12-2021

Estimating treatment effect on medical cost and examining medical cost trajectory using splines and change point techniques.

Indranil Ghosh
University of Louisville

ESTIMATING TREATMENT EFFECT ON MEDICAL COST AND EXAMINING MEDICAL COST TRAJECTORY USING SPLINES AND CHANGE POINT TECHNIQUES

By

Indranil Ghosh
B.Sc., University of Calcutta, 2012
M.Sc., Banaras Hindu University, 2014

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

December 2021

ESTIMATING TREATMENT EFFECT ON MEDICAL COST AND EXAMINING MEDICAL COST TRAJECTORY USING SPLINES AND CHANGE POINT TECHNIQUES

By

Indranil Ghosh
B.Sc., University of Calcutta, 2012
M.Sc., Banaras Hindu University, 2014

A Dissertation Approved on

November 22, 2021

by the following Dissertation Committee:

Dr. Maiying Kong, Dissertation Director

Dr. Shesh Nath Rai

Dr. Michael Egger

Dr. Jeremy Gaskins

Dr. Qi Zheng

Dr. Dongfeng Wu

DEDICATION

This dissertation is dedicated to my parents
Mr Animesh Ghosh
and
Mrs Jayashree Ghosh
who have given me invaluable support throughout my life.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Maiying Kong, for her guidance and patience for my dissertation research. I would also like to thank Dr. Michael Egger, Dr. Dongfeng Wu, and the Department of Bioinformatics and Biostatistics at University of Louisville for their financial support during my study. My gratitude also goes to Dr. Qi Zheng and Dr. Michael Egger for their valuable inputs for my dissertation, and goes to my dissertation committee members Dr. Shesh Rai and Dr. Jeremy Gaskins for their time, and invaluable comments and support.

ABSTRACT

ESTIMATING TREATMENT EFFECT ON MEDICAL COST AND EXAMINING MEDICAL COST TRAJECTORY USING SPLINES AND CHANGE POINT TECHNIQUES

Indranil Ghosh

November 22, 2021

In the world of growing medical needs, other than the clinical outcomes, the cost of healthcare is one of the important aspects to evaluate. The cost of treatment could act as a decisive factor on which one to choose from two equally likely effective treatment options. In literature, the most used quantity for cost of a treatment is cumulative lifetime cost since the diagnosis of a disease. While it provides a bird's eye view of the treatment cost, it fails to capture the underlying pattern of the treatment cost trajectory. We developed a marginal structural functional model (MSFM) using an I-spline basis to examine the accumulative cost trajectory over time. Further, to obtain a valid average treatment effect (ATE) estimator, we used the inverse probability of treatment weighting (IPTW) to control the confounding between the cost and the treatment groups. Penalized spline regressions were used to estimate the cost trajectory and ATE. We carried out extensive simulation studies to examine the performance of the proposed method. We also applied our proposed method to gastric cancers patients based SEER-Medicare $2005 - 2014$ database, and illustrated the cost pattern over time under different treatments.

Another important aspect of healthcare cost is to identify the underlying pattern of the cost due to a disease. The estimation of healthcare cost and locating the change points across the cost trajectory is important to policy makers and clinicians, given the increasing costs of healthcare delivery, budgetary constraints, and the aging population. While in the literature the lifetime cost was often studied, the estimation of cost patterns and change points for cost patterns are important to policy makers and insurance companies. We develop a piece-wise linear mixed effect change point model as well as a I-spline based non-parametric model to estimate the cost trajectory over time and evaluate the change points for cost. We model the patient-level cost trajectory as well as population-level cost trajectory by using the patient-level regression parameters, which depend on patient-level characteristics and treatment choices. We applied our proposed methods on pancreatic cancer patients in SEER-Medicare $2005 - 2014$ database and concluded that both models capture the cost trajectory as well the change points.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Estimating Treatment Effect on Medical Cost Trajectory using Propensity Scores and I-Splines

Medical cost due to different treatments is one of the important aspects to evaluate in addition to the clinical outcomes. In the literature, the most commonly used quantity for cost is the average lifetime cost since the diagnosis of a disease, which usually fails to capture the changing pattern of the cost trajectory over time. The pattern of the cost trajectory over time could be due to treatment received and it could also be due to the patients' own characteristics. In Chapter 2, we develop a marginal structural functional model (MSFM) to examine the average cost trajectory over time. We apply the MSFM to evaluate the cost difference over time due to treatment based on observational data, such as SEER-Medicare data. It is well known that the cost trajectory is not only related to the treatment but also patients' characteristics. In addition, the treatment selection is also related to patients' own health conditions and preferences. Thus, the relationship between treatment received and the outcome variable is confounded. To obtain the cost difference due to treatment, confounding variables must be controlled. We apply generalized propensity score based approaches, in particular, the inverse probability of treatment weighting (IPTW) method to estimate the cost trajectory and compare the treatment effect. In the

MSFM, we propose using I-splines to model the cost trajectory. I-splines is quite flexible to capture different cost patterns meanwhile I-splines basis functions are monotonic and suitable to model the accumulative cost trajectories. We carried out extensive simulation studies to examine the performance of the proposed method. We applied the method to investigate the cost patterns and cost difference due to treatment for patients with stage 2 and 3 gastric cancers based on SEER-Medicare $2005-2014$ database.

1.2 Estimating Healthcare Cost using Parametric Change Point Models

Estimation of healthcare cost due to a certain disease over a period of time, say from diagnosis of the disease to the end of life, is very important to policy makers and clinicians, given the increasing costs of healthcare delivery, budgetary constraints, and the aging population. It will be informative to understand cost pattern and trajectory over the course of disease progression and the recovery from the disease. Based on the recent works, the lifetime cost of a patient can be divided into four phases. The first phase is the time period before diagnosis of the disease (i.e., staging phase), where the cost increases possibly due to frequent health care visits and laboratory tests for diagnoses. The second phase is post diagnosis phase, with a high cost after diagnosis due to different treatments but decreasing over time and stabilizing into a third phase (i.e., a stable phase). The final phase is the pre-death phase where the cost increases from the third phase (i.e., stable phase). In chapter 3, we propose a five phase model by adding a pre-disease phase prior to the staging phase, and we develop a piece-wise linear mixed effect change point model to capture cost trajectory and detect the time points at which changes of phases occur. We use three population-level parameters to model change points that capture the transition of different phases, and use patient-level characteristics such as patient demographics, comorbidity, and treatments to capture the cost amount and time elapsed in each phase. A grid-search approach is applied to estimate the change point parameters in the model by minimizing the residual sum of squares. We applied the proposed method to estimate the cost trajectory for pancreatic cancer patients in SEER-Medicare $2005 - 2014$ database and the detailed analysis and results are provided in Chapter 3.

1.3 Non-parametric Models for Estimating Medical Cost and Change Points

In Chapter 4, we use the concepts of cost phases and change points developed in Chapter 3 and develop a flexible non-parametric spline model to capture subject-level as well as population-level cost trajectory and detect the change points. In the non-parametric approach, we propose using I-splines to model the cost-trajectory, where the I-splines coefficients are modeled depending on patient characteristics and treatment received. Note that the first derivative of I-splines is M-splines, and it is straight forward to get the first derivative of the cost-trajectory using the relationship between I-splines and M-splines. Hence the change points are easily identified by the cost trajectory function and their derivatives. We applied the proposed method to estimate the cost trajectory for pancreatic cancer patients in SEER-Medicare $2005 - 2014$ database and provide a detailed analysis and result in Chapter 4.

CHAPTER 2

ESTIMATING TREATMENT EFFECT ON MEDICAL COST TRAJECTORY USING PROPENSITY SCORES AND I-SPLINES

2.1 Introduction

Estimation of medical cost is crucial in the field of health economics and policy-making (Lin et al. 1997). Medical expenses are any costs incurred in the prevention or treatment of injury or disease. Healthcare providers and policy makers are often interested in the average costs for certain treatment and procedure, and compare the costs for new treatment or device versus standard treatment (Bang and Tsiatis 2002; Basu et al. 2011). The average cost for a treatment is understood as the average cost if the entire underlying population received this treatment (Li et al. 2016). The average treatment effect (ATE) on cost is defined as the cost difference if the entire population had been treated with a treatment A versus treatment B, or the average cost difference if certain policies are implemented in the population. It is well known that medical costs for individual patients are often impacted by treatment choice and the patients' own characteristics and comorbidity (Austin 2011). The patients' characteristics and comorbidities also impact the treatment choices. Thus, the relationship between medical cost and treatment is susceptible to confounding. To evaluate the ATEs due to a treatment or a policy for a certain population, confounding must be appropriately controlled (Li et al. 2016).

In literature, the lifetime cost since certain event (e.g., cancer diagnosis) is often used as a metric to measure the cost. Many approaches have been developed to estimate the lifetime cost (Lin et al. 1997; Bang and Tsiatis 2002; Basu et al. 2011; Li et al. 2016). However, lifetime cost only provides the overall cost of the population and often fails to capture the underlying cost pattern. In this project, we target to examine the cost profile over time, and compare their difference under different treatments. We propose a marginal structural functional model (MSFM) to model the average cost profile under standard care (say, control group), and also model the average cost difference between treatment and control. The MSFM is a function of time and treatment, which captures the cost trajectory and patterns under different treatments. We used the I-spline basis functions to construct the MSFM, which is quite flexible to capture different patterns of costs as well as capture cost difference due to treatments. To control the confounding, we propose using the inverse probability of treatment weighting (IPTW) method to estimate the ATE. In the following Section 2.2, we present our proposed method in details. In Section 2.3, we carried out simulation studies to examine the performance of the proposed method. In section 2.4, we applied the proposed method to study the cost difference due to different treatments for patients with stage 2 and 3 gastric cancers based on SEER-Medicare $2005-2014$ database. Section 2.5 is devoted to conclusions and discussions.

2.2 Marginal structural functional model for cost trajectory

Let us denote $(X, A, \mathcal{T}, Y)$ as the random variable to be observed for each patient, where X denotes a vector of all p confounding variables, A denotes the treatment assignment with G levels, (say $A \in \{0, 1, \cdots, G-1\}$), $\mathcal{T}$ denotes a vector of times elapsed since a certain event (e.g., diagnosis of gastric cancer) on which cost accumulates, and Y denotes the accumulated healthcare cost over the time course $\mathcal{T}$. In particular, for a patient (say, i^{th} patient for $i = 1, 2, \cdots, N$), the costs accumulated at time $\mathcal{T}_i = (t_{i1}, t_{i2}, \cdots, t_{in_i})^\top$ are denoted as $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i})^\top$. We assume that the covariate X and treatment assignment A are time invariant variables, and the accumulated cost is monotonically increasing as time increases. Let us denote the observation for i^{th} patient as $(X_i, A_i, \mathcal{T}_i, Y_i)$ with $\mathcal{T}_i = (t_{i1}, t_{i2}, \cdots, t_{in_i})^\top$ and $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i})^\top$, where X_i are the observed covariates at baseline (say at diagnosis) and A_i is the treatment received after diagnosis for the i^{th} patient ($i = 1, 2, \cdots, N$). The accumulative cost for the i^{th} patient at time t_{ij} is Y_{ij}, where $t_{i1} < t_{i2} < \cdots < t_{in_i}$ and $Y_{i1} \leq Y_{i2} \leq \cdots \leq Y_{in_i}$. We intend to compare the cost difference due to different treatments. That is, we target to answer the research question on whether the cost differs if all patients received treatment option a_1 versus if all patients received treatment option a_0? Here we assume $a_0 \neq a_1$, $a_0, a_1 \in \{0, 1, \cdots, G-1\}$. Let us denote the potential cost trajectory under treatment a as a function of t, denoted by $Y^{(a)}(t)$. The potential cost trajectory for patient i under treatment a_1 at time $\mathcal{T}_i = (t_{i1}, t_{i2}, \cdots, t_{in_i})^\top$ would be $Y_i^{(a_1)} = (Y_{i1}^{(a_1)}, Y_{i2}^{(a_1)}, \cdots, Y_{in_i}^{(a_1)})^\top$, and the cost trajectory under treatment a_0 at time $\mathcal{T}_i$ would be $Y_i^{(a_0)} = (Y_{i1}^{(a_0)}, Y_{i2}^{(a_0)}, \cdots, Y_{in_i}^{(a_0)})^\top$. The observed outcome $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i}) = Y_i^{(a_0)}$ if $A_i = a_0$, and $Y_i = Y_i^{(a_1)}$ if $A_i = a_1$. Although there are G potential outcomes for i^{th} patient, say $Y_i^{(0)}, Y_i^{(1)}, \cdots, Y_i^{(G-1)}$, only one of them is observed which corresponds to treatment A_i that the i^{th} patient receives. That is, $Y_i = Y_i^{(A_i)}$, which is also referred as consistency assumption in the literature (Li et al. 2016).

