theta}; \mathbf{Y}_i, \tilde{T}_i)$. To study the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$, we first consider the difference between the score function with true value $\boldsymbol{\theta}_0$ and the score function with

the consistent estimator $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned}
& \sqrt{n} \left\{ U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \hat{\eta}) - U_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}, \hat{\eta}) \right\} \\
& \propto \frac{1}{\sqrt{n}} \sum_{i=1}^n \Delta_i \left(\dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) - \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) \right) \\
& \quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - \Delta_i) \left\{ \frac{\int_{T_i}^{\infty} \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) d\hat{G}_i(\boldsymbol{\theta}_0; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} - \frac{\int_{T_i}^{\infty} \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) d\hat{G}_i(\hat{\boldsymbol{\theta}}; u)}{1 - \hat{G}_i(\hat{\boldsymbol{\theta}}; T_i)} \right\} \\
& = -\mathbf{B}_1 \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
& \quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - \Delta_i) \left\{ \frac{\int_{T_i}^{\infty} \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) d\hat{G}_i(\boldsymbol{\theta}_0; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} - \frac{\int_{T_i}^{\infty} \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) d\hat{G}_i(\boldsymbol{\theta}_0; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} \right. \\
& \quad \quad + \frac{\int_{T_i}^{\infty} \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) d\hat{G}_i(\boldsymbol{\theta}_0; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} - \frac{\int_{T_i}^{\infty} \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) d\hat{G}_i(\hat{\boldsymbol{\theta}}; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} \\
& \quad \quad \left. + \frac{\int_{T_i}^{\infty} \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) d\hat{G}_i(\hat{\boldsymbol{\theta}}; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} - \frac{\int_T^{\infty} \dot{l}(\hat{\boldsymbol{\theta}}; \mathbf{Y}_i, u) d\hat{G}_i(\hat{\boldsymbol{\theta}}; u)}{1 - \hat{G}_i(\hat{\boldsymbol{\theta}}; T_i)} \right\} \\
& = -\mathbf{B}_1 \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \mathbf{I} - \mathbf{II} - \mathbf{III},
\end{aligned}$$

where $\mathbf{B}_1 = -E \left\{ \Delta \ddot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, T) \right\}$. We simplify the term **I**, **II**, **III**. From delta method, term **I**, **II** can be shown to be asymptotically linear in $\hat{\boldsymbol{\theta}}$,

$$\mathbf{I} = \mathbf{B}_2 \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1),$$

$$\mathbf{II} = \mathbf{B}_3 \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1),$$

$$\mathbf{III} = \mathbf{B}_4 \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1),$$

where

$$\begin{aligned}\mathbf{B}_2 &= E \left\{ (1 - \Delta) \frac{\int_T^\infty \ddot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) dG_0(\boldsymbol{\theta}_0; u)}{1 - G_0(\boldsymbol{\theta}_0; T)} \right\} \\ \mathbf{B}_3 &= E \left\{ (1 - \Delta) \frac{\int_T^\infty \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) f'(Y|u; \boldsymbol{\theta}_0) d\eta_0(u)}{1 - G_0(\boldsymbol{\theta}_0; T)} \right\} \\ \mathbf{B}_4 &= E \left\{ (1 - \Delta) \frac{\left(\int_{T_i}^\infty \ddot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) dG_0(\boldsymbol{\theta}_0; u) \right) \left(\int_{T_i}^\infty f'(\mathbf{Y}_i|u; \boldsymbol{\theta}_0) d\eta_0(u) \right)}{(1 - G_0(\boldsymbol{\theta}_0; T))^2} \right\}\end{aligned}$$

Combining the above simplified forms for **I**, **II**, **III**, we find that $U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \hat{\eta})$ is asymptotically linear in $\hat{\boldsymbol{\theta}}$,

$$\sqrt{n}U_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}, \hat{\eta}) = \sqrt{n}U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \hat{\eta}) + \mathbf{B}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1),$$

where $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3 + \mathbf{B}_4$. Thus from Taylor expansion

$$\begin{aligned}\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 &= \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \mathbf{D}^{-1}U_n(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}, \hat{\eta}) \\ &= (\mathbf{I}_p + \mathbf{D}^{-1}\mathbf{B})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{D}^{-1}U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \hat{\eta}) + o_p(n^{-1/2}),\end{aligned}\tag{A.5}$$

where \mathbf{I}_p is an $p \times p$ identity matrix, and p is prespecified.

In addition, it follows from the martingale representation of the Kaplan-Meier estimator (Fleming and Harrington (1991), p.98) that, almost surely

$$\frac{\int_{T_i}^\infty \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) d\hat{G}_i(\boldsymbol{\theta}_0; u)}{1 - \hat{G}_i(\boldsymbol{\theta}_0; T_i)} - \frac{\int_{T_i}^\infty \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) dG_{0i}(\boldsymbol{\theta}_0; u)}{1 - G_{0i}(\boldsymbol{\theta}_0; T_i)} = n^{-1} \sum_{j=1}^n \int_0^\infty \xi_i(t) dM_j(t) + o_p(n^{-1/2})$$

where $M_i(t) = \Delta_i I(T_i \leq t) - \int_0^t I(T_i \geq u) d\Lambda(u)$, $\Lambda(u)$ is the cumulative hazard function.

$\xi_i(t)$ is some random process. Therefore,

$$U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \hat{\eta}) = U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \eta_0) + n^{-1} \sum_{i=1}^n \int_0^\infty \bar{\xi}(t) dM_i(t) + o_p(n^{-1/2})$$

where $\bar{\xi}(t)$ is the limit of $n^{-1} \sum_{i=1}^n (1 - \Delta_i) \xi_i(t)$. Then we have

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{\mathbf{I}_p + \mathbf{D}^{-1}\mathbf{B}}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{\gamma}_i + \mathbf{D}^{-1}U_n(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \eta_0) + \frac{\mathbf{D}^{-1}}{\sqrt{n}} \sum_{i=1}^n \int_0^\infty \bar{\xi}(t) dM_i(t) + o_p(1)$$

Denote

$$\begin{aligned} \boldsymbol{\zeta}_i = & (\mathbf{I}_p + \mathbf{D}^{-1}\mathbf{B})\boldsymbol{\gamma}_i + \mathbf{D}^{-1} \left\{ \Delta_i \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, T_i) + (1 - \Delta_i) \frac{\int_{T_i}^\infty \dot{l}(\boldsymbol{\theta}_0; \mathbf{Y}_i, u) dF_{\tilde{T}}(u | \mathbf{Y}_i; \boldsymbol{\theta}_0, \eta_0)}{1 - F_{\tilde{T}}(T_i | \mathbf{Y}_i; \boldsymbol{\theta}_0, \eta_0)} \right\} \\ & + \mathbf{D}^{-1} \int_0^\infty \bar{\xi}(t) dM_i(t) \end{aligned}$$

then multivariate central limit theorem implies that $\tilde{\boldsymbol{\theta}}$ is asymptotically normally distributed with variance $V_2 = E\boldsymbol{\zeta}_i^{\otimes 2}$.

Lastly, we show that two-stage estimator is asymptotically more efficient than one-stage estimator. $K = -E\{\partial U^u(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$, $L = -E\{\partial U^c(\boldsymbol{\theta}; \boldsymbol{\theta}_0, \eta_0)/\partial \boldsymbol{\theta}^T |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}$, $M = U^u(\boldsymbol{\theta}_0)$, $N = U^c(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0, \eta_0)$, then we have $K > 0$, $L \geq 0$, $M = N = 0$, and $EMM^T < 0$, $ENN^T \leq 0$.

$$(K + L)^{-1} = K^{-1} - L^{-1}(L^{-1} + K^{-1})L^{-1} \leq K^{-1}$$

Therefore,

$$\begin{aligned}
& (K + L)^{-1}\{E(M + N)(M + N)^T\}(K + L)^{-1} \\
& \leq K^{-1}\{E(M + N)(M + N)^T\}K^{-1} \\
& = K^{-1}\{EMM^T + ENN^T\}K^{-1} \\
& \leq K^{-1}\{EMM^T\}K^{-1}
\end{aligned}$$

A.2. Sensitivity analysis on misspecification of the data distribution

We conducted a simulation study to evaluate robustness of the proposed method under misspecification of the data distribution including misspecification of covariance matrix as well as violation of multivariate normal assumption.

The baseline scenario corresponds to the setting described in the main text. The censoring rate is 30% and the sample size is 500. First, we investigate the robust performances of the proposed estimators under misspecification of covariance structures. In the baseline scenario, the error $\epsilon_i \sim N(\mathbf{0}, \sigma^2 \Sigma_i)$ where $\sigma = 2$. We change Σ_i to be an $n_i \times n_i$ matrix, so that the correlation structure is misspecified. Under Scenario (1), Σ_i has AR-1 correlation structure, with an autocorrelation coefficient 0.8. Under Scenario (2), the diagonal elements of Σ_i is equal to $t_{ij}^{1/3}$ and other elements equal to 0.5. The Scenario (2) has practical meaning that the variation of medical cost often increases along with the measurement time. Second, we investigate the robust performances of the proposed method under violation of normal assumption Under Scenario (3), the error follows a mixture of two Normal distributions $\epsilon_{ij} \sim .5N(5, \sigma^2) + .5N(-5, \sigma^2)$; Under Scenario (4) the error follows a gamma distribution with shape=10, rate=1 and center around 0. We further simulate the response Y_{ij} to take value 0 with 10% probability (Scenario (5)).

Table A.1: Compare the descriptive, one-stage (1-stg) and two-stage (2-stg) estimates under model misspecification. A-AB: aggregated sample absolute bias; A-SD: aggregated empirical standard deviations; A-MSE: aggregated sample mean squared error. Results are averages from 1,000 mean point estimates of simulated samples. Runtime is in second.

Scenario	A-AB (desc, 1-stg, 2-stg)	A-SD (desc, 1-stg, 2-stg)	A-MSE (desc, 1-stg, 2-stg)	Runtime (seconds) (desc, 1-stg, 2-stg)
Baseline	(0.011, 0.050, 0.069)	(0.453, 0.311, 0.251)	(0.221, 0.109, 0.076)	(0.1, 7.9, 36.3)
(1)	(0.012, 0.037, 0.063)	(0.452, 0.347, 0.304)	(0.219, 0.133, 0.109)	(0.1, 4.8, 22.9)
(2)	(0.013, 0.053, 0.073)	(0.519, 0.346, 0.281)	(0.296, 0.139, 0.096)	(0.1, 4.7, 20.8)
(3)	(0.018, 0.050, 0.081)	(0.670, 0.361, 0.284)	(0.495, 0.149, 0.102)	(0.1, 5.4, 24.8)
(4)	(0.029, 0.100, 0.146)	(1.156, 0.488, 0.384)	(1.465, 0.285, 0.205)	(0.1, 5.5, 25.7)
(5)	(0.738, 0.749, 0.759)	(0.647, 0.384, 0.308)	(1.025, 0.753, 0.724)	(0.1, 4.8, 23.2)

Unless stated otherwise, the simulation settings are identical to the baseline scenario. Each scenario comprised $M = 1,000$ Monte Carlo replications. We applied the proposed method to data generated from the scenarios above, and the results are presented in Table A.1.

The simulation results for Scenario (1) and (2) show that under misspecification of variance-covariance structure, the A-ABs and A-SDs of our proposed one-stage and two stage estimators are appreciably small, and the computational burden is also small. Scenario (3) and (4) produce similar results when right-skewed and two-peak errors are misspecified as being normal. Therefore, the proposed estimators are robust against both misspecification of covariance structure and violation of normal assumption such as skewness and bimodality. Biases increase as the proportion of zero costs increases, which indicates the importance of model checking before using the two-stage method.

A.3. Model diagnosis on renal cell cancer care cost data

In this section, we present model diagnosis for the data application in Chapter 2 Section 2.4. Figure A.1 is a descriptive analysis averaging the pointwise log transformed cost trajectory stratified by survival year. First, the gray dots on the background represents the cost data from a random sample of 30 uncensored patients. These dots show a

substantial variation of medical cost data. Second, when there are adequate amount of cost data, the black curves have a smooth U-shape for survival years at 2, 4, 6, 8; the black curves have an L-shape for the LTS group. When the survival is close to the end of the maximum follow-up, the cost data become sparse, and we observe local variations. Therefore, we specify the LTS group in the data analysis in Chapter 2 Section 4, and apply second order penalties on two directions to achieve smooth cost trajectory estimates. Figure A.2 shows the Kaplan-Meier curves with a substantial difference between the survival probability of patients in local or regional stage and that for patients in distant stage. The histograms of log transformed cost by survival for uncensored patients suggest that the normal assumption holds for the quarterly cost data.