We first develop a MSFM using flexible I-splines to model the average cost trajectory under different treatments:

$$\mu_a(t) = \mu_0(t) + \mathbb{1}_{\{A=a\}}\Delta_a(t). \tag{2.1}$$

Here $\mu_a(t) = E(Y^{(a)}(t))$, $\mathbb{1}_{\{A=a\}} = 1$ if $A = a$ and $\mathbb{1}_{\{A=a\}} = 0$ if $A \neq a$, and $\Delta_a(t)$ denotes the cost difference between treatment a and control:

$$\Delta_a(t) = E\Big(Y^{(a)}(t)\Big) - E\Big(Y^{(0)}(t)\Big). \tag{2.2}$$

It is worth to mention that the expected value of the potential outcome is taken over the entire population. Note that, for each patient, we can only observe the cost over time under the treatment assigned. In an observational study, treatment received is often impacted by patient health conditions, which also impact the outcome variable in terms of cost or survival. Rosenbaum and Rubin (1983) proposed using propensity score to balance confounding variables between treatment group and control group. Imbens (2000) extended it to generalized propensity scores for multiple treatment groups. The generalized propensity score is defined as $Pr(A = a|X)$ for $a = 0, 1, \cdots, G-1$, which is often estimated by parametric method (e.g., multiple logistic regression) or non-parametric method (e.g., generalized boosting method). Once the generalized propensity score is estimated, the IPTW can be applied to the observed data. The weight for the i^{th} patient is obtained by $w_i = \sum\limits_{a=0}^{G-1} \frac{\mathbb{1}_{\{A_i=a\}}}{Pr(A_i=a|X_i)}$ for $i = 1, 2, \cdots, N$. Thus, we form a pseudo-sample where i^{th} subject in the original sample is considered as w_i subjects in the pseudo-sample. The sample size from each treatment group in the pseudo-sample is approximately same as the original sample size N, and the distributions of each covariate across different treatment groups in the pseudo-sample are similar (Yan et al. 2021). Thus, there is no confounding between treatment assignment and outcome in the pseudo-sample. The inference for average cost difference due to certain treatment is based on the pseudo-sample.

Note that the propensity score based methods are valid for causal inference on average treatment effect under the assumptions of exchangeability and positivity. The exchangeability assumes that the potential outcome is independent of the treatment assignment, given the confounding variables, *i.e.*, $\left(Y^{(0)}(t), Y^{(1)}(t)\right) \perp A|X$. The positivity assumes that each patient has a chance to receive any one of the treatments considered, that is, $0 < Pr(A = a|X) < 1$ for $a = 0, 1, \cdots, G-1$. In this project, we make the same assumptions for the exchangeability and positivity. The consistency of the estimated causal parameters in the proposed models holds under these assumptions (Robins et al. 2000; Hernan and Robins 2019).

2.2.1 Model for accumulative cost and its estimate

Note that the accumulative cost over time is a monotonic function of t. We propose using I-spline basis function to capture the monotonic nature of the accumulative cost over time. Let us take K knots in the range of time points, say, $0 = \tau_1 < \tau_2 < \cdots < \tau_K = \max\limits_{\substack{i=1,\cdots,N \\ j=1,\cdots,n_i}} (t_{ij})$ with the interior knots based on the equally-spaced quantiles of t_{ij} $(i = 1, 2, \cdots, N;\ j = 1, 2, \cdots, n_i)$. Let us denote $I_\kappa^3(t)$ $(\kappa = 1, \cdots, K-2)$ as the cubic I-spline basis functions based on the K knots, where $I_\kappa^3(t)$ is a smooth monotonic function ranged between $[0, 1]$ with support interval on $[\tau_\kappa, \tau_{\kappa+3}]$ (Wan et al. 2017). The I-spline basis functions can be constructed using the *iSpline* function from the R-package *splines2*.

We use the linear combination of the I-spline basis functions to model the potential cost trajectory for control group.

$$\mu_0(t) = E\Big(Y^{(0)}(t)\Big) = \beta_0 + \sum_{\kappa=1}^{K-2} \beta_\kappa I_\kappa^3(t) = \mathbf{B}^\top(t)\beta, \tag{2.3}$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_{K-2})^\top$ with $\beta_\kappa \geq 0$ for $\kappa = 1, \cdots, K-2$ and $\mathbf{B}(t) = (1, I_1^3(t), \cdots, I_{K-2}^3(t))^\top$. The constraint on β_κ is a sufficient condition for $\mu_0(t)$ to be monotonic. The difference of potential cost trajectory between treatment a and control is modeled by $\Delta_a(t)$:

$$\Delta_a(t) = E\Big(Y^{(a)}(t)\Big) - E\Big(Y^{(0)}(t)\Big) = \gamma_0^{(a)} + \sum_{\kappa=1}^{K-2} \gamma_\kappa^{(a)} I_\kappa^3(t) = \mathbf{B}^\top(t)\gamma^{(a)} \tag{2.4}$$

with $\gamma^{(a)} = (\gamma_0^{(a)}, \gamma_1^{(a)}, \cdots, \gamma_{K-2}^{(a)})^\top$.

We propose the following MSFM to model the potential average cost trajectory if all patients had received treatment a $(a = 0, 1, \cdots, G-1)$:

$$\mu_a(t) = \mu_0(t) + \sum_{g=1}^{G-1} \mathbb{1}_{\{a=g\}} \Delta_g(t) = \mathbf{B}^\top(t)\beta + \sum_{g=1}^{G-1} \mathbb{1}_{\{a=g\}} \mathbf{B}^\top(t)\gamma^{(g)}. \tag{2.5}$$

Here $\mathbb{1}_{\{a=g\}}$ is an indicator variable, and $\mathbb{1}_{\{a=g\}} = 1$ if $a = g$, and 0 otherwise. The MSFM could be considered as an extension of the marginal structural model (MSM) developed by Hernan and Robins (Hernan and Robins 2019). It is clear that equation (2.5) reduces to the potential cost trajectory $\mu_0(t)$ when $a = 0$, and equation (2.5) becomes $\mu_a(t) = \mu_0(t) + \Delta_a(t)$ for treatment a. That is, $\mu_a(t)$ models the average potential cost trajectory over time for treatment group a for $a = 0, 1, \cdots G-1$. A sufficient condition for $\mu_a(t)$ to be monotonic is that β_κ and $\gamma_\kappa^{(a)}$ satisfy $\beta_\kappa + \gamma_\kappa^{(a)} \geq 0$ for $\kappa = 1, \cdots, K-2$ and $a = 1, 2, \cdots G-1$.

To estimate β and $\gamma^{(a)}$ in equation (2.5), the penalized splines are used. That is, β and $\gamma^{(a)}$ are estimated by minimizing the following penalized residual sum of squares (PRSS):

$$
\begin{aligned}
PRSS(\lambda) =& \sum_{i=1}^{N} w_i \Big[\Big(Y_i - \mu_0(\mathcal{T}_i) - \sum_{g=1}^{G-1} \mathbb{1}_{\{A_i=g\}} \Delta_g(\mathcal{T}_i)\Big)^{\top} \Big(Y_i - \mu_0(\mathcal{T}_i) - \sum_{g=1}^{G-1} \mathbb{1}_{\{A_i=g\}} \Delta_g(\mathcal{T}_i)\Big)\Big] \\
& + \lambda\Big(\|\beta\|^2 + \sum_{g=1}^{G-1} \|\gamma^{(g)}\|^2\Big) \\
=& \sum_{i=1}^{N} w_i \Big\|\Big(Y_i - \mathbf{B}^{\top}(\mathcal{T}_i)\beta - \sum_{g=1}^{G-1} \mathbb{1}_{\{A_i=g\}} \mathbf{B}^{\top}(\mathcal{T}_i)\gamma^{(g)}\Big\|^2 + \lambda\Big(\|\beta\|^2 + \sum_{g=1}^{G-1} \|\gamma^{(g)}\|^2\Big) \\
=& \Big(\mathbf{Y} - \mathbf{X}\theta\Big)^{\top} \mathbf{W} \Big(\mathbf{Y} - \mathbf{X}\theta\Big) + \lambda\|\theta\|^2. \qquad (2.6)
\end{aligned}
$$

Here

$$
\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} \in \mathbb{R}^{N_{tot}}, \quad \theta = \begin{pmatrix} \beta \\ \gamma^{(1)} \\ \vdots \\ \gamma^{(G-1)} \end{pmatrix} \in \mathbb{R}^{G(K-1)},
$$

$$
\mathbf{X} = \begin{bmatrix} \mathbf{B}^{\top}(\mathcal{T}_1) & \mathbb{1}_{\{A_1=1\}}\mathbf{B}^{\top}(\mathcal{T}_1) & \cdots & \mathbb{1}_{\{A_1=G-1\}}\mathbf{B}^{\top}(\mathcal{T}_1) \\ \mathbf{B}^{\top}(\mathcal{T}_2) & \mathbb{1}_{\{A_1=2\}}\mathbf{B}^{\top}(\mathcal{T}_2) & \cdots & \mathbb{1}_{\{A_2=G-1\}}\mathbf{B}^{\top}(\mathcal{T}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}^{\top}(\mathcal{T}_N) & \mathbb{1}_{\{A_N=1\}}\mathbf{B}^{\top}(\mathcal{T}_N) & \cdots & \mathbb{1}_{\{A_N=G-1\}}\mathbf{B}^{\top}(\mathcal{T}_N) \end{bmatrix},
$$

and $\mathbf{W} = \mathbf{BlockDiag}\Big(\mathbf{I}_{n_1\times n_1}\, w_1, \mathbf{I}_{n_2\times n_2}\, w_2, \cdots, \mathbf{I}_{n_N\times n_N} w_N\Big)$. Here $N_{tot} = \sum_{i=1}^{N} n_i$, $\beta = (\beta_0, \beta_1, \cdots, \beta_{K-2})^{\top}$, $\gamma^{(g)} = (\gamma_0^{(g)}, \gamma_1^{(g)}, \cdots, \gamma_{K-2}^{(g)})^{\top}$ for $g = 1, \cdots, G-1$. To make sure $\mu_a(t)$ $(a = 0, 1, \cdots, G-1)$ being monotonic, the minimization of the $PRSS(\lambda)$ is subject to $\beta_\kappa \geq 0$ and $\beta_\kappa + \gamma_\kappa^{(g)} \geq 0$ for $\kappa = 1, \cdots, K-2$ and $g = 1, 2, \cdots, G-1$. The constrained optimization can be implemented by the function *gam* in R-package *mgcv*.

The tuning parameter λ in equation (2.6) controls the trade-off between the goodness-of-fit in the first term and the smoothness in the second term. For a fixed λ, the fitted value $\hat{\mathbf{Y}} = S_\lambda \mathbf{Y}$, where

$$S_\lambda = \mathbf{X}\Big(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{I}\Big)^{-1}\mathbf{X}^\top \mathbf{W}.$$

To obtain S_λ, we solve the following equation for $\hat{\theta}$:

$$\begin{aligned} \frac{\partial PSS(\lambda)}{\partial \theta} &= -2\mathbf{X}^\top \mathbf{W}\Big(\mathbf{Y} - \mathbf{X}\theta\Big) + 2\lambda\theta = 0 \\ &\Longrightarrow \Big(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{I}\Big)\theta = \mathbf{X}^\top \mathbf{W}\mathbf{Y} \\ &\Longrightarrow \hat{\theta} = \Big(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{I}\Big)^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{Y}. \end{aligned}$$

Thus, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta} = \mathbf{X}\Big(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{I}\Big)^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{Y} \triangleq S_\lambda \mathbf{Y}$. λ is chosen as the one which minimizes the generalized cross validation (GCV) criteria (Golub et al. 1979):

$$GCV(\lambda) = \frac{N_{tot}||\mathbf{Y} - \hat{\mathbf{Y}}||^2}{(N_{tot} - tr(S_\lambda))^2}.$$

Here $tr(S_\lambda)$ is the trace of the matrix S_λ.

The treatment effect for treatment a versus control can be estimated as $\hat{\Delta}_a(t) = \hat{\gamma}_0^{(a)} + \sum_{k=1}^{K-2} \hat{\gamma}_k^{(a)} I_k^3(t)$, and the inference on $\Delta_a(t)$ can be made based on its estimate $\hat{\Delta}_a(t)$ and its bootstrap variance.

In case that there are two treatment groups, say treatment ($a = 1$) and control ($a = 0$), the proposed method can be easily simplified. The cost under control is equation (2.3) is $\mu_0(t) = \mathbf{B}^\top(t)\beta$, and the cost difference due to treatment is:

$$\Delta(t) = E\Big(Y^{(1)}(t)\Big) - E\Big(Y^{(0)}(t)\Big) = \gamma_0 + \sum_{\kappa=1}^{K-2}\gamma_\kappa I_\kappa^3(t) = \mathbf{B}^\top(t)\gamma. \qquad (2.7)$$

Thus, the MSFM in equation (2.5) is simplified as:

$$\mu_a(t) = \mu_0(t) + a\Delta(t) = \beta_0 + \sum_{\kappa=1}^{K-2} \beta_\kappa I_\kappa^3(t) + a\big(\gamma_0 + \sum_{\kappa=1}^{K-2} \gamma_\kappa I_\kappa^3(t)\big) = \mathbf{B}^\top(t)\beta + a\mathbf{B}^\top(t)\gamma. \tag{2.8}$$

Thus equation (2.8) reduces to the potential cost trajectory $\mu_0(t)$ when $a = 0$, and $\mu_a(t) = \mu_0(t) + \Delta(t)$ when $a = 1$. The estimation procedure is carried out similarly by minimizing the penalized residual sum of squares and λ is selected by minimizing the GCV criteria.

2.2.2 Model for monthly cost and its estimate

Note that the first derivative of I-spline is a M-spline, that is $I_\kappa^{3\prime}(t) = M_\kappa^3(t)$ (Ramsay 1988; Wan et al. 2017). Once we obtain the accumulative cost for $\mu_a(t) = \mu_0(t) + \Delta^{(a)}(t)$, we can take the first derivative for $\mu_a(t)$ to obtain the average monthly cost under each treatment condition:

$$\mu_a{}'(t) = \mu_0{}'(t) + \sum_{g=1}^{G-1} \mathbb{1}_{\{a=g\}} \Delta_g{}'(t) = \sum_{\kappa=1}^{K-1} \beta_\kappa M_\kappa^3(t) + \sum_{g=1}^{G-1} \mathbb{1}_{\{a=g\}} \sum_{\kappa=1}^{K-2} \gamma_\kappa^{(g)} M_\kappa^3(t). \tag{2.9}$$

We can also obtain the average monthly cost difference between treatment a and control: $\Delta_a'(t) = \sum_{\kappa=1}^{K-2} \gamma_\kappa^{(a)} M_\kappa^3(t)$. The monthly cost difference between treatment a and a' can be obtained as $\Delta_a'(t) - \Delta_{a'}'(t)$. The inference for $\Delta_a'(t)$ and $\Delta_a'(t) - \Delta_{a'}'(t)$ can be obtained using their estimates and their bootstrap variance estimates.