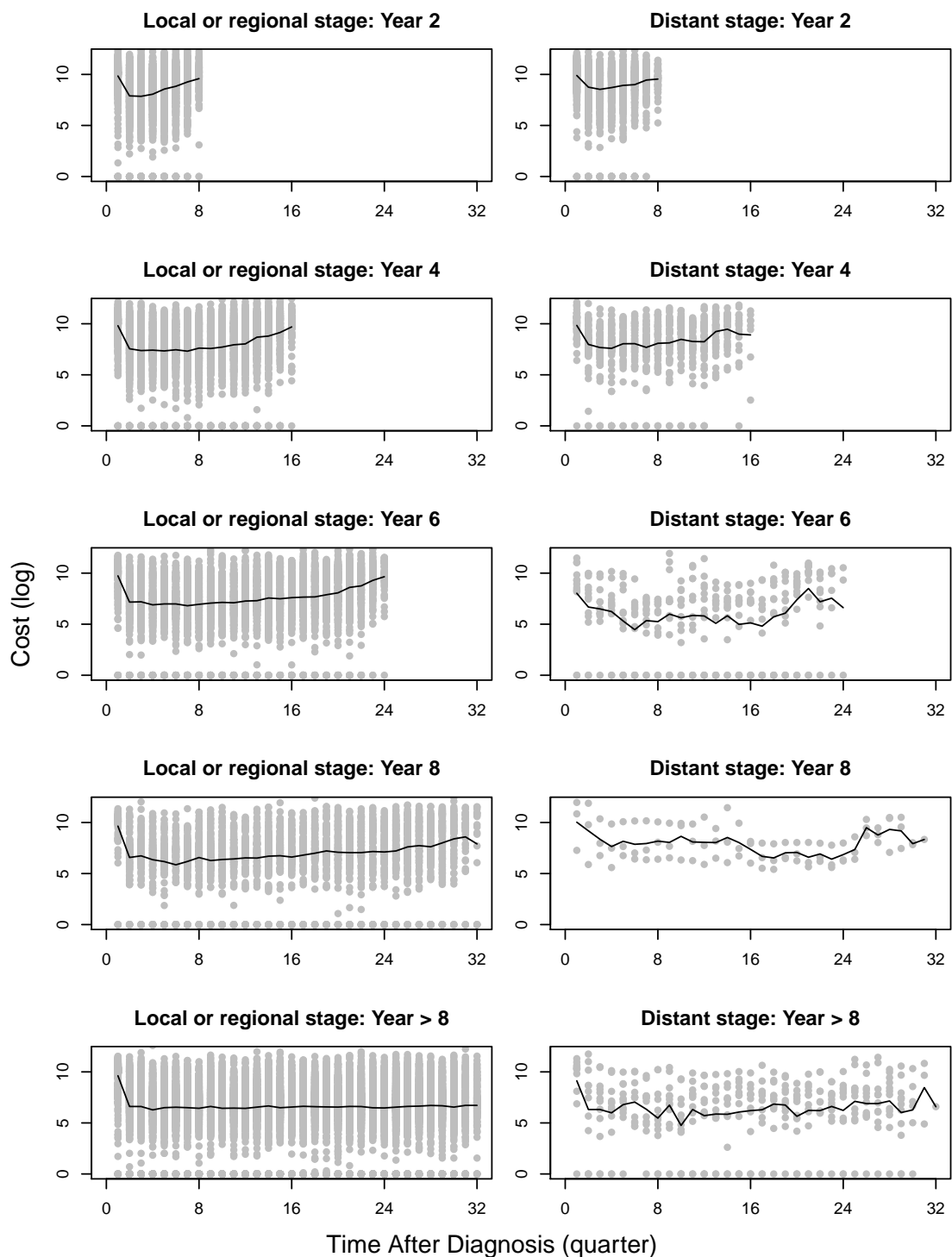


Figure A.1: Averaged log transformed quarterly cost stratified by year of survival. The black lines are the average quarterly costs among all the uncensored patients in each stratum. A random sample of 30 uncensored patients was selected from each stratum and their individual monthly cost trajectories are illustrated by the gray dots in the background.

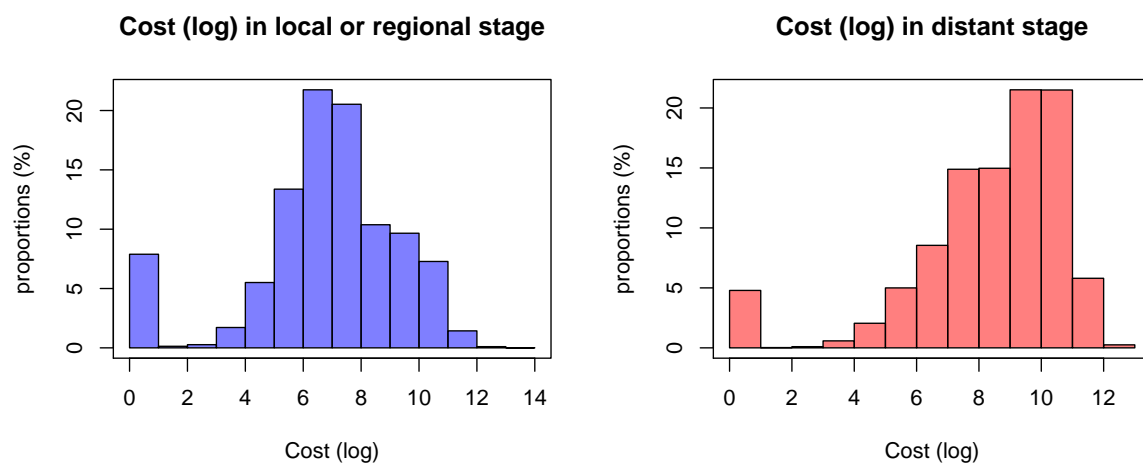
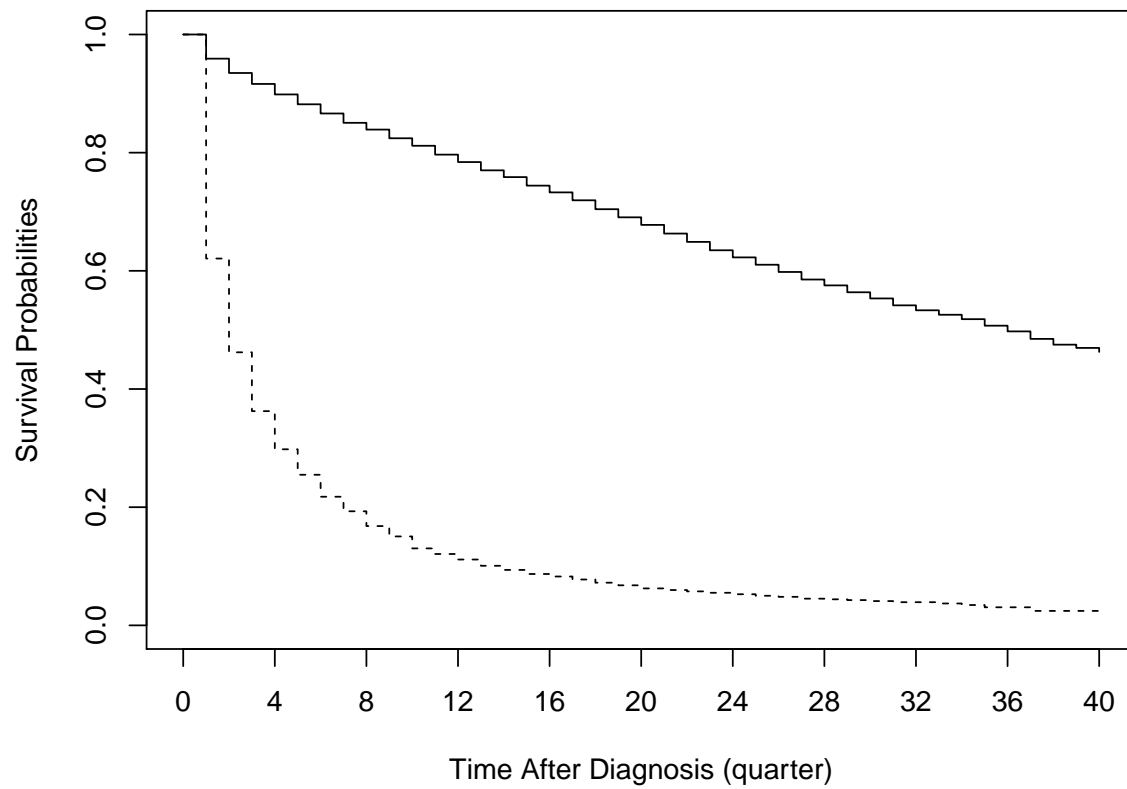


Figure A.2: Top panel: Kaplan-Meier estimates of the survival probabilities of renal cell cancer patients, solid curve: local or regional stage; dashed curve: distant stage; bottom panel: histograms of log transformed cost by cancer stage.

APPENDIX B

APPENDIX FOR CHAPTER 3

This Appendix includes three sections. In Appendix B.1, we present estimators and associated proofs. In Appendix B.2, we report coarsening approximation of survival and model checking results from an analysis of the prostate cancer cost data. In Appendix B.3, we provide additional simulation studies to explore the performance of proposed methods with (i) different censoring proportions, (ii) different zero proportions, and (iii) different skewness.

B.1. Asymptotic properties

Since the incidence cost is collected in a given time unit, and the maximum follow-up time $\tau < \infty$, the measurement time in our problem is finite and bounded, and the number of knots grows slower than the sample size n , and the maximum model flexibility is achieved when $K_j = \tau, j = 1, 2, 3$. We show that similar to the one-dimensional case (Chen and Wang, 2011; Chen et al., 2013), the asymptotics for the multi-dimensional case in the presence of right-censoring meet the small knots scenario analogous to Zhu et al. (2008). We show that the asymptotic variance is minimized when the true covariance is used (similar to Welsh et al. (2002)), and when the penalty parameter converges to infinity at a certain rate. The asymptotic normality of the proposed estimator is also proved.

Let $\text{cov}(\mathbf{Y}_i) = \Sigma$ be the true covariance matrix which does not vary across subjects; and V be the working covariance matrix of \mathbf{Y}_i . We assume that $(t, s) \in \mathcal{T}_1 \times \mathcal{T}_2$ with $\mathcal{T}_1 = [0, \tau]$ and $\mathcal{T}_2 = [0, \infty)$. For subjects who died prior to maximum follow-up time τ , $(t, s) \in \mathcal{T}_1^2$; and for LTS $\mathcal{T}_2 \setminus \mathcal{T}_1$ represent a mass domain i.e., for any $s \in (\tau, \infty)$,

$\mu(t, s)$ share the same function $\mu(t, \tau+) = \eta(t), t \in \mathcal{T}_1$. $\mathbf{B}_j(t) = [B_{-q_j}(t), \dots, B_{K_j}(t)]$ is the B-spline bases of order q_j with K_j knots $\{0 = \kappa_{j1} < \kappa_{j2} < \dots < \kappa_{j, K_j+1} = \tau\}$ for $j = 1, 2, 3$.