2.3 Simulation studies

To examine the performance of the proposed method, we carried out simulation studies under six different cost-profile generating models. For each cost-profile generating model, we generated 3 groups (say, one control group

($A = 0$) and two different treatment groups ($A = 1$ or 2)) based on multinomial distribution which was specified in step 2 under the simulation outlined later in this section. The cost generating model was taken as $Y(t) = b_0 + b_1 f(t) + I_{\{A=1\}}\Delta_1(t) + I_{\{A=2\}}\Delta_2(t) + \epsilon$, where $b_0 + b_1 f(t)$ mimics the cost profile for patient with covariate X at time t for control group, and $\Delta_a(t)$ captured the treatment effect due to treatment a $(a = 1, 2)$. We used two simulation settings for the cost-profile function in the control group:

F1. Linear increment rate per month: $f(t) = t$.

F2. Nonlinear increment rate per month: $f(t) = t + sin(t)$.

Under each increment setting, the cost differences due to treatment $\Delta_a(t)$ $(a = 1, 2)$ were simulated under each one of the following three settings:

S1. $\Delta_1(t) = 0$ and $\Delta_2(t) = 0$ for no treatment effect.

S2. $\Delta_1(t) = \gamma_0 + \gamma_1 t$ and $\Delta_2(t) = \gamma_0 + 10\gamma_1 t$ for cost difference being linear in time, where $\gamma_0 = 300$ and $\gamma_1 = -10$.

S3. $\Delta_1(t) = \gamma_0 + \gamma_1 sin(\frac{\pi}{4}t)$ and $\Delta_2(t) = \gamma_0 + 10\gamma_1 sin(\frac{\pi}{4}t)$ for cost difference being non-linear, where $\gamma_0 = 50$ and $\gamma_1 = -50$.

For each one of the six combinations, we also changed the magnitude of noise component $\epsilon \sim N(0, \sigma^2)$ by taking $\sigma = 10$ and $\sigma = 100$ respectively. Thus we had 12 simulation settings in total. We generated data under each simulation setting, we then used the proposed IPTW method to estimate the average treatment effect on the cost profile, and we also estimated the cost profile without IPTW. In the simulation study, we examined whether the proposed method could estimate the true cost profile appropriately and whether the estimates improved as the noise component was decreased. We followed the setting from Bang and Tsiatis (2002) and Li et al. (2016) with some modifications. The simulation studies were carried out by generating 1000 samples under each simulation setting, and each sample included 1000 patients ($N = 1000$). The simulation studies were carried out by the following steps:

Step 1. We generated a set of 4 covariates, $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ for i^{th} patient ($i = 1, 2, \cdots, N$), $X_{i1}, X_{i2} \sim Bernoulli(0.5) - 0.5$ and $X_{i3}, X_{i4} \sim Normal(0\,, 1)$. Here, we assume that $(X_{i1}, X_{i2}, X_{i3}, X_{i4})$ were time invariant covariates for the i^{th} patient.

Step 2. The treatment selection for the i^{th} patient (say A_i) was generated from the following multinomial distribution with parameter (p_{i0}, p_{i1}, p_{i2}), where $p_{ia} = Pr(A_i = a|X_i) = \frac{e^{X_i^\top \delta_a}}{1+e^{X_i^\top \delta_1}+e^{X_i^\top \delta_2}}$ for $a = (1, 2)$, $p_{i0} = 1 - p_{i1} - p_{i2}$, $\delta_1 = (1, -1, 1, -1)^\top$ and $\delta_2 = (-1, 1, -1, 1)^\top$.

Step 3. The survival time s_i (in months) for i^{th} patient was generated from an exponential distribution: $s_i \sim Exponential(20 + X_i^\top \delta_e)$, where $\delta_e = (0, 1, -1, 0)^\top$.

Step 4. We intended to examine the cost profile up to 24 months from diagnosis. We generated the accumulative cost response for i^{th} patient ($i = 1, 2, \cdots, N$) according to the cost-profile:

$$Y_{ij} = b_{i0} + b_{i1}\,f(t_{ij}) + I_{\{A_i=1\}}\Delta_1(t_{ij}) + I_{\{A_i=2\}}\Delta_2(t_{ij}) + \epsilon_{ij}.$$

Here the parameters $b_{i0} = 500 + X_i^\top \delta_{b_0}$ and $b_{i1} = 100 + X_i^\top \delta_{b_1}$ with $\delta_{b_0} = (100, -100, 10, -10)^\top$ and $\delta_{b_1} = (100, 10, 100, 10)^\top$, and the time sequence $t_{ij} = 0, 1, 2, \cdots, min(s_i, 24)$. For patients who died before 24 months from diagnosis (say, $s_i < 24$), the accumulative cost, Y_{ij}, would not change from the month of s_i to 24. That is, $Y_{ij} = Y_{is_i}$

for $s_i < t_{ij} \leq 24$.

Step 5. Using the data (X_i, A_i) generated from Steps 1 and 2, we estimated the generalized propensity score $Pr(A = A_i|X_i)$ using the multinomial regression model. We then calculated the weight for the i^{th} subject as the inverse probability of treatment received, that is, $w_i = \frac{1}{Pr(A=A_i|X_i)}$.

Step 6. We then applied the IPTW method to estimate cost profile (say, $\hat{\mu}_0(t)$) and treatment effects on cost (say, $\hat{\Delta}_a(t)$ for $a = 1, 2$). We also estimated the cost profile and treatment effect without using IPTW, which were used to examine the performance of the proposed method using IPTW versus without using IPTW.

Step 7. To examine the performance of each method we calculated the potential true cost for i^{th} patient under each treatment a ($a = 0, 1, 2$) as

$$Y^{(a)}(t_{ij}) = b_{i0} + b_{i1}\, f(t_{ij}) + I_{\{a=1\}}\Delta_1(t_{ij}) + I_{\{a=2\}}\Delta_2(t_{ij}) + \epsilon_{ij},$$

for $i = 1, \cdots, N$ and $j = 1, \cdots, min(s_i, 24)$. If the survival time for i^{th} patient is less than 24, then $Y_{ij}^{(a)} = Y_{in_i}^{(a)}$ for $s_i < t_{ij} \leq 24$. The true treatment effects for cost were calculated as:

$$\Delta_1^{true}(t) = \sum_{i=1}^{N} Y_{it}^{(1)} - \sum_{i=1}^{N} Y_{it}^{(0)} \text{ and } \Delta_2^{true}(t) = \sum_{i=1}^{N} Y_{it}^{(2)} - \sum_{i=1}^{N} Y_{it}^{(0)}$$

for $t = 1, 2, \cdots, 24$.

Step 8. We repeated the simulation $M = 1000$ times under each simulation setting. The resulting estimated cost profile for cohort group in m^{th} simulation was denoted as $\hat{\mu}_0^{(m)}(t)$ and the resulting estimated treatment effect in m^{th} simulation was denoted as $\hat{\Delta}_a^{(m)}(t)$ $(a = 1, 2; m = 1, \cdots, M)$.

Step 9. We calculated the true average treatment effect as the mean of $\Delta_1^{true}(t)$ and $\Delta_2^{true}(t)$ over the 1000 simulated data, denoted as $\bar{\Delta}_1^{true}(t)$ and $\bar{\Delta}_2^{true}(t)$. The performance of proposed methods with/without using IPTW were indicated by how close the ATE estimates to the true treatment effects $\bar{\Delta}_1^{true}$ and $\bar{\Delta}_2^{true}$ were. The following performance metrics were used for treatment a versus control $(a = 1, 2)$:

Mean absolute error (MAE),

$$MAE_a = \frac{1}{M} \sum_{m=1}^{M} \left(\sup_{1 \le t \le T} \left(|\hat{\Delta}_a^{(m)}(t) - \bar{\Delta}_a^{true}(t)| \right) \right);$$

Mean bias error (MBE),

$$MBE_a = \left| \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{T} \sum_{t=1}^{T} \left(\hat{\Delta}_a^{(m)}(t) - \bar{\Delta}_a^{true}(t) \right) \right) \right|;$$

and Root mean square error (RMSE)

$$RMSE_a = \frac{1}{M} \sum_{m=1}^{M} \left(\sqrt{ \frac{1}{T} \sum_{t=1}^{T} \left(\hat{\Delta}_a^{(m)}(t) - \bar{\Delta}_a^{T}(t) \right)^2 } \right)$$

where $T = 24$ and $M = 1000$.

Simulation results: The estimated cost trajectory for control group and the average treatment effect on cost between treatment a and cohort over time were presented in Figure 2.1 for linear increment rate function and in Figure 2.2 for non-linear increment rate function. The performance metrics under 12 different simulation settings, which included two types of increment rate for control group (F1 and F2), three types of cost difference functions (S1, S2, and S3), and two noise settings ($\sigma = 10$ and 100), were summarized in Table 2.1.

Table 2.1: Summarized performance metrics for estimating the average treatment effect on cost with IPTW (checked) and without IPTW under 12 different settings.

$f(t_{ij})$	$\Delta(t)$	σ	IPTW	$\Delta_1(t)$			$\Delta_2(t)$		
				MAE	MBE	RMSE	MAE	MBE	RMSE
F1	S1	10	✓	336.76	55.46	202.94	405.40	46.90	239.74
			×	1213.08	646.89	738.23	1218.85	646.66	739.75
		100	✓	437.83	128.12	274.85	421.34	124.77	244.97
			×	1275.49	713.23	796.05	1282.23	713.25	797.74
	S2	10	✓	368.27	72.33	220.88	363.24	49.06	217.11
			×	1221.67	651.57	743.48	1216.11	644.80	737.76
		100	✓	412.16	120.43	257.41	387.74	117.51	240.46
			×	1288.21	719.26	803.15	1277.49	710.25	794.42
	S3	10	✓	398.50	49.52	239.12	387.54	50.71	225.91
			×	1215.12	649.03	740.45	1205.89	639.21	731.22
		100	✓	423.50	103.34	263.99	411.15	110.53	247.74
			×	1281.08	716.61	799.87	1270.19	705.88	789.60
F2	S1	10	✓	346.78	66.68	216.50	384.42	27.10	234.87
			×	1179.28	650.98	740.57	1157.44	635.34	724.42
		100	✓	381.24	139.60	244.82	408.56	109.38	257.11
			×	1245.29	718.67	800.18	1230.28	707.25	788.66
	S2	10	✓	341.79	48.83	212.06	379.41	48.81	234.36
			×	1177.11	650.36	739.51	1153.24	632.80	721.71
		100	✓	391.15	108.76	249.74	381.75	139.65	245.26
			×	1247.10	719.20	800.82	1242.32	715.60	797.26
	S3	10	✓	357.94	48.02	221.42	383.13	47.19	232.75
			×	1165.86	643.91	732.40	1176.01	646.14	736.56
		100	✓	397.63	110.25	254.64	428.49	142.62	269.90
			×	1258.02	725.28	807.92	1238.21	714.98	796.04

Figure 2.1 showed the simulation results for the estimated cost trajectory for control group under the linear increment rate function F1: $f(t_{ij}) = t_{ij}$ (Panel A1) and the cost difference over time under three different settings for $\Delta_a(t)$: no treatment effect (Panel A2), linear treatment effect (Panel A3),

Figure 2.1: Simulation results for the estimated cost trajectory for control group under the linear increment rate function F1: $f(t_{ij}) = t_{ij}$ (Panel A1) and the cost difference over time under three different settings for $\Delta_a(t)$: no treatment effect (Panel A2), linear treatment effect (Panel A3), and non-linear treatment effect (Panel A4).