Let $V^{-1} = (v^{st})$ and $G = (g_{ij})$ with

$$g_{ij} = \sum_{s \neq t}^m \int_0^\infty \int_0^\infty \int_0^{y_2} \int_0^{y_1} B_\mu(x_1, y_1) v^{st} B_\mu(x_2, y_2) \rho_{st}(x_1, y_1, x_2, y_2) dx_1 dx_2 dy_1 dy_2 \\ + \sum_{s=1}^m \int_0^\infty \int_0^y B_\mu(x, y) v^{ss} B_\mu(x, y) \rho_s(x, y) dx dy.$$

Let $W = \{w_{ij}\} = V^{-1} \Sigma V^{-1}$ and $U = \{u_{ij}\}$ with

$$u_{ij} = \sum_{s \neq t}^m \int_0^\infty \int_0^\infty \int_0^{y_2} \int_0^{y_1} B_\mu(x_1, y_1) w^{st} B_\mu(x_2, y_2) \rho_{st}(x_1, y_1, x_2, y_2) dx_1 dx_2 dy_1 dy_2 \\ + \sum_{s=1}^m \int_0^\infty \int_0^y B_\mu(x_1, y_1) w^{ss} B_\mu(x, y) \rho_s(x, y) dx dy.$$

Denote $Q_{\Delta, n, jl}(x_1, y_1, x_2, y_2) = \frac{1}{n} \sum_{i=1}^n \Delta_i I(t_{i,j} \leq x_1, s_{i,j} \leq y_1, t_{i,l} \leq x_2, s_{i,l} \leq y_2)$, $Q_{\Delta, n, j}(x, y) = \frac{1}{n} \sum_{i=1}^n \Delta_i I(t_{i,j} \leq x, s_{i,j} \leq y)$; $Q_{n, jl}(x_1, y_1, x_2, y_2) = \frac{1}{n} \sum_{i=1}^n \Delta_i I(t_{i,j} \leq x_1, s_{i,j} \leq y_1, t_{i,l} \leq x_2, s_{i,l} \leq y_2) + (1 - \Delta_i) I(t_{i,j} \leq x_1, t_{i,l} \leq x_2) f(s_{i,j} \leq y_1) f(s_{i,l} \leq y_2)$, $Q_{n, j}(x, y) = \frac{1}{n} \sum_{i=1}^n \Delta_i I(t_{i,j} \leq x, s_{i,j} \leq y) + (1 - \Delta_i) I(t_{i,j} \leq x) f(s_{i,j} \leq y)$, $Q_{jl}(x_1, y_1, x_2, y_2)$ and $Q_j(x, y)$ are certain distribution functions with positive continuous density functions $\rho_{jl}(x_1, y_1, x_2, y_2)$ and $\rho_j(x, y)$ on \mathcal{T}^4 and \mathcal{T}^2 respectively.

Denote the approximation bias as $b_{a2}(t, s), b_{a1}(t, s), b_{a3}(t)$. Similar to Zhu et al.

(2008) and Xiao (2019), we use $\mu^{(i,j)}(t, s)$ to denote $\partial^{i+j}\mu(t, s)/\partial t^i \partial s^j$,

$$b_{a1}(t, s) = -\frac{\mu^{(q_1,0)}(\kappa_{1,i}, s)}{q_1!} \sum_{i=0}^{K_1} I(\kappa_{1,i} \leq t \leq \kappa_{1,i+1}) h_{1i}^{q_1} B_{p_1+1}\left(\frac{t - \kappa_{1,i}}{\delta_i}\right),$$

$$b_{a2}(t, s) = -\frac{\mu^{(0,q_2)}(t, \kappa_{2,i})}{q_2!} \sum_{i=0}^{K_2} I(\kappa_{2,i} \leq s \leq \kappa_{2,i+1}) h_{2i}^{q_2} B_{p_2+1}\left(\frac{s - \kappa_{2,i}}{\delta_i}\right),$$

where $B_{p_j+1}(t)$ is the $(p_j + 1)$ th Bernoulli polynomial (Barrow and Smith, 1978). $\mathbf{V} = \text{diag}(V, \dots, V)$ and $\mathbf{\Sigma} = \text{diag}\{\Sigma, \dots, \Sigma\}$. From the definition we have $P_{12} = \lambda_1 G_2 \otimes P_1 + \lambda_2 P_2 \otimes G_1$ and $P_3 = \lambda_3 P_3$ where $G_j = \int_x B_j(x) B_j^T(x) dx$, $P_j = \int_x B_j^{(q_j)}(x) B_j^{(q_j),T}(x) dx$, $j = 1, 2, 3$, and $B_j^{(q_j)}(x)$ denotes q_j th derivative of $B_j(x)$. Joint penalty matrix P has diagonal blocks P_{12} and P_3 .

B.1.1. Asymptotic behavior of $\tilde{\mu}(t, s)$

The estimator $\tilde{\mu}(t, s)$ depends on uncensored patients and LTS only, and data from patients who were censored prior to τ is not used. Let B be the joint design matrix of uncensored patients and LTS. Denote $n_\Delta = \sum_{i=1}^n \Delta_i$, $G_{\Delta,n} = \frac{1}{n_\Delta} \sum_{i=1}^n \Delta_i \mathbf{B}_{\mu i}^T V^{-1} \mathbf{B}_{\mu i}$, $H_{\Delta,n} = G_{\Delta,n} + P$, and $U_{\Delta,n} = \frac{1}{n_\Delta} \sum_{i=1}^n \Delta_i \mathbf{B}_{\mu i}^T V^{-1} \Sigma V^{-1} \mathbf{B}_{\mu i}$. Denote the shrinkage bias as $b_{\Delta,\lambda}(t, s, V)$ where

$$b_{\Delta,\lambda}(t, s, V) = -\frac{1}{n_\Delta} B_{12}^T(t, s) \left(G + \frac{P}{n_\Delta}\right)^{-1} P \psi_\Delta$$

where $\psi_\Delta = (B^T \mathbf{V}^{-1} B)^{-1} B^T \mathbf{V}^{-1} s_{\Delta,\mu}/n_\Delta$, and $s_{\Delta,\mu}(\cdot) = B_{12}^T(\cdot) \psi_\Delta$ is the best L_∞ approximation to the function $\mu(t, s)$. We provide some regularity assumptions.

Assumption B1. Let $h_{ji} = \kappa_{j,i+1} - \kappa_{j,i}$, $h_j = \max_{0 \leq i \leq K_j} h_{ji}$, there exists a constant $M_j > 0$, such that $h_j / \min_{0 \leq i \leq K_j} h_{ji} \leq M_j$ and $h_j = O(K_j^{-1})$, $j = 1, 2$.

Assumption B2. For any $j, l = 1, \dots, m$,

$$\sup_{(x_1, y_1), (x_2, y_2) \in \mathcal{T}_1^2} |Q_{\Delta, n, jl}(x_1, y_1, x_2, y_2) - Q_{jl}(x_1, y_1, x_2, y_2)| = o(h)$$

$$\sup_{(x, y) \in \mathcal{T}_1^2} |Q_{\Delta, n, j}(x, y) - Q_j(x, y)| = o(h)$$

Assumption B3. The number of knots $K_j = o(n_\Delta), j = 1, 2$.

Assumption B4. The eigenvalues of the working covariance V and true covariance Σ are bounded away from zero.

Theorem B1. Under the regularity assumptions B1-B4, $(t, s) \in \mathcal{T}^2$, $\lambda_j K_j^{2q_j} = o(n_\Delta), j = 1, 2$, the following statements hold

$$\begin{aligned} E[\tilde{\mu}(t, s)] - \mu(t, s) &= b_{a1}(t, s) + b_{a2}(t, s) + b_{\Delta, \lambda}(t, s, V) \\ &\quad + o(K_1^{-q_1} + K_2^{-q_2}) + o\left(\frac{\lambda_1 K_1^{q_1} + \lambda_2 K_2^{q_2}}{n_\Delta}\right) \\ \text{var}[\tilde{\mu}(t, s)] &= \frac{1}{n_\Delta} B_{12}^T(t, s) \left(G + \frac{P}{n_\Delta}\right)^{-1} U \left(G + \frac{P}{n_\Delta}\right)^{-1} B_{12}(t, s) + o\left(\frac{K_1 K_2}{n_\Delta}\right) \\ \text{MSE}[\tilde{\mu}(t, s)] &= O\left(\frac{K_1 K_2}{n_\Delta}\right) + O\left(\frac{\lambda_1^2 K_1^{2q_1} + \lambda_2^2 K_2^{2q_2}}{n_\Delta^2}\right) + O(K_1^{-2p_1} + K_2^{-2p_2}), \end{aligned} \tag{B.1}$$

and when $\lambda_j = o((n K_1 K_2)^{1/2} K_j^{-q_j}), j = 1, 2$ there exists

$$\frac{\tilde{\mu}(t, s) - (\mu + b_{a1} + b_{a2})(t, s) - b_{\Delta, \lambda}(t, s, V)}{\sqrt{\text{var}[\tilde{\mu}(t, s)]}} \rightarrow N(0, 1)$$

in distribution, as $n \rightarrow \infty$.

As the LTS estimate does not affect the estimation of θ_{12} , the proof of Theorem B1 is clear following Chen et al. (2013) with similar arguments but distinguishing the dimension of basis to be two (Xiao, 2019).

B.1.2. Asymptotic behavior of $\hat{\mu}(t, s)$ when residual lifetime distribution is known

Let $D = (\mathbf{D}_1^T, \dots, \mathbf{D}_n^T)^T$ be the joint design matrix of all patients where $\mathbf{D}_i = \Delta_i \mathbf{B}_{\mu i} + (1 - \Delta_i) E \mathbf{B}_{\mu i}$; $\mathbf{B}_{\mu i} = (B_{\mu i 1}^T, \dots, B_{\mu i n_i}^T)^T$; $B_{\mu i j} = I(\delta_i = 1) \mathbf{B}_{\mu}(t_{ij}, T_i) + I(\delta_i = 0, T_i = \tau) \mathbf{B}_{\mu}(t_{ij}, \tau +)$; $E \mathbf{B}_{\mu i} = (E B_{\mu i 1}^T, \dots, E B_{\mu i n_i}^T)^T$; $E B_{\mu i j} = \frac{1}{1 - F_{\tilde{T}}(T_i)} \left(\int_{T_i}^{\tau} \mathbf{B}_{\mu}(t_{ij}, s) dF_{\tilde{T}}(s) + \mathbf{B}_{\mu}(t_{ij}, \tau +) [1 - F_{\tilde{T}}(\tau)] \right)$. Denote $G_n = \frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T V^{-1} \mathbf{D}_i$, $H_n = G_n + P$, and $U_n = \frac{1}{n} \sum_{i=1}^n \mathbf{D}_i^T V^{-1} \Sigma V^{-1} \mathbf{D}_i$.

Denote the shrinkage bias as

$$b_{\lambda}(t, s, V) = -\frac{1}{n} B_{12}^T(t, s) \left(G + \frac{P}{n}\right)^{-1} P \psi$$

where $\psi = (D^T V^{-1} D)^{-1} D^T V^{-1} s_{\mu}/n$, and $s_{\mu}(\cdot) = B_{\mu}^T(\cdot) \psi$ is the best L_{∞} approximation to the function $\mu(t, s)$.

We impose additional regularity assumptions.

Assumption B5. Let $h_{ji} = \kappa_{j,i+1} - \kappa_{j,i}$, $h_j = \max_{0 \leq i \leq K_j} h_{ji}$, there exists a constant $M_j > 0$, such that $h_j / \min_{0 \leq i \leq K_j} h_{ji} \leq M_j$ and $h_j = O(K_j^{-1})$, $j = 1, 2, 3$. $h_3 = O(h)$.

Assumption B6. For any $j, l = 1, \dots, m$,

$$\sup_{(x_1, y_1), (x_2, y_2) \in \mathcal{T}_1 \times \mathcal{T}_2} |Q_{n,jl}(x_1, y_1, x_2, y_2) - Q_{jl}(x_1, y_1, x_2, y_2)| = o(h)$$

$$\sup_{(x, y) \in \mathcal{T}_1 \times \mathcal{T}_2} |Q_{n,j}(x, y) - Q_j(x, y)| = o(h)$$

Assumption B7. The number of knots $K_j = o(n)$, $j = 1, 2, 3$.