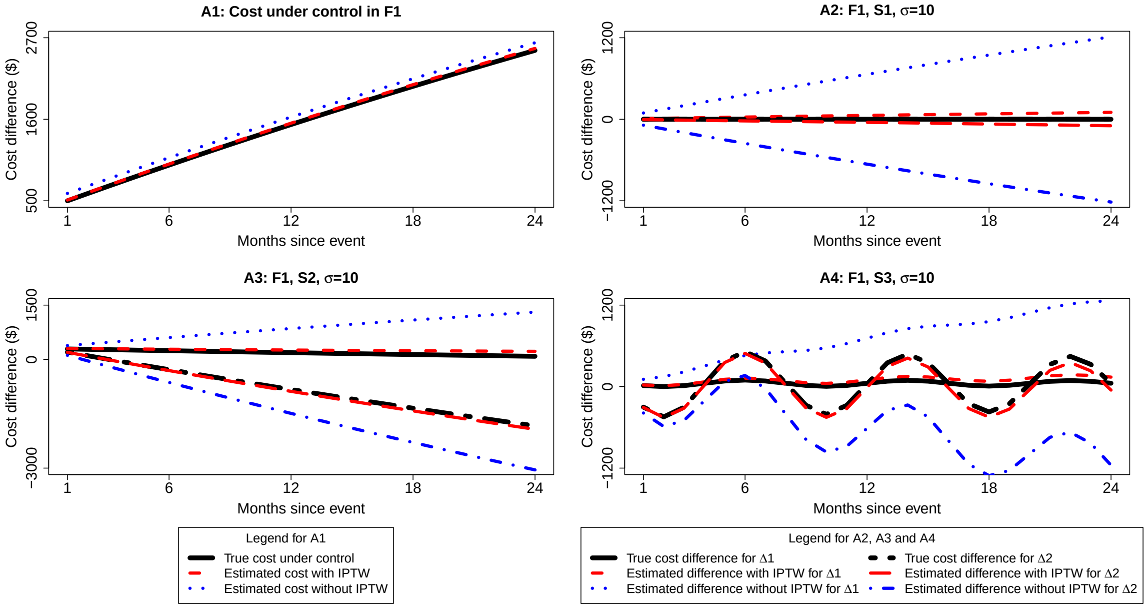
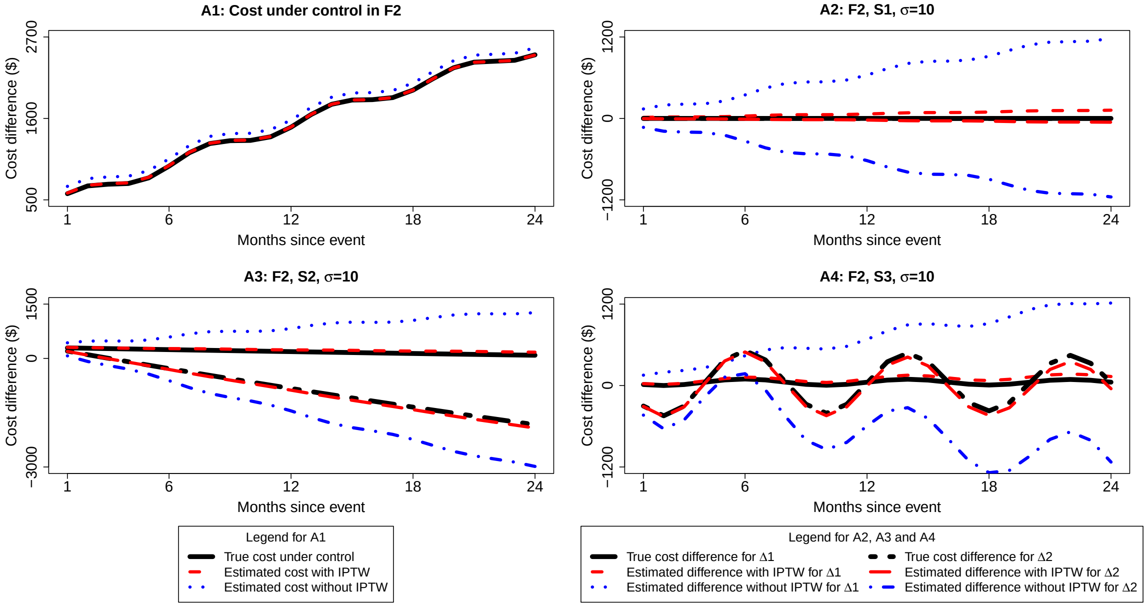


Figure 2.2: Simulation results for the estimated cost trajectory for control group under the non-linear increment rate function F2: $f(t_{ij}) = t_{ij} + sin(t_{ij})$ (Panel A1) and the cost difference over time under three different settings for $\Delta_a(t)$: no treatment effect (Panel A2), linear treatment effect (Panel A3), and non-linear treatment effect (Panel A4).



and non-linear treatment effect (Panel A4), where σ was set as 10. In Figure 2.1 Panel A1, the true cost trajectory under control was shown as solid line, the estimated cost trajectory using IPTW for control group was shown as a dashed line which overlaid with the true cost trajectory (solid line), indicating an unbiased estimation to the true cost trajectory. The estimated cost trajectory without using IPTW for control was shown as a dotted line in panel A1, which was far from the solid line, indicating a biased estimate for the true cost trajectory. Panel A2-A4 showed the true treatment effect on cost between treatment 1 and control (i.e. $\Delta_1(t)$) as a solid line, between treatment 2 and control (i.e. $\Delta_2(t)$) as a dashed line. The estimated cost difference using IPTW for $\hat{\Delta}_1(t)$ was shown as a short dashed line, and for $\hat{\Delta}_2(t)$ as a long dashed line. The estimated treatment effects without IPTW were shown as a dotted line for $\hat{\Delta}_1(t)$ and a dash-dotted line for $\hat{\Delta}_2(t)$. Panel A2 showed the results for no treatment effect, Panel A3 showed for linear treatment effect, and Panel A4 showed for non-linear treatment effect. In Panel A2, $\bar{\Delta}_1(t) = \bar{\Delta}_2(t) = 0$, the true treatment cost for treatment 1 and treatment 2 overlaid as the solid line. The estimated average treatment effect using IPTW $\hat{\Delta}_1(t)$ (dashed line) and $\hat{\Delta}_2(t)$ (long dashed line) Where close to the true $\bar{\Delta}_1(t)$ and $\bar{\Delta}_2(t)$, while the estimated ATE without using IPTW (dotted line for $\hat{\Delta}_1(t)$ and dash-dotted line for $\hat{\Delta}_2(t)$) were far away from the truth $\bar{\Delta}_1(t)$ and $\bar{\Delta}_1(t)$. Similar results were shown for panel A3 and A4. Figure 2.2 showed the simulation results under non-linear increment rate F2: $f(t_{ij}) = t_{ij} + sin(t_{ij})$ under different treatment effects. From Figure 2.1 and 2.2, it is clear that our estimated cost profile for control group and treatment effect with IPTW were close to the true values while the estimate without using IPTW were far away from the true values indicating that the proposed IPTW provided unbiased estimates, but not the method without

IPTW.

Table 2.1 summarizes the performance measures for each of the 12 simulation scenarios. From Table 2.1, it is clear that the performance metrics under each setting with the proposed IPTW method were better to estimate the average treatment effect than those without using IPTW. It was also clear that the estimation improved as the noise, σ, decreased from 100 to 10. Hence, the inference on ATE should be made based on the proposed method with IPTW.

2.4 Case study

Gastric cancer is a major health burden worldwide, and it is the second cause of cancer deaths after lung cancer (Correa 2013). The treatment options for the patients with stage II or III gastric cancer include chemotherapy and surgical procedures. It is interesting to examine whether the addition of chemotherapy to surgical procedure would benefit the patients' outcomes in terms of survival and overall medical costs. We consider three treatment groups: chemotherapy was given before surgery but not after (say, A1: Chemo $<$ Surgery); chemotherapy was given both before and after surgery (say, A2: Chemo $<$ Surgery $<$ Chemo); and chemotherapy was given after surgery but not before (say, A3: Surgery $<$ Chemo). Thus, we restricted our study cohort to patients with stage II or III gastric cancer who had surgical procedures and chemotherapy for gastric cancer using SEER Medicare $2005-2014$ gastric cancer data. It is well known that the cost is not only determined by the treatment received but also by patients' own comorbid conditions. We use the national cancer institute (NCI) comorbidity index to measure the patient's comorbidity (Klabunde et al. 2000), which were obtained from Medicare claims data one year prior diagnosis. We also examined the cost within two years after diagnosis. Thus, the study cohort was formed using the SEER-Medicare enrollment file (Pedsf) with the following inclusion criteria: (1) patients with gastric cancer specific primary site, histology, and behavioural code in the SEER database within year $2006 - 2012$; (2) patients in stage II and III, since patients in stage II and III are more likely to go through both chemotherapy and surgical procedure.

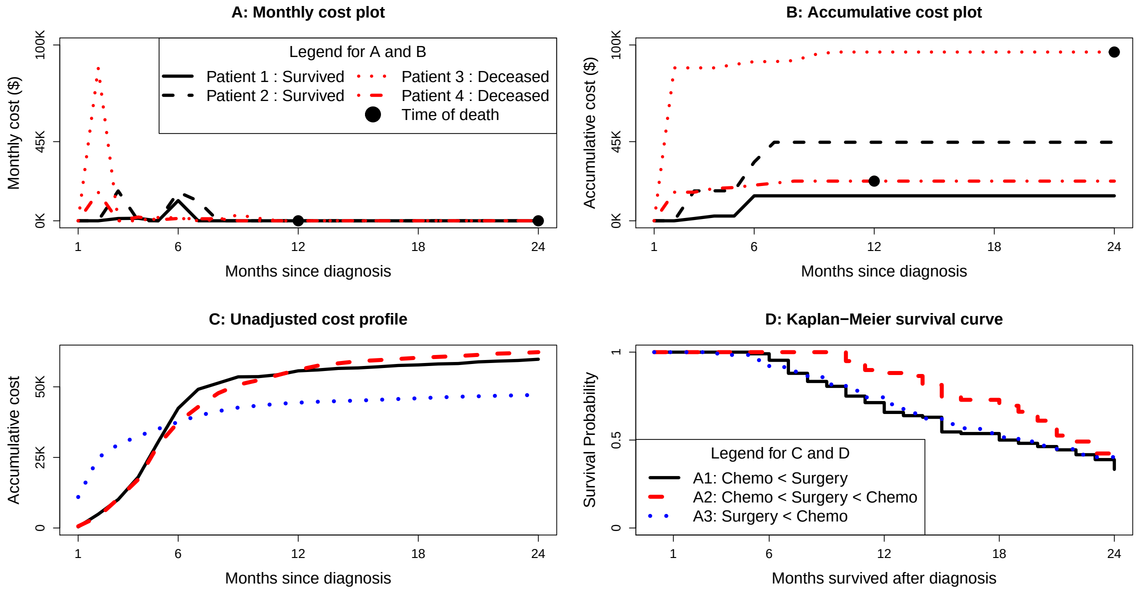
The NCI comorbidity index were calculated using the 2014 NCI SAS Macro from the NCI website (NCI 2014) using the SEER-Medicare enrollment file (Pedsf) and the diagnostic codes in the inpatient file (Medpar), the outpatient file (Outpat) and the carrier claims file (NCH). The other covariates obtained from Pedsf included demographic variables (i.e. race, age, sex), geographical variable (such as urban/rural and state) and cancer variables (such as specific primary site, histology, behavioural code and indicator for first diagnosis).

The three different treatment groups (say A1, A2, and A3) were formed by following algorithms from literature (Yeh et al. 2017; Wu et al. 2019). Group A1 for "Chemo < Surgery" included patients who had chemotherapy between diagnosis and surgery but not after surgery. More specifically, a patient was counted in group $A1$ if all of the following inclusion/exclusion criteria were met: (i) first chemotherapy was after diagnosis but within six months of diagnosis, (ii) surgery was after the first chemotherapy but within one year of the first chemotherapy, and (iii) No chemotherapy seen between surgery and the following six months. Group $A2$ for "Chemo < Surgery < Chemo" included patients who had the same criteria as (i) and (ii) from group A1, however the patient had chemotherapy after surgery within six months period. Group $A3$ for "Surgery < Chemo" included patients who had (i) surgery after diagnosis within one year, (ii) the first chemo therapy was after surgery within six months period. We had 6447 Patients with stage II and III Gastric cancer. By applying the inclusion/exclusion criteria (flow chart in Appendix, Figure A0.1), we had $N = 959$ patients for this study. Among them, 156 patients were under treatment A1, 92 patients were under treatment A2, and 711 patients were under treatment A3.

All medical cost from Medpar, Outpat and NCH files from the day of diagnosis up to two years of follow-up were obtained as the primary outcome in the analysis. The costs were standardized to 2014 cost rates using the Consumer Price Index (CPI-U) data (U.S. Department of Labor Bureau of Labor Statistic 2021). For each patient, monthly costs were calculated as the sum of all costs occurring during that month. Once we have the monthly costs, we calculated the accumulative monthly costs for two years since diagnosis. Figure 2.3 showed the monthly cost (Panel A) and cumulative cost (Panel B) for four randomly selected patients. Figure 2.3 Panel C portrayed the observed average accumulative cost trajectory under the three treatment groups (solid line for A1, dashed line A2, and dotted line A3), and Panel D showed the Kaplan–Meier survival curves for the three treatment groups.

Descriptive statistics and significance tests for the covariates associated with treatment and cost were provided in Table 2.2 under the column “Original sample”. It was clear that age, NCI index, race, and geographic variables (say, different states) were significantly associated with treatment groups. The survival time were not significantly associated with treatment groups, and the accumulative two-year costs were significantly associated with treatment groups. However, there were possible confounding between treatment and outcome. We applied the proposed IPTW method to estimate the cost trajectory over time. To do that, we first estimated the generalized

Figure 2.3: Illustration of monthly cost (Panel A) and accumulative cost (Panel B) for four randomly selected patients, and the average cost trajectory (Panel C) and the Kaplan–Meier survival curves (Panel D) for the three different treatment groups.



propensity score using the multinomial regression model and obtained the inverse probability of treatment received as a weight for each subject to form a weighted sample. The summarized statistics in the weighted sample were reported in Table 2.2 under the column “Weighted sample”. It was clear that in the original sample some of the covariates were significantly different among the three treatment groups, while covariates were not significantly different any more among the three treatment groups in the weighted sample. Thus, the confounding due to the variables in Table 2.2 was more likely removed, and the estimates for the accumulative costs based on the weighted sample were less biased. We estimated the accumulative cost trajectory under each treatment group using the sample mean (solid line), the method without IPTW (dotted line) and the methods with IPTW (dashed line) (see Figure 2.4 Panels A1-A3). We also estimated the accumulative cost difference due to treatments using the sample mean difference (solid line), the method without IPTW (dotted line) and the methods with IPTW (dashed line) (see Figure 2.4 Panels B1-B3).

The monthly cost profiles were obtained by the average of the monthly cost in the original sample. We obtained the estimated monthly cost by taking the derivative of the accumulative costs for each associated MSFM with IPTW and without IPTW, which were presented in Figure 2.4 Panels C1-C3. We also estimated the monthly cost difference due to treatments using the sample mean difference (solid line), the method without IPTW (dotted line) and the methods with IPTW (dashed line) (see Figure 2.4 Panels D1-D3).

In all panels, the thin dashed line running on two sides of the estimated curves provided a 95% point-wise confidence band for the estimated quantities based on the proposed method with IPTW.

From Table 2.2, it was clear that some confounding variables were significantly different in the original sample. However, they were not significantly different anymore in the weighted sample. Thus, we drew the conclusion based on the MSFM with IPTW. From Figure 2.4 (Panels A1-A3 and B1-B3), we conclude that the accumulative cost over time were not significantly different between A1 (Chemo $<$ Surgery) and A2 (Chemo $<$ Surgery $<$ Chemo), however the accumulative cost for A3 (Surgery $<$ Chemo) was significantly higher for the first five months than A1 and A2 groups, but the accumulative cost at A3 was significantly lower after seven months from diagnosis. From 2.4 (Panels C1-C3 and D1-D3), we saw the pattern and difference in monthly cost. While the average monthly cost differed around six months between A1 and A2, the overall monthly cost pattern between A1 and A2 remained similar. However, A3 differed significantly from A1 and A2 during the first eight months. While, initially for the first three months, the cost for A3 was significantly higher than A1 and A2, the cost for A3 dropped rapidly from second month to sixth month and the cost for A3 was significantly lower than A1 and A2 during third to eighth month. The monthly cost after eight month was similar among the three groups. The current approach not only helped analyse the cost pattern over time but also helped us detect the cost spikes in cost trajectory over time. It was evident from Figure 2.4 (Panels C1-C3) that the average monthly costs over time were substantial in the initial phase after diagnosis but reduced after a year of treatment.

Figure 2.4: The estimated accumulative cost trajectory under three different treatment combinations (Panels A1-A3) and the comparisons between different treatments (Panels B1-B3) and the estimated average monthly cost trajectory under three different treatment combinations (Panels C1-C3) and the comparisons between different treatments (Panels D1-D3) for stomach cancer patients.

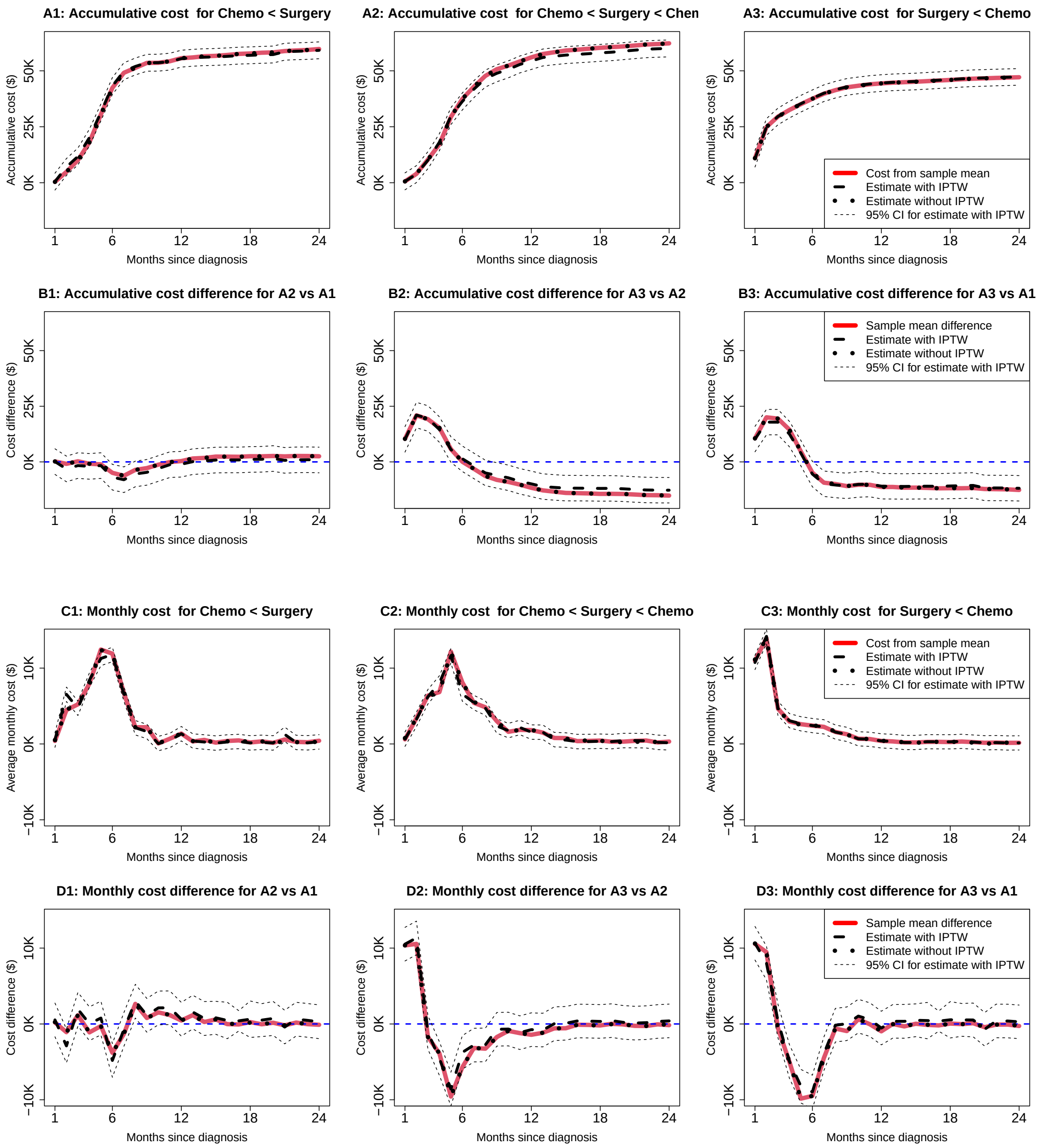


Table 2.2: Summarized statistics for possible confounding variables and point outcome variables in the original sample and weighted sample, stratified by treatment group

Confounding variables		Original sample				Weighted sample			
		A1	**A2**	**A3**	**p-value**	**A1**	**A2**	**A3**	**p-value**
Sample Size [n]		156	711	92		933	959	925	
Age [a][$Q2(Q1, Q3)$]		70(67, 75)	72(68, 77)	70(66, 74)	0.009	72(68, 76)	71(67, 77)	72(66, 76)	0.263
NCI Index [$Q2(Q1, Q3)$]		0(0, 1.3)	1.1(0, 1.7)	0(0, 1.3)	0.049	0(0, 1.6)	1.1(0, 1.7)	0(0, 1.3)	0.714
Gender [n(%)]	Male	107(68.6%)	447(62.9%)	66(71.7%)	0.131	571(61.3%)	619(64.6%)	587(63.5%)	0.673
	Female	49(31.4%)	264(37.1%)	26(28.3%)		361(38.7%)	340(35.4%)	338(36.5%)	
Race [n(%)]	White	138(88.5%)	477(67.1%)	71(77.2%)	< 0.001	683(73.2%)	686(71.5%)	665(71.9%)	0.889
	Others	18(11.5%)	234(32.9%)	21(22.8%)		250(26.8%)	273(28.5%)	260(28.1%)	
State [n(%)]	California	44(28.2%)	268(37.7%)	39(42.4%)	< 0.001	384(41.2%)	353(36.8%)	368(39.8%)	0.615
	Connecticut	15(9.6%)	47(6.6%)	< 11		72(7.8%)	70(7.3%)	68(7.4%)	
	Georgia	20(12.8%)	66(9.3%)	< 11		83(9%)	91(9.5%)	88(9.5%)	
	Hawaii	< 11	25(3.5%)	< 11		< 11	25(2.6%)	< 11	
	Iowa	< 11	15(2.1%)	< 11		29(3.1%)	29(3.1%)	27(3%)	
	Kentucky	< 11	39(5.5%)	< 11		55(5.9%)	50(5.2%)	53(5.8%)	
	Louisiana	< 11	65(9.1%)	< 11		57(6.1%)	73(7.6%)	70(7.6%)	
	Michigan	< 11	52(7.3%)	< 11		43(4.6%)	62(6.4%)	53(5.7%)	
	New Jersey	29(18.6%)	105(14.8%)	18(19.6%)		153(16.4%)	151(15.8%)	147(15.9%)	
	New Mexico	< 11	< 11	< 11		< 11	< 11	< 11	
	Utah	< 11	< 11	< 11		< 11	< 11	< 11	
	Washington	< 11	19(2.7%)	< 11		36(3.9%)	32(3.3%)	31(3.3%)	
Urban/Rural [n(%)]	Big Metro	96(61.5%)	412(57.9%)	57(62%)	0.583	599(64.2%)	565(59%)	546(59%)	0.386
	Metro	38(24.4%)	220(30.9%)	23(25%)		236(25.3%)	280(29.2%)	249(26.9%)	
	Urban	< 11	31(4.4%)	< 11		32(3.4%)	45(4.7%)	49(5.3%)	
	Less Urban	13(8.3%)	37(5.2%)	< 11		45(4.8%)	56(5.8%)	75(8.1%)	
	Rural	< 11	11(1.5%)	< 11		20(2.2%)	14(1.4%)	< 11	
Outcome variables		**A1**	**A2**	**A3**	**p-value**	**A1**	**A2**	**A3**	**p-value**
Survival Time [b][$Q2(Q1, Q3)$]		18(11, 31)	19(11, 34)	22(16, 32)	0.896	19(12, 36)	19(11, 34)	21(15, 31)	0.615
Accumulative cost [c][$Q2(Q1, Q3)$]		40.58(28.46, 69.84)	36.51(27.023, 55.15)	54.42(37.41, 80.01)	< 0.001	46.07(31.13, 69.845)	36.62(27.01, 55.16)	53.55(35.45, 80.01)	< 0.001

[a] Calculated in years; [b] Calculated in months; [c] Calculated in multiples of $1000 for 2 years after diagnosis

2.5 Discussion

In this study, we developed a MSFM model to evaluate the population-level cost trajectory under certain treatment and to estimate cost difference over time using penalized splines and IPTW. The proposed method provides consistent estimates for population-level cost trajectory and the average treatment effect on cost. We applied the proposed method to study the cost trajectory and treatment effects for patients with gastric cancer based on SEER Medicare database. We illustrated how the accumulative cost and monthly cost over time provided more information than a single time-point cost evaluation. Based on the proposed model, we can easily obtain monthly cost estimates by using the first derivative of the cumulative function modeled by I-splines, which are M-spline and provide valid inference on monthly cost.

The proposed method does not incorporate with censoring. However, in case of censoring, the probability of censoring can be modelled, and the inverse probability of non-censoring weighting along with IPTW could be applied to obtain unbiased population-level cost and evaluate treatment effect on cost. (Li et al. 2016).

CHAPTER 3

ESTIMATING HEALTHCARE COST USING PARAMETRIC CHANGE POINT MODELS

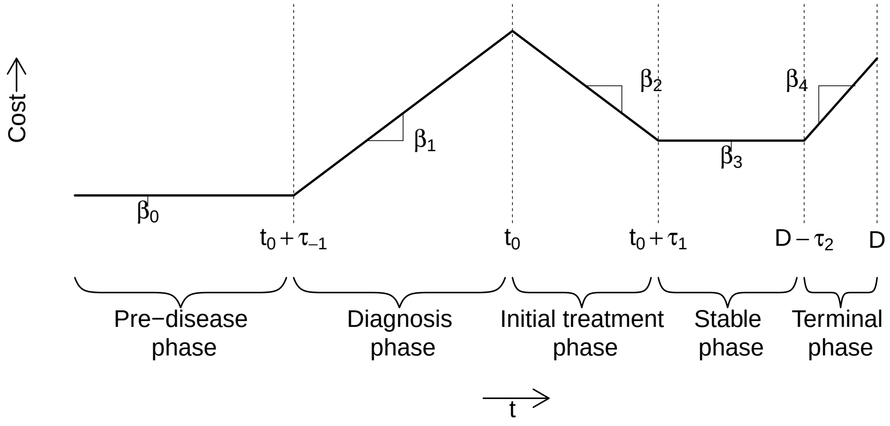
3.1 Introduction

Estimation of healthcare costs is crucial in the medical field. Healthcare cost associated with a certain disease could vary according to the received treatments and patients' characteristics and comorbidities. The healthcare cost could also change dramatically due to certain events such as diagnosis of cancer, intensive treatment, and death. With the increasing costs of healthcare delivery, budgetary constraints, and the aging population, it is important for policymakers and clinicians to know the cost trajectory in regards to patients diagnosed with certain diseases that have high incidences and are expensive to treat, e.g., cancer (Mihaylova et al. 2011; Wijeysundera et al. 2012). In the literature, lifetime cost due to certain disease is often studied (Lin et al. 1997; Bang and Tsiatis 2002; Basu et al. 2011; Li et al. 2016). However, it will be more informative to understand cost patterns and trajectories throughout disease progression and recovery.

Recent literature has suggested that there are multiple phases for the cost of a patient diagnosed with cancer (see, e.g., Brown et al. 2002; Wijeysundera et al. 2012; Tramontano et al. 2019). In particular, Wijeysundera et al. (2012) and Tramontano et al. (2019) suggested that the healthcare cost related to cancer can be divided into 4 phases. Understanding the different cost phases and identifying the points at which cost phases occur is thus crucial for policymakers and health insurance companies (Tramontano et al. 2019). However, while cost phases are considered critical parameters for cost analysis in most literature, the rigorous methods to estimate these parameters are lacking. For example, both Wijeysundera et al. (2012) and Tramontano et al. (2019) defined the cost phases and estimated the cost using the sample means of different sub-cohorts. This approach lacks the robustness to provide a statistical model for estimating the change points with varying patient cohort and different patients' characteristics and treatment choices. In this project, we provide a statistical framework for estimating the change points along with cost trajectory. Although change point detection techniques are widely used in economics and meteorology (Reeves et al. 2007; Paulus et al. 2015), the use of change points in medical cost is novel.