Theorem B2. Under the regularity assumptions B1-B7, $(t, s) \in \mathcal{T}^2$, $\lambda_j K_j^{2q_j} =$

$o(n), j = 1, 2, 3$, the following statements hold

$$\begin{aligned}
E[\hat{\mu}(t, s)] - \mu(t, s) &= b_{a1}(t, s) + b_{a2}(t, s) + b_{\lambda}(t, s, V) \\
&\quad + o(K_1^{-q_1} + K_2^{-q_2}) + o\left(\frac{\lambda_1 K_1^{q_1} + \lambda_2 K_2^{q_2} + \lambda_3 K_3^{q_3}}{n}\right) \\
\text{var}[\hat{\mu}(t, s)] &= \frac{1}{n} B_{12}^T(t, s) (G + \frac{P}{n})^{-1} U (G + \frac{P}{n})^{-1} B_{12}(t, s) + o\left(\frac{K_1 K_2}{n}\right) \\
\text{MSE}[\hat{\mu}(t, s)] &= O\left(\frac{K_1 K_2}{n}\right) + O\left(\frac{\lambda_1^2 K_1^{2q_1} + \lambda_2^2 K_2^{2q_2} + \lambda_3^2 K_3^{2q_3}}{n^2}\right) + O(K_1^{-2p_1} + K_2^{-2p_2}),
\end{aligned} \tag{B.2}$$

and when $\lambda_j = o((nK_1K_2)^{1/2}K_j^{-q_j}), j = 1, 2, 3$ there exists

$$\frac{\hat{\mu}(t, s) - (\mu + b_{a1} + b_{a2})(t, s) - b_{\lambda}(t, s, V)}{\sqrt{\text{var}[\hat{\mu}(t, s)]}} \rightarrow N(0, 1)$$

in distribution, as $n \rightarrow \infty$.

The proof of Theorem B2 is given as follows. We first derive

$$\hat{\mu}(t, s) = \frac{1}{n} B_{12}^T(t, s) H_n^{-1} D^T \mathbf{V}^{-1} Y = \hat{\mu}_{\text{reg}}(t, s) - \frac{1}{n} B_{12}^T(t, s) H_n^{-1} (P/n) G_n^{-1} D^T \mathbf{V}^{-1} Y,$$

where $\hat{\mu}_{\text{reg}}(t, s) = \frac{1}{n} B_{12}^T(t, s) G_n^{-1} D^T \mathbf{V}^{-1} Y$. Then we consider the bias

$$[E\hat{\mu} - \mu](t, s) = [s_{\mu} - \mu](t, s) + [E\hat{\mu}_{\text{reg}} - s_{\mu}](t, s) - \frac{1}{n} B_{12}^T(t, s) H_n^{-1} (P/n) G_n^{-1} D^T \mathbf{V}^{-1} Y.$$

Xiao (2019) showed in Proposition 3.1 that $[s_{\mu} - \mu - b_{a1} - b_{a2}](t, s) = o(K_1^{-q_1} + K_2^{-q_2})$.

Analogous to the proof of Lemma A.8 in Xiao (2019), from Assumption 2 we have

$\|G_n^{-1}\|_{\max} = o(h^{-1})$, and thus the second term has order $o(K_1^{-q_1} + K_2^{-q_2})$. The third term

is given as

$$\begin{aligned}
& \frac{1}{n} B_{12}^T(t, s) H_n^{-1} (P/n) G_n^{-1} B^T \mathbf{V}^{-1} s_\mu = \frac{1}{n} B_{12}^T(t, s) H_n^{-1} (P/n) \boldsymbol{\theta} \\
& = \frac{1}{n} B_{12}^T(t, s) H^{-1} (P/n) \boldsymbol{\theta} + \frac{1}{n} B_{12}^T(t, s) (H_n^{-1} - H^{-1}) (P/n) \boldsymbol{\theta} \\
& = b_\lambda(t, s, V) + \frac{1}{n} B_{12}^T(t, s) (H_n^{-1} - H^{-1}) (P/n) \boldsymbol{\theta}.
\end{aligned}$$

The first $(p_1 + p_2)$ elements of $P\boldsymbol{\theta}$ are $(\lambda_1 G_2 \otimes P_1 + \lambda_2 P_2 \otimes G_1) \boldsymbol{\theta}_{12}$ with order $O(\lambda_1 n h_1^{-q_1} + \lambda_2 h_2^{-q_2})$ according to Theorem 5.1 in Xiao (2019), and the rest elements are $\lambda_3 P_3 \boldsymbol{\theta}_3$ with order $O(\lambda_3 h_3^{-q_3})$ from Theorem 1 in Chen and Wang (2011). Assumption 2 and the definition of P leads to $\|H^{-1}\|_{\max} = o(h^{-1})$, $\|H_n^{-1}\|_{\max} = o(h^{-1})$. We have $\frac{1}{n} B_{12}^T(t, s) (H_n^{-1} - H^{-1}) (P/n) \boldsymbol{\theta}$ and $\frac{1}{n} B_{12}^T(t, s) H_n^{-1} (P/n) G_n^{-1} D^T \mathbf{V}^{-1} (Y - s_\mu)$ both are asymptotically ignorable. Therefore, $[E\hat{\mu} - \mu](t, s) = [b_{a1} + b_{a2}](t, s) + b_\lambda(t, s, V) + o(K_1^{-q_1} + K_2^{-q_2}) + o(\lambda_1 n^{-1} h_1^{-q_1} + \lambda_2 n^{-1} h_2^{-q_2} + \lambda_3 n^{-1} h_3^{-q_3})$.

Next, the variance is given by

$$\begin{aligned}
\text{var}(\hat{\mu}(t, s)) &= \frac{1}{n_\Delta} B_{12}^T(t, s) H_n^{-1} U_n H_n^{-1} B_{12}(t, s) \\
&= \frac{1}{n_\Delta} B_{12}^T(t, s) \{H^{-1} U H^{-1} + H_n^{-1} (U_n - U) H_n^{-1} \\
&\quad + H^{-1} U (H_n^{-1} - H^{-1}) + (H_n^{-1} - H^{-1}) U H_n^{-1}\} B_{12}(t, s)
\end{aligned}$$

where the second, third and forth term can be shown to have order $o(n^{-1} h^{-1})$ by similar arguments for the bias. Since the shrinkage bias is negligible, the asymptotic bias does not depend on the choice of working covariance matrix, or the density function $Q(x)$. The asymptotic variance therefore is minimized when the true covariance is used, and the asymptotic MSE is minimized when $V = \Sigma$.

Next we prove the asymptotic normality of $\tilde{\mu}(t, s)$. From Assumption B3 and $\lambda_j =$

$o((nK_1K_2)^{1/2}K_j^{-q_j}), j = 1, 2, 3$ we have

$$\begin{aligned} & \frac{E[\hat{\mu}(t, s)] - (\mu + b_{a1} + b_{a2})(t, s) - b_\lambda(t, s, V)}{\text{var}(\tilde{\mu}(t, s))} \\ &= \frac{o(K_1^{-q_1} + K_2^{-q_2}) + o(n^{-1}(\lambda_1 K_1^{q_1} + \lambda_2 K_2^{q_2} + \lambda_3 K_3^{q_3}))}{o(\sqrt{n^{-1}K_1K_2})} = o(1) \end{aligned}$$

Therefore, it is sufficient to show that

$$\frac{\hat{\mu}(t, s) - E[\hat{\mu}(t, s)]}{\text{var}(\hat{\mu}(t, s))} \rightarrow N(0, 1)$$

in distribution. From definition $\hat{\mu}(t, s) = B_{12}^T(t, s)H_n^{-1} \sum_{i=1}^n D_i^T V^{-1} Y_i$ we can represent

$$\hat{\mu}(t, s) - E[\hat{\mu}(t, s)] = B_{12}^T(t, s)H_n^{-1} \sum_{i=1}^n D_i^T V^{-1} \epsilon_i = \sum_{i=1}^n C_{n,i} \epsilon_i$$

where $C_{n,i}^T = B_{12}^T(t, s)H_n^{-1} D_i^T V^{-1}$, and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^T$. It suffices to verify

$$\frac{\max_{1 \leq i \leq n} \|C_{n,i}\|^2}{\sum_{i=1}^n \|C_{n,i}\|^2} = o\left(\frac{1}{nh}\right)$$

where $\|C_{n,i}\|^2 = B_{12}^T(t, s)H_n^{-1} \{D_i^T V^{-2} D_i\} H_n^{-1} B_{12}^T(t, s)$. Therefore, the Linderberg condition holds, and this proves the theorem.

B.3. Asymptotic behavior of $\hat{\mu}(t, s)$ when residual lifetime distribution is unknown

In practice, the survival distribution is often estimated, i.e. $\hat{F}_{\hat{T}}(s) = 1 - \exp\{-\int_0^s \hat{h}(t)dt\}$ with $\hat{h}(t) = \exp\{B_h(t)\hat{\beta}\}$. By the result in Kauermann (2005) that under mild regularity assumptions, the penalized estimate $\hat{\beta}$ is consistent in the sense that $E\hat{\beta} - \beta = o(n^{-1/2})$. The next Corollary B1 shows the asymptotic properties when the survival distribution is unknown.

Corollary B1. Under the assumption of Theorem 2, the following statements hold

$$\begin{aligned}
E[\hat{\mu}_*(t, s)] - \mu(t, s) &= b_{a1}(t, s) + b_{a2}(t, s) + b_\lambda(t, s, V) + o(K_1^{-q_1} + K_2^{-q_2}) \\
&\quad + o\left(\frac{\lambda_1 K_1^{q_1} + \lambda_2 K_2^{q_2} + \lambda_3 K_3^{q_3}}{n}\right) + o(K_1^{-1} K_2^{-1} K_4^{-1} n^{-1/2}) \\
\text{var}[\hat{\mu}_*(t, s)] &= \frac{1}{n} B_{12}^T(t, s) (G_* + \frac{P_*}{n})^{-1} U_* (G_* + \frac{P_*}{n})^{-1} B_{12}(t, s) + o\left(\frac{K_1 K_2}{n}\right) \\
\text{MSE}[\hat{\mu}_*(t, s)] &= O\left(\frac{K_1 K_2}{n}\right) + O\left(\frac{\lambda_1^2 K_1^{2q_1} + \lambda_2^2 K_2^{2q_2} + \lambda_3^2 K_3^{2q_3}}{n^2}\right) \\
&\quad + O(K_1^{-2p_1} + K_2^{-2p_2}) + o(K_1^{-2} K_2^{-2} K_4^{-2} n^{-1}),
\end{aligned} \tag{B.3}$$

and when $\lambda_j = o((nK_1K_2)^{1/2}K_j^{-q_j})$, $j = 1, 2, 3$ there exists

$$\frac{\hat{\mu}_*(t, s) - (\mu + b_{a1} + b_{a2})(t, s) - b_\lambda(t, s, V)}{\sqrt{\text{var}[\hat{\mu}_*(t, s)]}} \rightarrow N(0, 1)$$

in distribution, as $n \rightarrow \infty$.

Let h_4 be the maximum length between adjacent knots for hazard estimation, the plug-in estimate of design matrix is consistent with $E\hat{D} - D = o(hh_4n^{-1/2})$. We have

$$\hat{\mu}_*(t, s) - \hat{\mu}(t, s) = \frac{1}{n} B_{12}^T(t, s) \{ \hat{H}_n^{-1} \hat{D}^{-1} - H_n^{-1} D^{-1} \} V^{-1} Y,$$

where $\hat{H}_n^{-1} \hat{D}^{-1} - H_n^{-1} D^{-1}$ has order $o(hh_4n^{-1/2})$. To account for the variation from hazard estimation, Let $G_* = (g_{*ij})$ and $V_*^{-1} = (v_*^{st})$ with

$$\begin{aligned}
g_{*ij} &= \sum_{s \neq t}^m \int_0^\infty \int_0^\infty \int_0^{y_2} \int_0^{y_1} B_{*,i}(x_1, y_1) v_*^{st} B_{*,j}(x_2, y_2) \rho_{st}(x_1, y_1, x_2, y_2) dx_1 dx_2 dy_1 dy_2 \\
&\quad + \sum_{s=1}^m \int_0^\infty \int_0^y B_{*,i}(x, y) v_*^{ss} B_{*,j}(x, y) \rho_s(x, y) dx dy,
\end{aligned}$$

where $B_*(x, y) = (B_\mu(x, y)^T, B_h(t)^T)^T$. Let $W_* = \{w_{*ij}\} = V_*^{-1}\Sigma_*V_*^{-1}$ and $U_* = \{u_{*ij}\}$ with

$$u_{*ij} = \sum_{s \neq t}^m \int_0^\infty \int_0^\infty \int_0^{y_2} \int_0^{y_1} B_{*,i}(x_1, y_1) w_*^{st} B_{*,j}(x_2, y_2) \rho_{st}(x_1, y_1, x_2, y_2) dx_1 dx_2 dy_1 dy_2 \\ + \sum_{s=1}^m \int_0^\infty \int_0^y B_{*,i}(x_1, y_1) w_*^{ss} B_{*,j}(x, y) \rho_s(x, y) dx dy.$$

P_* is the joint penalty matrix of P and the penalty for hazard estimation. The rest proof is similar to that for Theorem B2.