In evaluating the cost related to cancer, Wijeysundera et al. (2012) and Tramontano et al. (2019) proposed that the cancer attributable cost can be categorized into 4 different phases due to the required medical cares which include the diagnosis phase, initial treatment phase, stable phase and a terminal phase. It is also very important to capture the baseline cost prior to the diagnosis phase as it helps us identify the time point when the cost starts to increase and thus help in estimating the first change point before diagnosis. In this project, we propose a model with 5 healthcare cost phases, as illustrated in Figure 3.1. The 5 different cost phases are defined as: (P0) Pre-disease phase, a phase where patients may not be aware of any such disease or do not have any health conditions until a certain time $(t_0 + \tau_{-1})$ before diagnosis time at t_0, and the baseline cost parameter in this phase is denoted by β_0; (P1) Diagnosis phase, which is defined from

Figure 3.1: Illustration of cost phases and change points



the change point before diagnosis $(t_0 + \tau_{-1})$ to the time of diagnosis of the disease at t_0. During this time, extensive diagnostic tests and procedures could be performed, which drives the medical cost higher. The cost parameter associated with this phase is denoted by β_1; (P2) Initial treatment phase, which starts from the diagnosis time t_0 and lasts till the end of intensive treatment at time $(t_0 + \tau_1)$. The cost during this phase continues to decrease with time until it reaches a comparatively stable phase, and the cost parameter associated with this phase is denoted by β_2; (P3) Stable phase, which lies between the change point after diagnosis $(t_0+\tau_1)$ and becoming severely ill again right before the end of life at time $(D-\tau_2)$ which is τ_2 months before the time of death at time D. The cost parameter associated with this phase is denoted by β_3; and (P4) Terminal phase, which spans from becoming severely ill after the stable phase, at the terminal change point $(D - \tau_2)$ to the end of life at time D. The cost parameter associated with this phase is denoted as β_4.

It is well-known that medical costs are often impacted by treatment received and patient comorbidities (Austin 2011). To incorporate the patient level characteristics and treatment, we propose a piece-wise linear mixed model with change points at τ_{-1}, t_0, τ_1, and τ_2 and cost parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top$ to estimate the cost pattern over time, where cost parameters are based on the patients' covariates and are modeled and estimated as patient specific.

The rest of the paper is organized as follows. In section 3.2, we provide detailed information for the proposed method. In section 3.3, we apply the proposed method to estimate the cost trajectory for pancreatic cancer patients in SEER-Medicare $2005 - 2014$ database. The final section 3.4 is reserved for discussion.

3.2 The proposed model for change point detection and cost trajectory estimation

3.2.1 A simple change point model

Without loss of generality, t_0 is fixed as time 0, since we can always align the patients cost and standardize them from the time of diagnosis. Let $C(t)$ denote the cost during month t and D denote the time of death. Under the five phase assumption as illustrated in Figure 3.1, with the established change points models in literature (Reeves et al. 2007; Paulus et al. 2015), the cost trajectory during pre-disease phase and diagnosis phase can be captured by a 3-parameter model $C(t) = \beta_0 + \beta_1 \big(t - \tau_{-1}\big)^+$ for $t \leq 0$, where $A^+ = A$ if $A > 0$ and 0 otherwise, for a generic quantity A. The cost profile from diagnosis to death can be captured by a 5-parameter model, $C(t) = \beta_3 + \beta_2 \big(\tau_1 - t\big)^+ + \beta_4 \big(t - (D - \tau_2)\big)^+$ for $t > 0$, where D is the time of death. That is, the proposed piece-wise change model can be written as:

$$\begin{aligned} C(t) = &\Big[\beta_0 + \beta_1\big(t - \tau_{-1}\big)^+\Big] I_{(t<0)} \\ &+ \Big[\beta_3 + \beta_2\big(\tau_1 - t\big)^+ + \beta_4\big(t - (D - \tau_2)\big)^+\Big] I_{(t\geq 0)} + \epsilon(t) \end{aligned} \tag{3.1}$$

where $I_{(\cdot)}$ is the indicator function such that $I_{(A)} = 1$ if a generic event A is true and 0 otherwise, and $\epsilon(t) \sim N(0, \sigma^2 R)$ where R is an auto-regressive correlation matrix. It is clear that $E\big[C(t)\big]$ is a continuous function of time t with a possible discontinuity at 0, where the 3-parameter model for the first two phases and 5-parameter model for the last three phases meet. Since the cost function in practice is often continuous, we further impose the constraint $\beta_0 + \beta_1(0 - \tau_{-1}) = \beta_3 + \beta_2(\tau_1 - 0)$ to ensure the continuity of $E\big[C(t)\big]$ at $t = 0$. Thus, by replacing $\beta_3 = \beta_0 - \beta_1\tau_{-1} - \beta_2\tau_1$ in $C(t)$ the expectation of the cost function in equation (3.1) can be written as :

$$\begin{aligned} &E\big[C(t)\big] \\ =&\Big[\beta_0 + \beta_1\big(t - \tau_{-1}\big)^+\Big] I_{(t<0)} \\ &+ \Big[\beta_0 + \beta_1\big(0 - \tau_{-1}\big) - \beta_2\big(\tau_1 - 0\big) + \beta_2\big(\tau_1 - t\big)^+ + \beta_4\big(t - (D - \tau_2)\big)^+\Big] I_{(t\geq 0)} \\ =&\beta_0 + \beta_1\Big[\big(t - \tau_{-1}\big)^+ I_{(t<0)} - \big(\tau_{-1}\big) I_{(t\geq 0)}\Big] \\ &+ \beta_2\Big[\Big\{-\tau_1 + \big(\tau_1 - t\big)^+\Big\} I_{(t\geq 0)}\Big] + \beta_4\Big[\big(t - (D - \tau_2)\big)^+ I_{(t\geq 0)}\Big] = \mathbf{Z}^\top \boldsymbol{\beta} \end{aligned} \tag{3.2}$$

where $\mathbf{Z} = \Big(1, (t - \tau_{-1})^+ I_{(t<0)} - (\tau_{-1}) I_{(t\geq 0)}, \big(-\tau_1 + (\tau_1 - t)^+\big) I_{(t\geq 0)}, \big(t - (D - \tau_2)\big)^+ I_{(t\geq 0)}\Big)^\top$, $\boldsymbol{\beta} = \big(\beta_0, \beta_1, \beta_2, \beta_4\big)^\top$ are the regression parameters related to cost, and $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)^\top$ denote the three change points. Both $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are required to be estimated.

It is expected that the cost profile also depends on patients' characteristics such as age and comorbid conditions (Austin 2011). We further propose a model where the regression parameters are dependent on patients' variables and are patient specific, while the change points are population specific.

3.2.2 The proposed patient-level change point cost models

Let $(\mathbf{X}, \mathcal{T}, C, \delta)$ denote the random variable observed for a patient. $\mathbf{X}$ denotes a vector of p time invariant covariates of the patient, including patients' characteristics, medical history, and treatment information. $\mathcal{T} = (t_{-a}, \cdots, t_0, t_1, \cdots, t_b)^\top$ denotes the vector of time points where medical costs $C = (C_{-a}, \cdots, C_0, C_1, \cdots, C_b)^\top$ occurs. δ is an indicator variable on whether the patient died at the last observed time t_b.

For patients who died during the study period, we use the observed survival time in the change point model. For patients who are censored, we use the predicted survival time in the change point model. That is, if $\delta = 1$, then the survival time D is same as the last observed time point t_b. If $\delta = 0$, then the survival time D is the predicted survival time from a working model, like accelerated failure time (AFT) model. Based on equation (3.2), we have a terminal cost phase for a patient if the time for the last observation t_b satisfies the condition $t_b > D - \tau_2$.

Let $\{(\mathbf{X}_i, \mathcal{T}_i, C_i, \delta_i)\}_{i=1}^N$ denote the observed data for N patients in the study. The expected cost trajectory for each patient is assumed to theoretically follow the pattern specified in Figure 3.1. We propose that the regression parameters $\boldsymbol{\beta}$ in equation (3.2) are patient specific, say $\boldsymbol{\beta}_i$ for i^{th} patient, while the change point parameters are population specific. The cost for patient i at time t_{ij} can be written as $C_{ij} = \mathbf{Z}_{ij}^\top \boldsymbol{\beta}_i + \epsilon_{ij}$, where $C_{ij} = C(t_{ij})$, $\mathbf{Z}_{ij} = \Big(1, (t_{ij} - \tau_{-1})^+ I_{(t_{ij} < t_{i0})} + (t_{i0} - \tau_{-1}) I_{(t_{ij} \geq t_{i0})}, \big(- (\tau_1 - t_{i0}) + (\tau_1 - t_{ij})^+ \big) I_{(t_{ij} \geq t_{i0})}, \big(t_{ij} - (D_i - \tau_2)\big)^+ I_{(t_{ij} \geq t_{i0})} \Big)^\top$, $\boldsymbol{\beta}_i = \big(\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{4i}\big)^\top$. We model $\boldsymbol{\beta}_i$ as a linear function of the patient covariates $X_i = \big(X_{i1}, X_{i2}, \cdots, X_{ip}\big)^\top$, that is

$$\boldsymbol{\beta}_i = \begin{bmatrix} \gamma_{01} & \gamma_{02} & \cdots & \gamma_{0p} \\ \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \gamma_{41} & \gamma_{42} & \cdots & \gamma_{4p} \end{bmatrix} X_i \triangleq \Sigma\, X_i,$$

where $\Sigma \in \mathbb{R}^{4 \times p}$ is the parameter matrix to be estimated. Once, we have Σ estimated, we can use the estimated γ_{bp} $(b = 0, 1, 2, 4)$ to predict the significance of the covariate X_p on the patient-level cost parameter B_b for the i^{th} patient. Also, we can evaluate β_i from the estimated Σ and provide predictions of baseline cost β_{0i}, and the rate of change in the other cost phases from β_{1i}, β_{2i} and β_{4i} as well. Note that

$$C_{ij} = \mathbf{Z}_{ij}^\top \boldsymbol{\beta}_i + \epsilon_{ij} = \mathbf{Z}_{ij}^\top \Sigma\, X_i + \epsilon_{ij}. \tag{3.3}$$

To estimate Σ, we apply Roth's Columns Lemma (Roth 1934), which is popularly known as the "vec trick", and rewrite equation (3.3) as,

$$C_{ij} = \big[{X_i}^\top \bigotimes \mathbf{Z}_{ij}^\top\big]\, \mathbf{vec}\big(\Sigma\big) + \epsilon_{ij},$$

where

$$\big[{X_i}^\top \bigotimes \mathbf{Z}_{ij}^\top\big] = \big[X_{i1}\mathbf{Z}_{ij}^\top,\, X_{i2}\mathbf{Z}_{ij}^\top,\, \cdots,\, X_{ip}\mathbf{Z}_{ij}^\top\big] \in \mathbb{R}^{4p}$$

and $\mathbf{vec}\big(\Sigma\big) = \big(\gamma_{01}, \gamma_{11}, \gamma_{21}, \gamma_{41}, \gamma_{02}, \gamma_{12}, \gamma_{22}, \gamma_{42}, \cdots, \gamma_{0p}, \gamma_{1p}, \gamma_{2p}, \gamma_{4p}\big) \in \mathbb{R}^{4p}$. The costs of the i^{th} patient at the time sequence $\mathcal{T}_i = (t_{i(-a_i)}, \cdots, t_{i0}, t_{i1}, \cdots, t_{ib_i})^\top$ are

$$C_i = \Big[{X_i}^\top \bigotimes \mathbf{Z}_i^\top\Big]\, \mathbf{vec}\big(\Sigma\big) + \epsilon_i, \tag{3.4}$$

where

$$\mathbf{Z}_i^\top = \left(\mathbf{Z}_{i(-a_i)}^\top, \mathbf{Z}_{i(0)}^\top, \mathbf{Z}_{ib_i}^\top\right) \in \mathbb{R}^{(a_i+b_i+1)\times 4},$$

$$\left[{X_i}^\top \bigotimes \mathbf{Z}_i^\top\right] = \left(X_{i1}\mathbf{Z}_i^\top, X_{i2}\mathbf{Z}_i^\top, \cdots, X_{ip}\mathbf{Z}_i^\top\right) \in \mathbb{R}^{(a_i+b_i+1)\times 4p},$$

and ϵ_i is the vector of random noises for the i^{th} patient. To evaluate Σ and the change points, we expand equation (3.4) for the entire sample as,

$$C = \begin{bmatrix} {X_1}^\top \bigotimes \mathbf{Z}_1 \\ \vdots \\ {X_i}^\top \bigotimes \mathbf{Z}_i \\ \vdots \\ {X_N}^\top \bigotimes \mathbf{Z}_N \end{bmatrix} \mathbf{vec}\big(\Sigma\big) + \epsilon \tag{3.5}$$

where $C = \Bigg(C_1, \cdots, C_i, \cdots, C_N\Bigg)^\top \in \mathbb{R}^{\sum\limits_{i=1}^{N}(a_i+b_i+1)\times 1}$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_N)^\top$ is the vector of random noises, and $\epsilon_i \sim MVN(0, \sigma^2 R)$.