B.2. Coarsening approximation and model checking

The integration in equation (3.6) can be computationally intensive, thus, we use coarsening of survival time to approximate the integration of mean cost trajectories with respect to time to death. Suppose patient i is censored at time $T_i < \tau$, we assume the true time to death \tilde{T}_i falls in one of G_i prespecified time intervals denoted as L_{ig} ($g = 1, \dots, G_i$), such that $\cup_{g=1}^{G_i-1} L_{ig} = (T_i, \tau]$ and $L_{iG_i} = (\tau, \infty)$. Let l_{ig} be a representative point of corresponding interval; we specify l_{ig} as the middle point of L_{ig} , $g = 1, \dots, G_i-1$, and $l_{iG_i} = \tau^+$ throughout this paper. To balance between approximation accuracy and computational feasibility, we suggest the intervals of each censored subject to have equal length of at least one month, and a cap $G_i \leq 10$ to be the maximum number of intervals that a censored patient could have. We approximate the expectation of mean cost trajectories for censored subjects by a weighted average of possible mean cost trajectories,

$$E\{Y_i(t) | \tilde{T} > T_i\} \approx \sum_{g=1}^{G_i} E(Y(t) | \tilde{T} = l_{ig}) P_{ig}(\boldsymbol{\beta}),$$

where $P_{ig}(\boldsymbol{\beta}) = P(\tilde{T} \in L_{ig} | \tilde{T} > T_i)$ denotes the conditional probability weight for the actual time to death of subject i to fall in interval L_{ig} .

Figure B.1 shows a sensitivity analysis with estimated mean cost trajectories $\tilde{\mu}(t, s)$ at survival months 24, 48, 72, 96 and for LTS (120+). This figure can be viewed as complete case analysis when the data right-censored prior to maximum follow-up time τ is not utilized, while the $\hat{\mu}(t, s)$ method uses such data. The results and conclusions are similar to that of $\hat{\mu}(t, s)$. We conclude that our proposed method is robust against informative censoring.

B.3. Additional Simulation Studies

In response to this comment, we added simulation studies to evaluate the proposed method. The simulation design is shown as follows. We used our default simulation setting with gamma cost data, and sample size $n=500$. We alter

- (i) the censoring proportions to be 20% and 40%, and results are in Table B.1;
- (ii) the zero proportions to be 10% and 20%, and results are in Table B.2;
- (iii) the variance function to be $\text{var}(Y) = (1.5\mu)^2$ (moderate skewness) and $(2\mu)^2$ (high skewness), and results are in Table B.3.

Pointwise absolute error (AE), mean squared error (MSE), and coverage probability (CP) are based on aggregated results from 1000 Monte Carlo repetitions. GEE1-N is the estimator assuming constant variance ignoring subjects censored prior to τ ; GEE1-Y is the estimator assuming constant variance using all available data; and GEE2-Y is the proposed estimator assuming flexible variance using all available data.

All estimators show little bias, and coverage probabilities close to 95%. The efficiency gain from the use of censored data GEE1-Y compared to GEE1-N for the proposed estimator is clearly seen in all scenarios. We see increased MSE with increased censoring rate. This is expected since there is increasing loss of cost information. Larger MSEs for settings with larger zero proportions could be due to the increased variation. When the

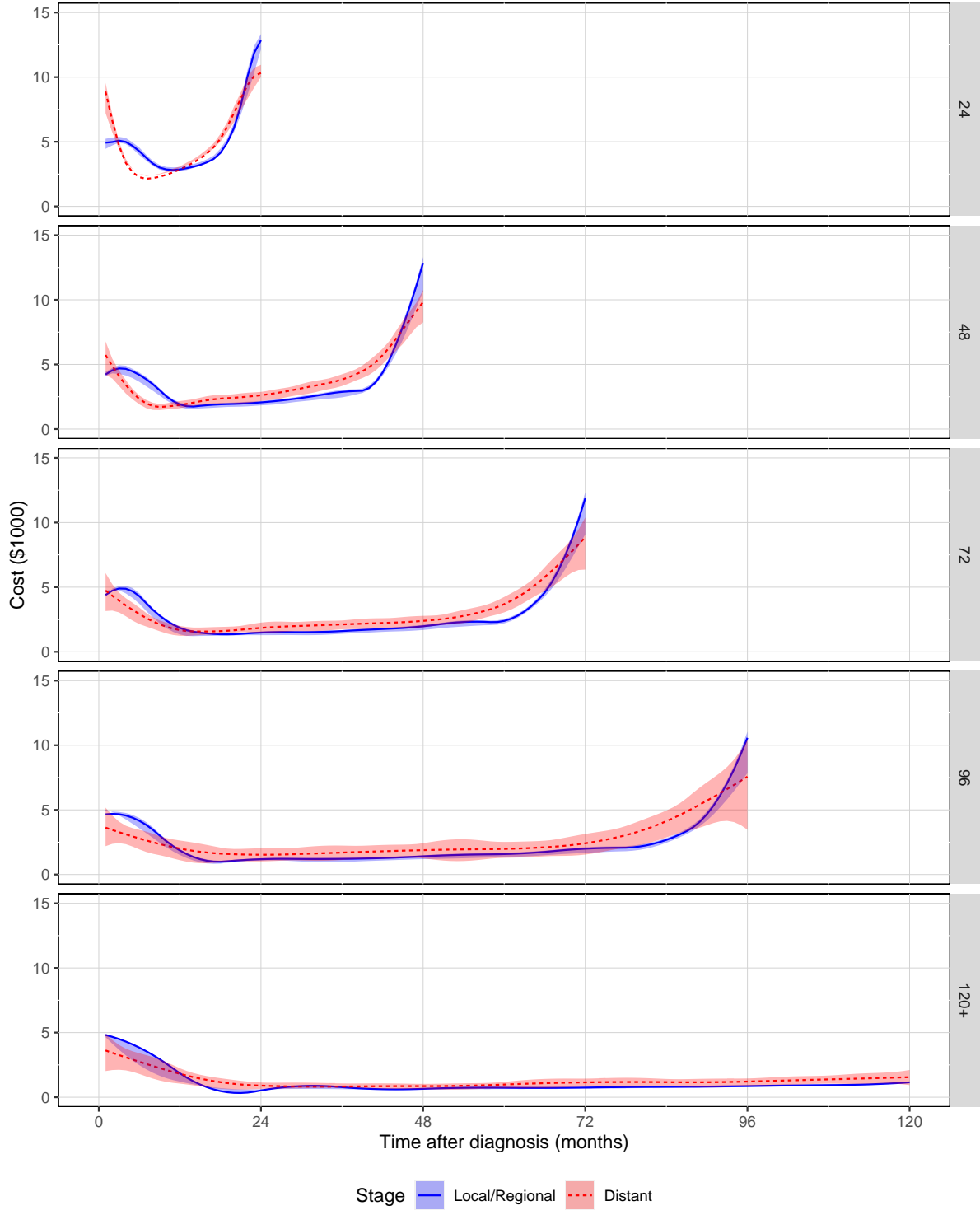


Figure B.1: SEER-Medicare data application results for estimated monthly cost trajectories $\tilde{\mu}(t, s)$ at survival months 24, 48, 72, 96 and for LTS (120+). Estimates at local/regional stage (blue) are plotted against estimates at distant stage. The shaded areas are 95% confidence intervals.

skewness of cost data increases, the MSE also slightly increases for both GEE1-Y and GEE2-Y. Consistent estimation of the cost trajectory is achieved without explicitly modeling the full distribution of the cost data. This makes the proposed method convenient to use.

The conclusion is that the proposed GEE2-Y method estimate the cost trajectory consistently. There is loss of efficiency with higher censoring rate, higher zero proportion or higher level of skewness.

Table B.1: Simulation results under different skewness for the estimation of cost trajectory $\mu(t, s)$ at selected t and s . In response to Reviewer 1 Comment 1 (i) different censoring proportions. Compared to the default setting, the censoring proportion is changed to 20% and 40%. The sample size $n = 500$. The cost data follow Gamma distributions. The sample size $n = 500$.

GEE1-N					GEE1-Y			GEE2-Y		
t	s	AE	MSE	CP	AE	MSE	CP	AE	MSE	CP
Censoring rate = 20%										
0.10	0.4	0.045	0.231	0.967	0.050	0.223	0.969	0.002	0.211	0.964
0.20	0.4	0.236	0.212	0.946	0.236	0.209	0.949	0.134	0.163	0.954
0.30	0.4	0.045	0.233	0.964	0.045	0.227	0.967	0.017	0.216	0.956
0.15	0.6	0.025	0.100	0.960	0.031	0.095	0.960	0.009	0.092	0.959
0.30	0.6	0.140	0.093	0.924	0.148	0.093	0.916	0.115	0.079	0.937
0.45	0.6	0.029	0.106	0.948	0.034	0.101	0.952	0.017	0.096	0.957
0.20	0.8	0.106	0.100	0.942	0.087	0.090	0.946	0.062	0.087	0.947
0.40	0.8	0.113	0.093	0.947	0.101	0.087	0.945	0.054	0.083	0.948
0.60	0.8	0.110	0.101	0.942	0.090	0.094	0.943	0.060	0.093	0.948
Censoring rate = 40%										
0.10	0.4	0.042	0.268	0.964	0.059	0.251	0.967	0.011	0.238	0.960
0.20	0.4	0.237	0.234	0.955	0.243	0.228	0.957	0.141	0.181	0.953
0.30	0.4	0.044	0.255	0.966	0.051	0.243	0.966	0.012	0.235	0.964
0.15	0.6	0.030	0.125	0.948	0.039	0.113	0.947	0.007	0.111	0.956
0.30	0.6	0.139	0.109	0.929	0.149	0.106	0.921	0.114	0.093	0.941
0.45	0.6	0.028	0.127	0.948	0.035	0.116	0.957	0.016	0.116	0.962
0.20	0.8	0.101	0.125	0.957	0.058	0.099	0.941	0.019	0.096	0.958
0.40	0.8	0.113	0.120	0.943	0.096	0.097	0.931	0.029	0.095	0.950
0.60	0.8	0.103	0.128	0.955	0.057	0.104	0.932	0.003	0.105	0.950

Table B.2: Simulation results under different zero proportions for the estimation of cost trajectory $\mu(t, s)$ at selected t and s . Compared to the default setting, the zero proportion is changed to 10% and 30%. The sample size $n = 500$.

GEE1-N					GEE1-Y			GEE2-Y		
t	s	AE	MSE	CP	AE	MSE	CP	AE	MSE	CP
Zero proportion = 10%										
0.10	0.4	0.053	0.356	0.957	0.089	0.316	0.971	0.056	0.307	0.959
0.20	0.4	0.236	0.314	0.945	0.262	0.296	0.955	0.163	0.246	0.958
0.30	0.4	0.030	0.351	0.964	0.054	0.317	0.972	0.002	0.310	0.958
0.15	0.6	0.028	0.209	0.947	0.053	0.167	0.942	0.016	0.171	0.955
0.30	0.6	0.148	0.173	0.934	0.168	0.155	0.888	0.128	0.141	0.933
0.45	0.6	0.032	0.204	0.951	0.055	0.177	0.937	0.041	0.191	0.950
0.20	0.8	0.075	0.264	0.960	0.034	0.162	0.525	0.085	0.184	0.866
0.40	0.8	0.096	0.239	0.965	0.054	0.154	0.555	0.047	0.182	0.891
0.60	0.8	0.082	0.277	0.952	0.055	0.188	0.492	0.138	0.235	0.858
Zero proportion = 30%										
0.10	0.4	0.062	0.453	0.964	0.093	0.403	0.977	0.067	0.391	0.965
0.20	0.4	0.230	0.384	0.952	0.258	0.358	0.965	0.167	0.311	0.961
0.30	0.4	0.044	0.433	0.958	0.066	0.390	0.968	0.011	0.381	0.961
0.15	0.6	0.023	0.254	0.955	0.051	0.206	0.945	0.024	0.207	0.950
0.30	0.6	0.160	0.208	0.933	0.182	0.184	0.906	0.147	0.168	0.929
0.45	0.6	0.030	0.254	0.942	0.053	0.217	0.946	0.048	0.223	0.951
0.20	0.8	0.079	0.316	0.957	0.023	0.188	0.602	0.071	0.205	0.886
0.40	0.8	0.108	0.281	0.971	0.062	0.174	0.616	0.033	0.195	0.918
0.60	0.8	0.087	0.322	0.949	0.048	0.222	0.561	0.125	0.259	0.894

Table B.3: Simulation results under different levels of skewness for the estimation of cost trajectory $\mu(t, s)$ at selected t and s . The variance function is changed to $\text{var}(Y) = (1.5\mu)^2$ and $(2\mu)^2$. The cost data follow Gamma distributions. The sample size $n = 500$.