3.2.3 The estimation procedure

The proposed change point model involves the estimation of the regression parameters $\{\boldsymbol{\beta}_i\}_{i=1}^N$, the change points $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)$, and the time to death D_i for patients with $\delta_i = 0$. We first use the predicted survival time from the AFT model based on the covariates X_i to estimate D_i for patients with $\delta_i = 0$. Next, we specify how to estimate the change points, $\boldsymbol{\tau}$, which are population specific parameters in our proposed model. According to existing literature (see, e.g., Tramontano et al. 2019), for cancer patients, τ_{-1}, τ_1, and τ_2 are often considered as 2 months prior diagnosis, 6 months after diagnosis, and 6 months before death, respectively. We expand the possible set of values for each change point parameter. Namely, we assume that τ_{-1} is within a grid of possible values $(-6, -5, -4, -3, -2, -1)$, τ_1 is within a grid of possible values $(1, 2, 3, 4, 5, 6)$, and τ_2 is within a grid of possible values $(5, 6, 7, 8, 9, 10)$. Thus we totally have $6 \times 6 \times 6 (= 216)$ possible combinations for the value of $\boldsymbol{\tau}$. For each combination of $\boldsymbol{\tau}$, we fit a linear mixed-effects model (3.5) to estimate Σ and $\boldsymbol{\beta}_i$ and calculate the residual mean square errors (RMSE). $\hat{\boldsymbol{\tau}}$, the optimal choice for $\boldsymbol{\tau}$ is the one minimizing the RMSE, and the final estimate for Σ and $\boldsymbol{\beta}_i$ are obtained by fitting (3.5) using the optimally chosen $\hat{\boldsymbol{\tau}}$.

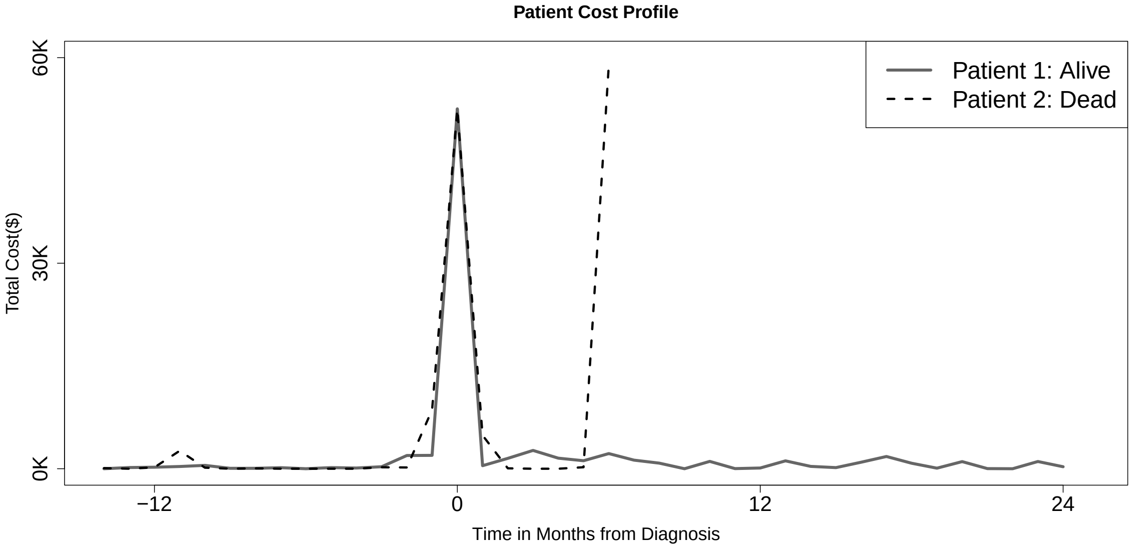
To make inference on the change points $\boldsymbol{\tau}$, we use non-parametric bootstrap re-sampling scheme to evaluate the accuracy and distribution of the change points. We obtain B bootstrap samples (say $B = 1000$) from the observed data, and then repeat the same estimation procedure for each bootstrap sample. Subsequently, we can get the distribution of $\hat{\boldsymbol{\tau}}$, which provides insight on the accuracy of the optimal selection for $\boldsymbol{\tau}$. In addition, we provide cost distributions in each phase by calculating the median and interquartile range (IQR) of the monthly cost in each phase. In the following section, we present a case study to illustrate the quantities involved.

3.3 Case Study

We applied our proposed method on SEER Medicare $2005 - 2014$ pancreatic cancer data to study the healthcare cost patterns over time and their relationship with different covariates. Our main goal is to estimate the population-specific change points along the patient-level cost trajectory. This will provide greater details about the cost pattern related to pancreatic cancer treatment over the course of time in different health care cost phases. These cost estimates and the estimation of phases will be important for healthcare systems and cancer control policy leaders and aid in the guidance of resource allocation for cancer care and research in the future (Tramontano et al. 2019). The study cohort included patients with pancreatic cancer specific diagnosis at primary site, histology, and behavioral code in the SEER database within year $2006-2013$ and having at least one treatment after diagnosis. If multiple combinations of diagnosis were found, the first such occurrence was considered. The comorbidities for a patient were obtained using the NCI comorbidity index based on one year data prior to pancreatic cancer diagnosis. NCI comorbidity index is cancer specific and excludes solid tumors, leukemia, and lymphomas as comorbid conditions (Klabunde et al. 2000). NCI comorbidity index was calculated by the 2014 NCI SAS Macro (NCI 2014) using the SEER-Medicare enrollment file (Pedsf), the inpatient file (Medpar), the outpatient file (Outpat) and the carrier claims file (NCH). The other covariates obtained from Pedsf included demographic variables (i.e., race, age, and sex), geographical variable (i.e, urban/rural and state) and cancer specific variables (i.e, stage of cancer). Treatment assigned for each patient was categorized as chemotherapy or surgery whichever came first after diagnosis. After applying the inclusion/exclusion criteria, we had 2899 patients in the sample, of whom 2277 patients died during the study period and 622 patients survived during the study period.

All costs from Medpar, Outpat and NCH files were obtained as the observed costs in the analysis. The costs were adjusted to 2014 cost rates using the Consumer Price Index (CPI-U) data (U.S. Department of Labor Bureau of Labor Statistic 2021). For each patient, we set the time of diagnosis as origin (say, $t_0 = 0$) and calculated the monthly cost as the sum of all costs occurred during that month. We restricted our time period of observations from to 14 months prior diagnosis to have just enough data to capture the pre-disease phase cost. Examples of cost trajectory for two randomly chosen patients from SEER Medicare $2005 - 2014$ pancreatic cancer data were illustrated in Figure 3.2. From this figure, it is clear that the patient who died (dashed line) had experienced five phases with a cost spike in the last phase (i.e., terminal phase), while the patient who survived had a very long period of stable phase (solid line) and did not enter the terminal phase during the study period.

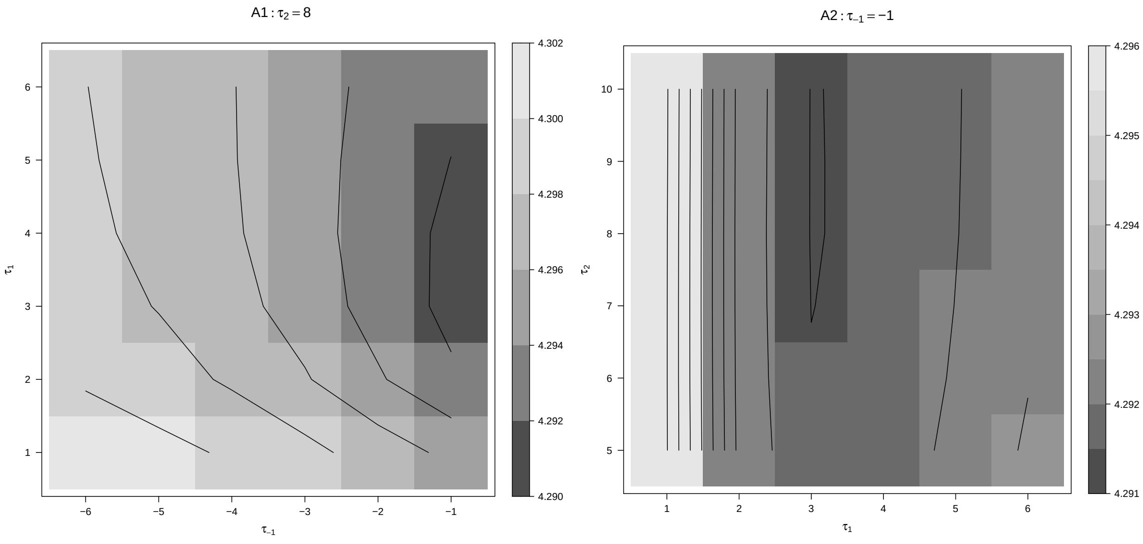
Figure 3.2: Examples of patient-level cost trajectory with diagnosis of pancreatic cancer at time $t_0 = 0$, from SEER Medicare $2005 - 2014$ pancreatic cancer database



As healthcare cost data are often highly skewed, we transformed the costs into a logarithmic scale to fit our proposed model, following previous literature (see, e.g., Manning and Mullahy 2001; Başer et al. 2004). We set 6 possible values for each change point: τ_{-1}, τ_1 and τ_2 range from -6 to -1, 1 to 6, and 5 to 10, respectively. In this case study, $\hat{\boldsymbol{\tau}}$, the optimal choice of $\boldsymbol{\tau}$, that minimized the RMSE of our proposed models were obtained as

$(-1, 3, 8)$. Figure 3.3 Panel A1 illustrated the contour plot of RMSE for the different choices of τ_{-1} and τ_1 with τ_2 fixed at the optimal value 8. Figure 3.3 Panel A2 provided the contour plot of RMSE for different choices of τ_1 and τ_2 with the optimal choice of τ_{-1} at -1. From Figure 3.3 is it evident that our optimal choice of $\hat{\boldsymbol{\tau}} = (-1, 3, 8)$ minimized the RMSE of the model among all other chosen values of $\boldsymbol{\tau}$. We further estimated the distribution of $\hat{\boldsymbol{\tau}}$ by performing 1000 bootstrap sampling. The relative frequency of the selected change points $\boldsymbol{\tau}$ was shown in Table 3.1. It is clear that our optimal selection for each change point $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2) = (-1, 3, 8)$ is the mode of the distribution of each change point.

Figure 3.3: The contour plot of RMSE for τ_{-1} versus τ_1 with τ_2 fixed at the optimal value 8 (Panel A1), and the contour plot of RMSE for τ_2 versus τ_1 with τ_{-1} fixed at the optimal value -1.



Once we obtained the optimal choice of change points, we then used it to estimate the regression parameter matrix Σ. We further estimated the patient level monthly cost during each phase, and summarized the distribution of monthly cost in-terms of median cost and interquartile range

Table 3.1: Distribution of $\boldsymbol{\tau}$ based on 1000 bootstrap samples

$\boldsymbol{\tau_{-1}}$						
Choices	-1	-2	-3	-4	-5	-6
Occurrence %	80%	20%	0%	0%	0%	0%

$\boldsymbol{\tau_{1}}$						
Choices	1	2	3	4	5	6
Occurrence %	0%	0%	84%	16%	0%	0%

$\boldsymbol{\tau_{2}}$						
Choices	5	6	7	8	9	10
Occurrence %	0%	2%	6%	38%	34%	20%

(IQR) for the study cohort during different cost phases in Table 3.2. It is clear that the monthly cost during the diagnosis phase was the highest and followed by initial treatment phase then terminal phase. The cost during stable phase was higher than the pre-disease phase.

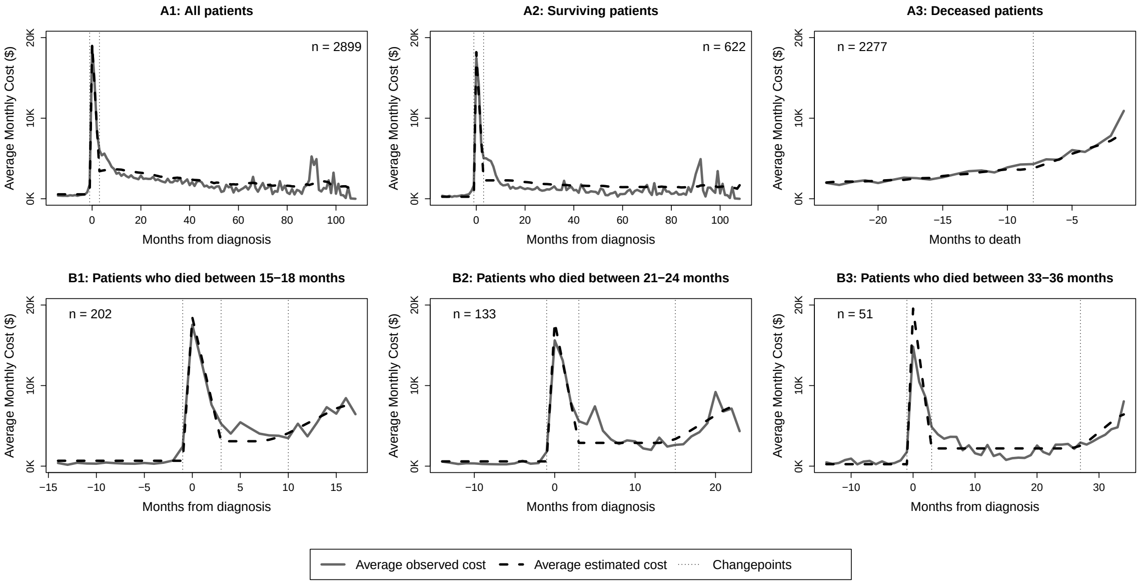
Table 3.2: Distribution of patient-level monthly costs during the five different cost phases