GEE1-N					GEE1-Y			GEE2-Y		
t	s	AE	MSE	CP	AE	MSE	CP	AE	MSE	CP
$\text{var}(Y) = (1.5\mu)^2$										
0.10	0.4	0.048	0.460	0.959	0.084	0.404	0.972	0.036	0.389	0.963
0.20	0.4	0.234	0.383	0.951	0.264	0.355	0.965	0.156	0.301	0.960
0.30	0.4	0.037	0.450	0.959	0.062	0.404	0.968	0.007	0.393	0.959
0.15	0.6	0.029	0.275	0.943	0.055	0.218	0.934	0.016	0.220	0.952
0.30	0.6	0.156	0.220	0.928	0.173	0.193	0.890	0.134	0.175	0.928
0.45	0.6	0.040	0.268	0.952	0.062	0.226	0.935	0.046	0.240	0.944
0.20	0.8	0.069	0.347	0.955	0.039	0.211	0.548	0.109	0.239	0.891
0.40	0.8	0.097	0.317	0.964	0.052	0.197	0.558	0.062	0.231	0.908
0.60	0.8	0.067	0.346	0.948	0.069	0.230	0.504	0.166	0.285	0.886
$\text{var}(Y) = (2\mu)^2$										
0.10	0.4	0.052	0.592	0.956	0.088	0.516	0.974	0.033	0.494	0.964
0.20	0.4	0.239	0.479	0.960	0.272	0.438	0.971	0.158	0.376	0.963
0.30	0.4	0.038	0.578	0.963	0.063	0.514	0.971	0.014	0.497	0.957
0.15	0.6	0.031	0.352	0.942	0.057	0.275	0.942	0.022	0.277	0.955
0.30	0.6	0.161	0.276	0.926	0.177	0.237	0.907	0.139	0.216	0.935
0.45	0.6	0.044	0.341	0.952	0.065	0.284	0.950	0.052	0.296	0.944
0.20	0.8	0.062	0.434	0.952	0.044	0.264	0.583	0.122	0.294	0.907
0.40	0.8	0.096	0.398	0.964	0.050	0.242	0.584	0.070	0.278	0.921
0.60	0.8	0.062	0.432	0.949	0.072	0.283	0.547	0.177	0.340	0.899

APPENDIX C

APPENDIX FOR CHAPTER 3

This Appendix includes two sections. First, we present the theoretical properties of the proposed estimator and a sketch of the proof. Second, we present the model checking results from analysis of the prostate cancer cost data.

C.1. Asymptotics

we consider the consistency and asymptotic normality for fixed-knots penalized generalized estimating equations of the varying-coefficient single-index models with (possibly) right-censored covariates.

Let the estimating equations $\{\mathbf{U}_\theta(\boldsymbol{\theta})^T, \mathbf{U}_\gamma(\boldsymbol{\gamma})^T\}^T =: \mathbf{U}_\xi(\boldsymbol{\xi}) = \mathbf{0}$ depend on uncensored patients and LTS only, and data from subjects who were censored prior to τ is not used. Let the estimating equations $\{\mathbf{u}_\theta(\boldsymbol{\theta})^T, \mathbf{u}_\gamma(\boldsymbol{\gamma})^T\}^T =: \mathbf{u}_\xi(\boldsymbol{\xi}) = \mathbf{0}$ incorporate all available subjects given that the residual lifetime distribution for subjects who were censored prior to τ is known. Both joint estimating equations have the true parameter vector $\boldsymbol{\xi}^0 = (\boldsymbol{\theta}^{0T}, \boldsymbol{\gamma}^{0T})^T$, and $\boldsymbol{\xi}_{-1}^0 = (\boldsymbol{\theta}_{-1}^{0T}, \boldsymbol{\gamma}^{0T})^T$ excludes the first elements for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. In reality, the residual lifetime distribution is unknown, and needs to be estimated through a proper survival model. Thus, the corresponding joint estimating equations for the longitudinal and survival parameter vectors is $\{\mathbf{u}_\theta(\boldsymbol{\theta}; \boldsymbol{\beta})^T, \mathbf{u}_\gamma(\boldsymbol{\gamma}; \boldsymbol{\beta})^T, \mathbf{U}_\beta(\boldsymbol{\beta})^T\}^T =: \{\mathbf{u}_\xi(\boldsymbol{\xi}; \boldsymbol{\beta}), \mathbf{U}_\beta(\boldsymbol{\beta})^T\}^T = \mathbf{0}$, and the true parameter vector is $\boldsymbol{\phi}^0 = (\boldsymbol{\xi}^{0T}, \boldsymbol{\beta}^{0T})^T$. We express $\{\mathbf{u}_\theta(\boldsymbol{\theta})^T, \mathbf{u}_\gamma(\boldsymbol{\gamma})^T\}^T$ as $\mathbf{u}_\xi(\boldsymbol{\xi}; \boldsymbol{\beta})$ to emphasize its dependence on $\boldsymbol{\beta}$. Let $\boldsymbol{\phi}_{-1}^0 = (\boldsymbol{\xi}_{-1}^{0T}, \boldsymbol{\beta}^{0T})^T$. Denote $n_\Delta = \sum_{i=1}^n \{\delta_i + (1 - \delta_i)1(T_i = \tau)\}$.

C.1.1. Regularity Conditions

We adopt the following regularity conditions for the consistency and asymptotic normality of the proposed estimator (Yu and Ruppert (2002); Bai et al. (2009)).

C1. The number of observations each subject $\{n_i\}$ is bounded for some positive integers.

C2. The parameter space Ξ is compact, and the true parameter vector ξ^0 is an interior point of Ξ .

C3. The eigenvalues of the working covariance V and the true covariance Σ are bounded away from zero.

C4. $Pr(\tilde{T} < C) > 0$, such that $n_\Delta = O(n)$.

C5. $U_\xi(\xi) = \mathbf{0}$ has a unique root, and the first order derivative is bounded and positive.

C6. $\mathcal{U}_\xi(\xi) = \mathbf{0}$ has a unique root, and the first order derivative is bounded and positive.

Theorem C1. Under mild regularity conditions C1-C5,

1. If the smoothing parameter $\lambda_n = o(1)$, the model parameter $\tilde{\xi}$ estimated by solving equations $U_\xi(\xi) = \mathbf{0}$ converges to ξ^0 in probability.
2. If the smoothing parameter $\lambda_n = o(n_\Delta^{-1/2})$, the model parameter $\tilde{\xi}$ estimated by solving equations $U_\xi(\xi) = \mathbf{0}$ is asymptotically normal, that is

$$\sqrt{n} \left(\tilde{\xi} - \xi^0 \right) \rightarrow_d N(\mathbf{0}, \Sigma)$$

where $\Sigma = J_{\xi}(\xi^0) (\mathbf{H}_{0,\lambda}^T \mathbf{M}_{0,\lambda}^{-1} \mathbf{H}_{0,\lambda})^{-1} J_{\xi}^T(\xi^0)$, and $J_{\xi}(\xi^0) = \text{diag}\{J(\theta_1), J(\theta_2), \mathbf{I}_{\gamma}\}$.

3. For $k = 1, 2$ and $l = 1, \dots, p$,

$$\sqrt{n} \left(\tilde{\theta}_{kl} \tilde{\gamma}_k - \theta_{kl}^0 \gamma_k^0 \right) \rightarrow_d N \left(\mathbf{0}, \text{diag}\{\theta_{kl}^0 \gamma_k^0\}^T \mathbf{R}^T \text{diag}\{\xi^0\}^{-T} \Sigma \text{diag}\{\xi^0\}^{-1} \mathbf{R} \text{diag}\{\theta_{kl}^0 \gamma_k^0\} \right)$$

where \mathbf{R} is a index matrix corresponding to the parameters θ_{kl} and γ_k . The degrees of freedom df equals the number of parameters.

4. The Wald test statistic testing multiple parameters equal to zero simultaneously,

$$T_n = (\mathbf{C}\tilde{\xi})^T [\mathbf{C}\tilde{\Sigma}\mathbf{C}^T]^{-1} (\mathbf{C}\tilde{\xi}) \rightarrow \chi^2(df)$$

where \mathbf{C} is a index matrix corresponding to the parameters. The degrees of freedom df equals the number of parameters (= number of ones in \mathbf{C}).

Proof of Theorem C1:

Under independent censoring assumption, $E\{\mathbf{U}_{\xi 1}(\xi^0)\} = \mathbf{0}$ holds, so $n_{\Delta}^{-1} E\{\mathbf{H}_{n,\lambda}\}$ and $n_{\Delta}^{-1} E\{\mathbf{M}_{n,\lambda}\}$ converges uniformly in ξ_{-1} in a neighborhood of ξ_{-1}^0 as $n_{\Delta} \rightarrow \infty$, and the limits are $\mathbf{H}_{0,\lambda}$ and $\mathbf{M}_{0,\lambda}$, correspondingly. Since $\mathbf{U}_{\xi}(\xi) = \mathbf{0}$ is the score function for the penalized quasi-likelihood McCullagh and Nelder (1989) $\tilde{Q}(\mu, \mathbf{Y}) - \frac{1}{2} \xi^T \tilde{\Omega}_{\lambda} \xi$ where $\tilde{Q}(\mu, \mathbf{Y}) = \int_{\mathbf{Y}_i}^{\mu_i} (\mathbf{Y}_i - u)^T \mathbf{V}_i^{-1} du$, and $\tilde{\Omega}_{\lambda} = \text{diag}\{I_{\theta}, \Omega_{\lambda}(\gamma)\}$.

The proof for consistency and asymptotic normality of parameters in THEOREM (1 and 2) follows from the result that was obtained by penalized quadratic inference functions Bai et al. (2009) immediately. The asymptotic property does not depend on the choice of working covariance matrix, therefore, the asymptotic variance and MSE are minimized when the true covariance is used.

The proof for the extended asymptotic normality and Wald test statistic in THEOREM (3 and 4) is based on the asymptotic normality of $\tilde{\boldsymbol{\xi}}$. The asymptotic covariance of $\tilde{\theta}_{kl}\tilde{\boldsymbol{\gamma}}_k$ can be obtained via the multivariate delta method. It is useful for the inference on trajectory surfaces given baseline covariates. Finally,

Theorem C2. Under mild regularity conditions C1-C6,

1. If the smoothing parameter $\lambda = o(1)$, then the model parameter $\hat{\boldsymbol{\phi}}$ estimated by solving equations $\{\boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{\xi}; \boldsymbol{\beta})^T, \boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta})^T\}^T = \mathbf{0}$ exists, and converges to $\boldsymbol{\phi}^0$ in probability.
2. If the smoothing parameter $\lambda_n = o(n^{-1/2})$, then the model parameter $\hat{\boldsymbol{\phi}}$ estimated by solving equations $\{\boldsymbol{U}_{\boldsymbol{\xi}}(\boldsymbol{\xi}; \boldsymbol{\beta})^T, \boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta})^T\}^T = \mathbf{0}$ is asymptotically normal, that is

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^0) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = J(\boldsymbol{\phi}^0) (\boldsymbol{\mathcal{H}}_{0,\lambda_n}^T \boldsymbol{\mathcal{M}}_{0,\lambda}^{-1} \boldsymbol{\mathcal{H}}_{0,\lambda})^{-1} J^T(\boldsymbol{\phi}^0)$, and $J(\boldsymbol{\phi}^0) = \text{diag}\{J(\boldsymbol{\theta}_1), J(\boldsymbol{\theta}_2), \boldsymbol{I}_{\boldsymbol{\gamma}}, \boldsymbol{I}_{\boldsymbol{\beta}}\}$.

3. For $k = 1, 2$,

$$\sqrt{n} \left(\hat{\theta}_{kl} \hat{\boldsymbol{\gamma}}_k - \theta_{kl}^0 \boldsymbol{\gamma}_k^0 \right) \rightarrow_d N \left(\mathbf{0}, \text{diag}\{\theta_{kl}^0 \boldsymbol{\gamma}_k^0\}^T \boldsymbol{\mathcal{R}}^T \text{diag}\{\boldsymbol{\phi}^0\}^{-T} \boldsymbol{\Sigma} \text{diag}\{\boldsymbol{\phi}^0\}^{-1} \boldsymbol{\mathcal{R}} \text{diag}\{\theta_{kl}^0 \boldsymbol{\gamma}_k^0\} \right)$$

where $\boldsymbol{\mathcal{R}}$ is a index matrix corresponding to the parameters θ_{kl} and $\boldsymbol{\gamma}_k$. The degrees of freedom df equals the number of parameters.

4. The Wald test statistic testing multiple parameters equal to zero simultaneously,

$$T_n = (\boldsymbol{C}\hat{\boldsymbol{\phi}})^T [\boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^T]^{-1} (\boldsymbol{C}\hat{\boldsymbol{\phi}}) \rightarrow \chi^2(df)$$

where \mathbf{C} is a index matrix corresponding to the parameters. The degrees of freedom df equals the number of parameters.

Proof of Theorem C2: The derivation of the induced estimating equation leads to $E\{\mathbf{U}_{\xi_1}(\xi^0)\} = \mathbf{0}$, so $n^{-1}E\{\mathbf{H}_{n,\lambda}\}$ and $n^{-1}E\{\mathbf{M}_{n,\lambda}\}$ converges uniformly in ϕ_{-1} in a neighborhood of ϕ_{-1}^0 as $n \rightarrow \infty$, and the limits are $\mathbf{H}_{0,\lambda}$ and $\mathbf{M}_{0,\lambda}$, correspondingly. By the first-order Taylor expansion,

$$\hat{\xi} - \xi_0 = \tilde{\xi} - \xi_0 + \left\{ \frac{\partial \mathbf{U}_{\xi}(\xi; \hat{\beta})}{\partial \xi^T} \bigg|_{\xi=\tilde{\xi}} \right\}^{-1} \mathbf{U}_{\xi}(\tilde{\xi}; \hat{\beta}) + o_p(n^{-1/2}).$$

From the regularity conditions, we have the consistency and asymptotic normality of $\hat{\beta}$ Kauermann (2005) that leads to the unbiased estimator of conditional survival distribution $F_{\tilde{T}}(s|T_i, \mathbf{X}_i)$. Thus, $\mathbf{U}_{\xi}(\tilde{\xi}; \hat{\beta})$ converges to zero in probability. Then it follows the consistency of $\hat{\xi}$.

An application of the multivariate central limit theorem and delta method implies that $\hat{\phi}$ is asymptotically normal with mean zero and variance $J(\phi^0) (\mathbf{H}_{0,\lambda}^T \mathbf{M}_{0,\lambda}^{-1} \mathbf{H}_{0,\lambda})^{-1} J^T(\phi^0)$.

C.2. Model Checking

In this section, we check the model assumption of the proposed longitudinal varying coefficient single-index model.

The proposed model is meant to fit the best nonlinear curves that explains the data given a set of parameters. Therefore, the model relies on the assumption that the incident cost data follows nonlinear curves. In Figure C.1 (a), we verify that there are complicated dependence between nonlinear cost trajectory and survival. The histogram of quarterly costs visualized in Figure C.1 (b) shows that the data is skewed with 11.7% zero costs, and thus, our method may be appropriate without assuming a cost data distribution such

as normal. The estimated survival distribution visualized in Figure C.1 (c) illustrate the data features including heavy censoring, and the presence of a LTS group, which verifies that our two-part model may be particularly useful for local/regional cancer stage group. There is clear difference of the survival distribution in different cancer stages.

Our model also assumes that incident cost data follows nonlinear curves given a set of baseline covariates. Graphically we test in Figure C.2, plotting the estimated mean curves of quarterly medical cost by quadratic splines with five equally spaced knots, when the death time is 16 quarters after cancer diagnosis. First, the covariate effect is different for different cancer stages, suggesting two separate model is needed. Second, there is clearly nonlinear effect on cost trajectories given different baseline covariates, suggesting that using varying coefficient may be appropriate.

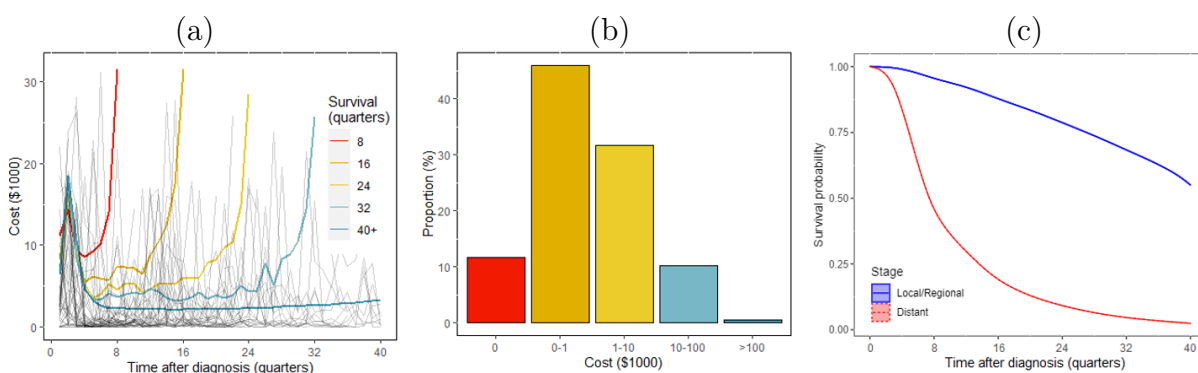


Figure C.1: Descriptive analysis of the quarterly medical costs for the SEER-Medicare prostate cancer cohort. (a) The trajectory of average quarterly costs among subjects who died at selected quarters. A random sample of 50 uncensored subjects was selected and their individual quarterly cost data are plotted in the background as gray lines. (b) The histogram of quarterly costs per 1,000 US dollars. (c) The estimated survival distributions by cancer stages. A substantial proportion of individuals with local regional stage are censored by the end of follow-up.

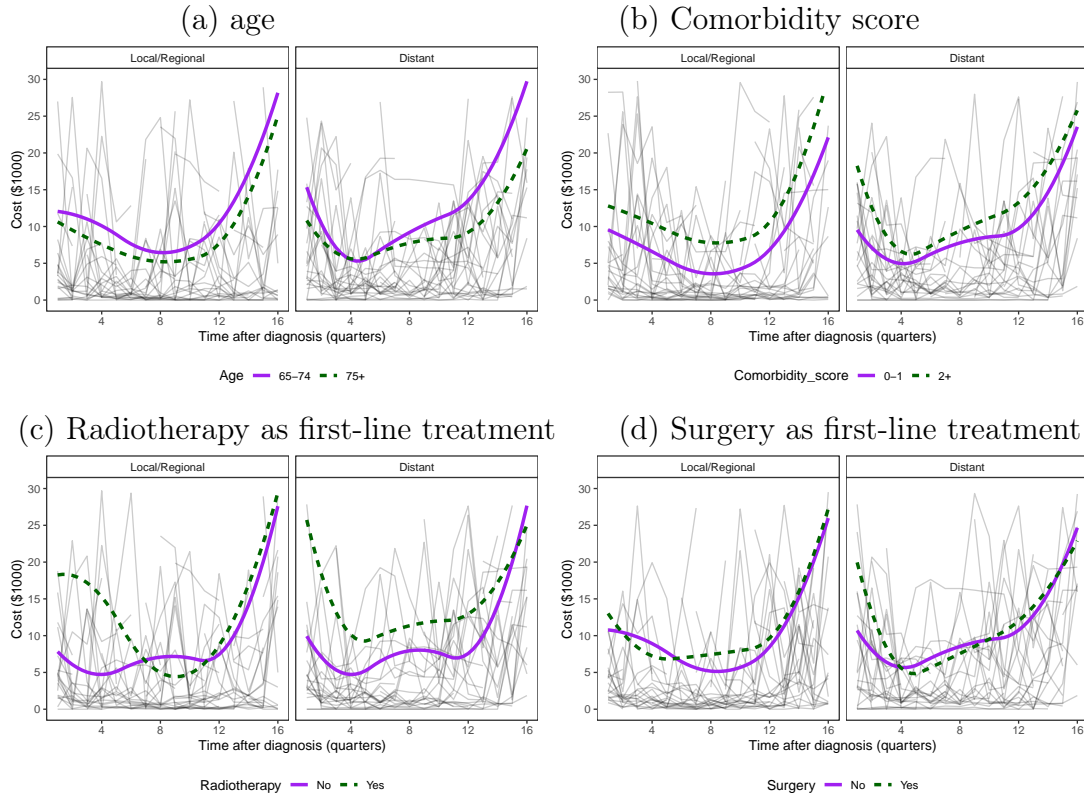


Figure C.2: Plots of the estimated mean curves of quarterly medical cost by quadratic splines with five equally spaced knots for the SEER-Medicare prostate cancer cohort by baseline covariates: (a) age, (b) comorbidity score, (c) radiotherapy as first-line treatment, and (d) surgery as first-line treatment. A random sample of 50 uncensored subjects was selected and their individual quarterly cost data are plotted in the background as gray lines. Death time is 16 quarters after cancer diagnosis.

C.3. Additional simulation results

We present results for simulation studies in addition to the two examples in the main text to investigate the effect of zero-inflation on either ideal setting (Normal outcome) or real setting (Gamma outcome). Without description, all settings are the same as the simulation settings in the main text. In Example 1*, we randomly set 10% cost data to zero in Example 1 (outcome is zero-inflated Normal instead of Normal). In Example 2*, we do not set cost data to zero in Example 2 (outcome is Gamma instead of zero-inflated Gamma). Tables C.1 show the pointwise absolute error (AE), MSE, and the coverage probability (CP) of a 95% confidence interval of linear coefficients on the estimated cost trajectory surface. The results are compared among the proposed “all data” estimator, and the reduced “partial data” estimator from Section 2. Both estimators show little bias, decreased MSE with increased sample size, and coverage probabilities close to 95%. The efficiency gain from the use of censored data in “all data” estimator compared to “partial data” estimator is clearly seen in all scenarios. Slightly larger MSEs of proposed estimation for zero-inflated settings may be due to the increased variation. Figure C.3 shows that consistent estimation of the baseline cost trajectory is achieved without explicitly modeling the full distribution of the cost data. This makes the proposed method convenient to use.

Table C.1: Bias, MSE and CP of linear coefficients in the simulation studies. (a) Data generated from zero-inflated Normal distribution, by randomly setting 10% zero costs in Example 1; (b) Data generated from Gamma distribution, similar to Example 2 without setting zero costs.

Example 1*: zero-inflated Normal data							
		All data			Partial data		
n		Bias	MSE ($\times 10^{-2}$)	CP	Bias	MSE ($\times 10^{-2}$)	CP
2000	θ_{11}	0.000	0.024	0.937	-0.001	0.026	0.938
	θ_{12}	-0.000	0.024	0.933	0.000	0.026	0.937
	θ_{21}	-0.000	0.011	0.945	-0.000	0.012	0.948
	θ_{22}	0.000	0.079	0.943	0.001	0.155	0.930
4000	θ_{11}	0.002	0.011	0.941	-0.000	0.012	0.948
	θ_{12}	-0.000	0.011	0.947	0.000	0.012	0.950
	θ_{21}	-0.001	0.042	0.939	-0.001	0.077	0.942
	θ_{22}	0.000	0.042	0.935	0.000	0.076	0.940
Example 2*: Gamma data							
		All data			Partial data		
n		Bias	MSE ($\times 10^{-2}$)	CP	Bias	MSE ($\times 10^{-2}$)	CP
2000	θ_{11}	-0.010	0.491	0.945	-0.015	0.777	0.907
	θ_{12}	0.004	0.439	0.944	0.004	0.666	0.897
	θ_{21}	-0.013	0.895	0.917	-0.028	1.997	0.896
	θ_{22}	0.001	0.793	0.913	0.004	1.483	0.891
4000	θ_{11}	-0.005	0.249	0.934	-0.007	0.363	0.922
	θ_{12}	0.001	0.234	0.937	0.002	0.325	0.928
	θ_{21}	-0.005	0.397	0.936	-0.011	0.851	0.914
	θ_{22}	-0.000	0.378	0.938	-0.000	0.769	0.917

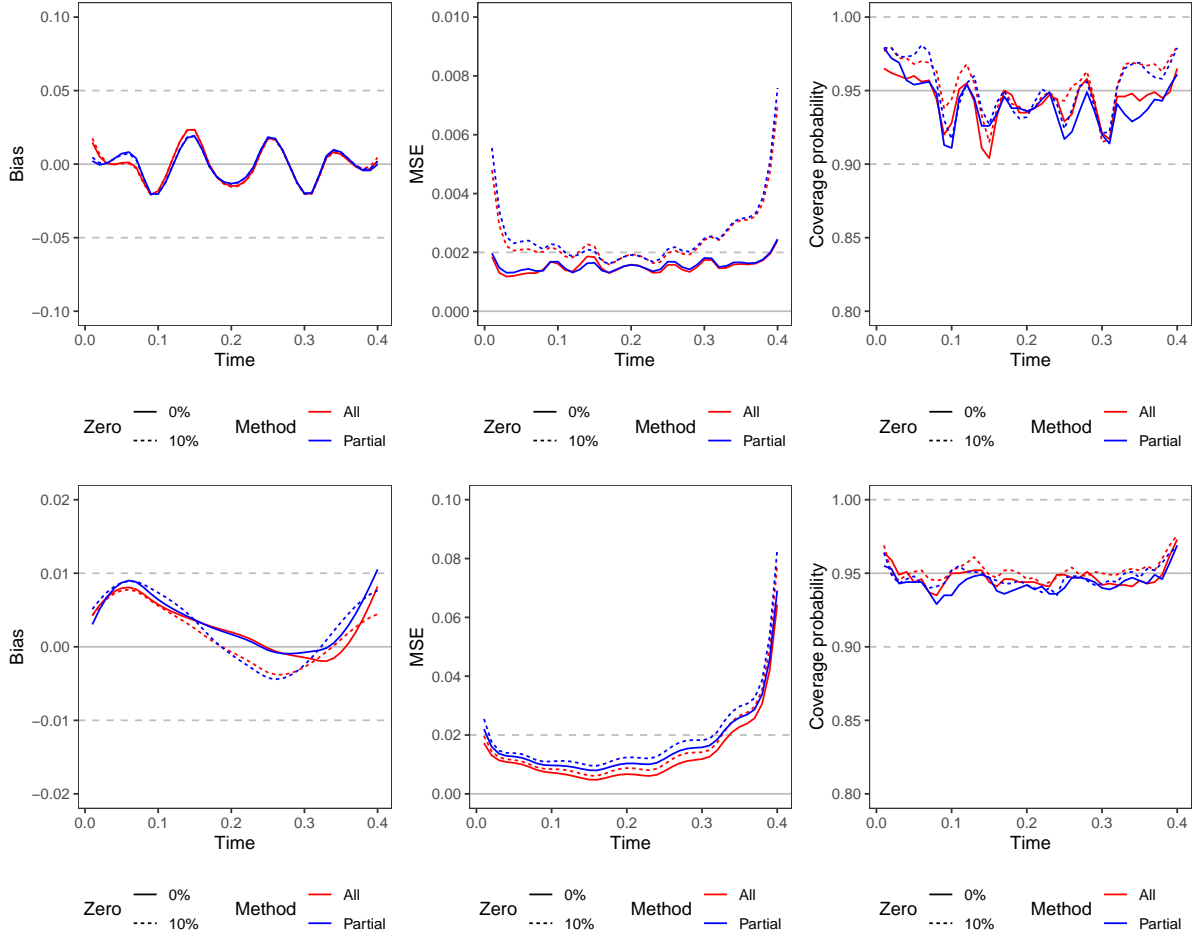


Figure C.3: Simulation results of estimated trajectories for Normal (top) or Gamma (bottom) data. $n=4000$, $x=(0,0)$, $s=0.4$. Left: plot of point-wise biases; middle: plot of point-wise mean squared errors; right: plot of pointwise empirical coverage probabilities. We compare lines for methods using all data (red) or using partial data (blue) having 0% (solid) or 10% (dashed) zero values.

BIBLIOGRAPHY

- Treating prostate cancer. *American Cancer Society*, Accessed Apr 13, 2022. URL <http://cancerstatisticscenter.cancer.org>.
- Y. Bai, W. K. Fung, and Z. Y. Zhu. Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis*, 100(1):152–161, 2009.
- D.L. Barrow and P.W. Smith. Asymptotic properties of best $l_2[0, 1]$ approximation by splines with variable knots. *Quarterly of applied mathematics*, 36(3):293–304, 1978.
- J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.
- N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.
- M. L. Brown, G. F. Riley, N. Schussler, and R. Etzioni. Estimating health care costs related to cancer treatment from seer-medicare data. *Medical Care*, 40:IV104–IV117, 2002.
- T. Cai, R. J. Hyndman, and M. P. Wand. Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, 11(4):784–798, 2002.
- K. C. G. Chan and M. C. Wang. Backward estimation of stochastic processes with failure events as time origins. *The Annals of Applied Statistics*, 4(3):1602–1620, 2010.
- H. Chen and Y. Wang. A penalized spline approach to functional mixed effects model analysis. *Biometrics*, 67(3):861–870, 2011.
- H. Chen, Y. Wang, M. C. Paik, and H. A. Choi. A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data. *Journal of the American Statistical Association*, 108(504):1216–1229, 2013.
- Kathleen A Cronin, Andrew J Lake, Susan Scott, Recinda L Sherman, Anne-Michelle Noone, Nadia Howlader, S Jane Henley, Robert N Anderson, Albert U Firth, Jiemin Ma, et al. Annual report to the nation on the status of cancer, part i: National cancer statistics. *Cancer*, 124(13):2785–2800, 2018.
- C. De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.
- Ian Duncan, Tamim Ahmed, Henry Dove, and Terri L Maxwell. Medicare cost at end of life. *American Journal of Hospice and Palliative Medicine*, 36(8):705–710, 2019.
- H. C. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.
- P. H. C. Eilers and B. D. Marx. Generalized linear additive smooth structures. *Journal of computational and graphical statistics*, 11(4):758–783, 2002.
- P. H. C. Eilers and B. D. Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653, 2010.
- G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal data analysis*. CRC press, 2008.
- T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 1991.
- W. J. Fu. Penalized estimating equations. *Biometrics*, 59(1):126–132, 2003.
- K. Goldfeld and J. Wujciak-Jens. simstudy: Illuminating research methods through data generation. *Journal of Open Source Software*, 5(54):2763, 2020.
- Michael W Kattan. Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer. *Current Opinion in Urology*, 13(2):111–116, 2003.
- F. Kauermann. Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49(1):169–186, 2005.
- S. Kong, B. Nan, J. D. Kalbfleisch, R. Saran, and R. Hirth. Conditional modeling of longitudinal data with terminal event. *Journal of the American Statistical Association*, 113(521):357–368, 2018.
- E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media, 2004.
- L. Li and T. Greene. Varying coefficients model with measurement error. *Biometrics*, 64

- (2):519–526, 2008.
- L. Li, C. H. Wu, J. Ning, X. Huang, Y. C. Shih, and Y. Shen. Semiparametric estimation of longitudinal medical cost trajectory. *Journal of the American Statistical Association*, 113(522):582–592, 2018.
- Y. Li and D. Ruppert. On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436, 2008.
- K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986a.
- K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986b.
- K. Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):3–24, 1992.
- Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- L. Liu. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*, 28(6):972–986, 2009.
- L. Liu, R. A. Wolfe, and J. D. Kalbfleisch. A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*, 26(1):139–155, 2007.
- L. Liu, X. Huang, and J. O’Quigley. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, 64(3):950–958, 2008.
- A. B. Mariotto, K. R. Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown. Projections of the cost of cancer care in the united states: 2010–2020. *Journal of the National Cancer Institute*, 103(2):117–128, 2011.
- A. B. Mariotto, L. Enewold, J. Zhao, C. A. Zeruto, and K. R. Yabroff. Medical care costs associated with cancer survivorship in the united states. *Cancer Epidemiology and Prevention Biomarkers*, 29(7):1304–1312, 2020.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models II*. Chapman and Hall, 1989.

- Alan W Partin, John Yoo, H Ballentine Carter, Jay D Pearson, Daniel W Chan, Jonathan I Epstein, and Patrick C Walsh. The use of prostate specific antigen, clinical stage and gleason score to predict pathological stage in men with localized prostate cancer. *The Journal of urology*, 150(1):110–114, 1993.
- J. Pinheiro and D. Bates. *Mixed-effects Models in S and S-PLUS*. Springer Science & Business Media, 2006.
- D. Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press, 2012.
- Oliver Karl Schilling. Cohort-and age-related decline in elder’s life satisfaction: is there really a paradox? *European Journal of Ageing*, 2(4):254–263, 2005.
- Marianne Schmid, Christian P Meyer, Gally Reznor, Toni K Choueiri, Julian Hanske, Jesse D Sammon, Firas Abdollah, Felix KH Chun, Adam S Kibel, Reginald D Tucker-Seeley, et al. Racial differences in the surgical care of medicare beneficiaries with localized prostate cancer. *JAMA oncology*, 2(1):85–93, 2016.
- Y. T. Shih, Y. Xu, C. R. Chien, B. Kim, Y. Shen, L. Li, and D. M. Geynisman. Rising economic burden of renal cell carcinoma among elderly patients in the usa: Part ii—an updated analysis of seer-medicare data. *PharmacoEconomics*, 37(12):1495—1507, 2019.
- Z. Shun and P. McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760, 1995.
- David A Siegel, Mary Elizabeth O’Neil, Thomas B Richards, Nicole F Dowling, and Hannah K Weir. Prostate cancer incidence and survival, by stage and race/ethnicity—united states, 2001–2017. *Morbidity and Mortality Weekly Report*, 69(41):1473, 2020.
- Justin G Trogon, Aaron D Falchook, Ramsankar Basak, William R Carpenter, and Ronald C Chen. Total medicare costs associated with diagnosis and treatment of prostate cancer in elderly men. *JAMA oncology*, 5(1):60–66, 2019.
- S. Wang, Y. Shen, Y. T. Shih, Y. Xu, and L. Li. Statistical modeling of longitudinal medical cost trajectory: renal cell cancer care cost analyses. *Biostatistics*, 2020.
- A. H. Welsh, X. Lin, and R. J. Carroll. Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*, 97(458):482–493, 2002.

- J. Wu, H. Peng, and W. Tu. Large-sample estimation and inference in multivariate single-index models. *Journal of multivariate analysis*, 171:382–396, 2019.
- L. Xiao. Asymptotics of bivariate penalised splines. *Journal of Nonparametric Statistics*, 31(2):289–314, 2019.
- M. Yamin. Counting the cost of covid-19. *International Journal of Information Technology*, 12(2):311–317, 2020.
- Y. Yu and D. Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002.
- L. P. Zhao and R. L. Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648, 1990.
- Zhiyuan Zheng, Ahmedin Jemal, Xuesong Han, Gery P Guy Jr, Chunyu Li, Amy J Davidoff, Matthew P Banegas, Donatus U Ekwueme, and K Robin Yabroff. Medical financial hardship among cancer survivors in the united states. *Cancer*, 125(10):1737–1747, 2019.
- Z. Zhu, W. K. Fung, and X. He. On the asymptotics of marginal regression splines with longitudinal data. *Biometrika*, 95(4):907–917, 2008.

VITA

Shikun Wang was born and grew up in China. After completing her high school education at Guangzhou No.6 Middle School, She went to Zhejiang University for her Bachelor's degree in Mathematics and Applied Mathematics from Sept, 2012 to July, 2016. She received the Master's degree in Applied Statistics at the University of Michigan, Ann Arbor in May, 2018. She joined the Ph.D. program in Biostatistics at the University of Texas MD Anderson Cancer UTHHealth Graduate School of Biomedical Sciences in 2018.