Cost phases [a]	Monthly cost: Median (Q1, Q3)
Pre-disease phase ($-14 \leq t \leq t_0 + \tau_{-1}$)	$\$0(\$0, \$185)$
Diagnosis phase ($t_0 + \tau_{-1} < t \leq t_0$)	$\$6654(\$1388, \$27351)$
Initial treatment phase ($t_0 < t \leq t_0 + \tau_1$)	$\$3368(\$181, \$9694)$
Stable phase ($t_0 + \tau_1 < t \leq D - \tau_2$)	$\$195(\$0, \$1652)$
Terminal phase ($D - \tau_2 < t \leq D$)	$\$683(\$0, \$5092)$

Finally, for the optimal choice of $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2) = (-1, 3, 8)$, we fitted our proposed model and estimate the monthly costs at log-scale, and we then transformed the cost into the original scale in dollar amount. Figure 3.4 provided the average observed monthly cost (solid line) and average predicted monthly cost (dashed lines) for different selected cohorts of the sample, along with the estimated change points at $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2) = (-1, 3, 8)$ (dotted lines). Figure 3.4 Panel A1 showed the cost and its estimation for the entire sample while A2 showed the results for the patients who survived during the study period. Panels B1-B3 provided the similar results but for different cohorts of the sample according to their survival time since diagnosis. Panel B1 showed the results from patients who died between $15-18$ months after diagnosis, Panel B2 showed the results from patients who died between $21-24$ months after diagnosis, and Panel B3 showed the results from patients who died between $33-36$ months after diagnosis. Note that, all these plots were aligned to the time of diagnosis as origin (at time 0) on the x-axis. Hence, we cannot provide τ_2 in the plot in Panel A1 and A2, as the death time D varies across patients. We provided similar results in Panels B1-B3 but with relative narrow range of survival time, with x-axis ranged to the longest survival time in their respective cohort. However, these representations did not accurately illustrate τ_2. All these panels are standardized to the time of diagnosis and since the time of death varies for each patient, the best representation of τ_2 would be in a panel where the survival time on the x-axis is aligned with respect to the time of death. Figure 3.4 Panel A3 presented the summarized cost aligned with death which illustrated the role of τ_2. Here the x-axis had been aligned to the time of death as origin (at time 0) and represented months prior the time of death. Hence, we plotted the optimal value τ_2 at -8 month, and it clearly showed an uptrend of cost starting from 8 months before death, indicating that the estimated τ_2 did capture the change point.

In literature, we often see that the major contributing factor to the change in cost are due to cancer stages (Wijeysundera et al. 2012; Tramontano et al. 2019). We have already included the cancer stage as a factor variable in our model. To further examine the impact of cancer stages on the cost, we stratify the 2899 pancreatic cancer patients into two groups: the first group consisting of patients with stage 1 and 2 pancreatic cancer and the second group consisting of patients with stage 3 pancreatic cancer.

Figure 3.4: Average observed and estimated monthly cost along with change points for different cohorts of pancreatic cancer patients using the proposed parametric change point approach



We used our model to predict the costs for both groups. The cost profiles for the stratified groups are provided in Figure 3.5 Panels A1 and B1 respectively, and each group is further divided into cohorts of surviving and deceased patients in Panels A2, A3 and B2, B3 respectively. It is clear that the cancer stage does have a significant impact on the cost. In particular, the cost after treatment for patients with stage 1 and 2 cancer stabilized after an initial high treatment phase as compared to the patients with stage 3 cancer. This is likely due to surgical treatment of stage 1 and 2 pancreatic cancer at the time of diagnosis, whereas stage 3 pancreatic cancer is more likely treated with chemotherapy only and hence have larger cost variation after initial treatment period, indicating more costs due to medical needs and possibly early deaths. Our model is able to capture the first and second change points as shown in Figure 3.5 Panels A1, A2, B1 and B2 and the final change point before death (see Panel A3 and B3).

Figure 3.5: Average observed and estimated monthly cost along with change points for different stages of pancreatic cancer patients using the proposed parametric change point approach

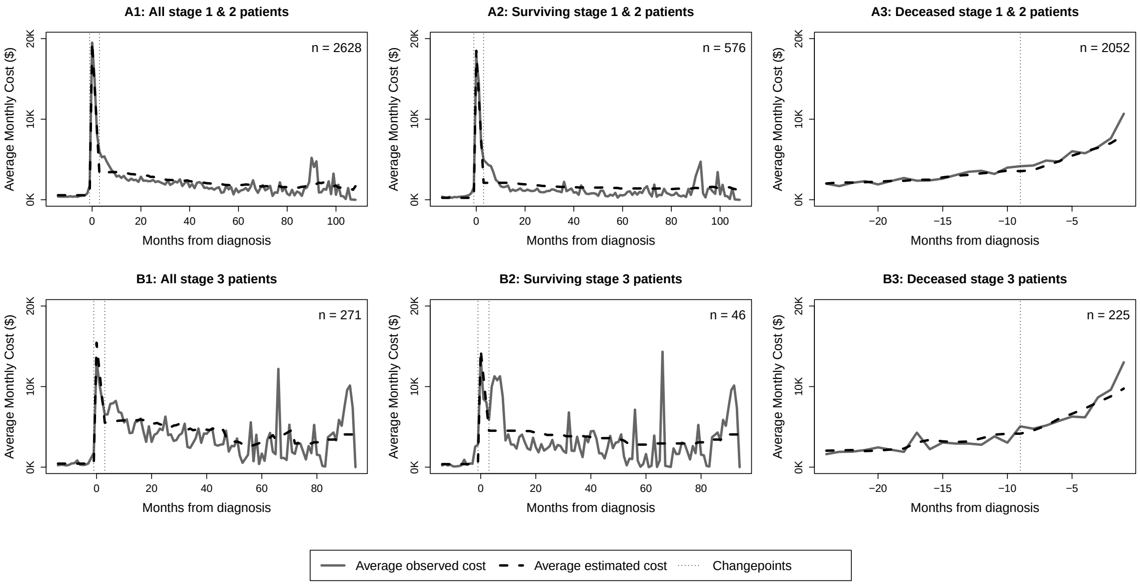


Table 3.1 and Figures 3.4 and 3.5 clearly showed that our estimation of the change points $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2) = (-1, 3, 8)$ matched the change patterns of the observed cost trajectories. We also see that our method captured the upward and downward trends in the cost trajectories with the peaks around a month after diagnosis and stabilizing after around 3 months. We also see how well the model captured the rise of the cost before death. Thus, the proposed model could help to understand the cost pattern, make plans for cost expenses, and improve an awareness for certain events such as diagnosis of disease or possible mortality.

3.4 Discussion

In this study, we proposed a parametric change point model to estimate population-level change points and patient level cost trajectory. The patient level cost trajectory could be associated with different patient's demographics, comorbidities, and treatment choices. The novel idea here was to model the cost trajectory based on change point detection models and further add another layer of accuracy by making the regression parameters depend on the patients' characteristics. The results from the case study showed that our proposed method provided a good estimate of the change points in the cost pattern which enabled us to accurately estimate the cost pattern over time. We demonstrated our method on pancreatic cancer data from SEER Medicare database. The estimation of the change points helped us accurately infer about the cost patterns for patients with certain disease. It further helped us understand the change points in cost pattern, which enable us to make more informed decisions regarding the funds for treatment over a period of time. Further, an upward trend in cost pattern after a stabilized cost period could also act as a warning sign of deteriorating health of the patient or recurrences of the disease.

CHAPTER 4

NON-PARAMETRIC MODELS FOR ESTIMATING MEDICAL COST AND CHANGE POINTS

4.1 Introduction

Estimating healthcare costs due to a certain disease over the course of the disease from diagnosis to the end of life, is very important to policy makers and clinicians, given the increasing costs of healthcare delivery, budgetary constraints, and the aging population (Mihaylova et al. 2011; Wijeysundera et al. 2012). It could provide important information if the cost pattern and trajectory over the course of disease progression and the recovery from the disease can be predicted accurately (Lin et al. 1997; Bang and Tsiatis 2002; Basu et al. 2011; Li et al. 2016). Based on the recent works, the lifetime cost of a patient can be divided into different phases due to diagnosis, treatment and mortality (Brown et al. 2002; Wijeysundera et al. 2012; Tramontano et al. 2019). In general, medical cost depends on patients' variable such as age, gender, and comorbid conditions (Austin 2011). It is of great interest to develop a flexible and suitable model that can capture the patient level cost as well as the population cost with great accuracy, which ca provide information for decision making.

Here, we investigate a flexible non-parametric approach to capture the cost patterns and change points. In particular, we propose to use I-spline basis functions along with patient-level regression coefficients to capture patient-level cost trajectory as well as cohort-level cost trajectory. Note that the first derivative of I-splines is M-splines, and it is straight forward to get the first derivative of the cost-trajectory using the relationship between I-splines and M-splines (Wan et al. 2017). Built on our previous investigation on cost phases and change points, we propose a flexible non-parametric approach to estimate the cost trajectory and change points using the penalized regressions splines and its first derivatives. In the following Section 4.2 we propose a non-parametric model for cost trajectory and provide estimating procedure for the trajectory and change points. In Section 4.3, we apply the proposed method to estimate the cost trajectory for pancreatic cancer patients in SEER-Medicare $2005-2014$ database. Section 4.4 is reserved for discussions.

4.2 The proposed non-parametric model for cost trajectory estimation and change point detection

4.2.1 The proposed non-parametric model

Let $(\mathbf{X}, \mathcal{T}, C, D)$ denote the random variable observed for a patient, where $\mathbf{X}$ denotes a vector of p time invariant covariates of the patient, including patients' characteristics, medical history, and treatment information. $\mathcal{T} = (t_{-a}, \cdots, t_0, t_1, \cdots, t_b)^\top$ denotes the vector of time points where medical costs $C = (C_{-a}, \cdots, C_0, C_1, \cdots, C_b)^\top$ occurred. D is an indicator variable whether the patient died at the last observed time point t_b. Without loss of generality, we align the time of diagnosis of a patient t_0 as time 0. Let $\{(\mathbf{X}_i, \mathcal{T}_i, C_i, D_i)\}_{i=1}^N$ denote the observed data for N patients in the study.

We propose a non-parametric cost model to capture the monthly cost trajectory over time which is expressed as the linear combination of the I-spline basis functions. Let us denote K knots in the range of observed time points as $(\tau_1, \tau_2, \cdots, \tau_K)$ with

$$\min_{i=1,\cdots,N}(t_{i(-a_i)}) = \tau_1 < \tau_2 < \cdots < \tau_K = \max_{i=1,\cdots,N}(t_{ib_i}).$$

The interior knots were taken based on the equally-spaced quantiles of t_{ij} $(i = 1, 2, \cdots, N;\ j = -a_i, \cdots, 0, \cdots, b_i)$. Let us denote $I^3_\kappa(t)$ $(\kappa = 1, \cdots, K-2)$ as the cubic I-spline basis functions based on the K knots, where $I^3_\kappa(t)$ is a smooth monotonic function ranged between $[0, 1]$ with support interval on $[\tau_\kappa, \tau_{\kappa+3}]$ (Wan et al. 2017). The I-spline basis functions can be constructed using the *iSpline* function from the R-package *splines2*. We propose to use the linear combination of the I-spline basis functions to model the cost trajectory over time for the i^{th} patient $(i = 1, 2, \cdots, N)$ as

$$C_i = \begin{bmatrix} I^3_1(t_{i(-a_i)}) & I^3_2(t_{i(-a_i)}) & \cdots & I^3_{K-2}(t_{i(-a_i)}) \\ \vdots & \vdots & \vdots & \vdots \\ I^3_1(t_{ij}) & I^3_2(t_{ij}) & \cdots & I^3_{K-2}(t_{ij}) \\ \vdots & \vdots & \vdots & \vdots \\ I^3_1(t_{ib_i}) & I^3_2(t_{ib_i}) & \cdots & I^3_{K-2}(t_{ib_i}) \end{bmatrix} \boldsymbol{\beta}_i + \epsilon_i = \mathcal{I}(\mathcal{T}_i)^\top \boldsymbol{\beta}_i + \epsilon_i, \qquad (4.6)$$

where $C_i = C(\mathcal{T}_i)$, $\mathcal{I}(\mathcal{T}_i)$ is known once the knots and $\mathcal{T}_i = (t_{i(-a_i)}, \cdots, t_{i0}, \cdots, t_{ib})^\top$ are specified, the regression parameter $\boldsymbol{\beta}_i = \left(\beta_{1i}, \beta_{2i}, \cdots \beta_{K-2,i}\right)^\top$ are patient specific which are assumed to depend on patient's variables X_i, and $\epsilon_i = (\epsilon_{i(-a_i)}. \cdots, \epsilon_{ij}, \cdots, \epsilon_{ib_i})^\top \sim MVN(0, \sigma^2 R)$, where R is an auto-regressive correlation matrix.

It is noted that the cost usually increases and has an upward trend towards the end of life for patients who are dying, which results in quite different cost trajectory than other time periods in their lives. To incorporate this upward trend, we constructed a smooth function for patients who died during the study period (i.e. for those with $D_i = 1$) to capture the cost pattern at the end of life. That is, the cost over time for patients who died during the study would be the combination of two smooth functions. To be specific, we add another smoothing spline function aligned to the time of death t_{ib_i} for the patients who died (i.e., $D_i = 1$). That is, we standardize the time points as $\mathcal{T}_i^* = (t^*_{i(-(a_i+b_i))}, \cdots, t^*_{i(-1)}, t^*_{i0})^\top$ for patients who died during the study period, where $t^*_{ij} = t_{ij} - t_{ib_i}$ if $D_i = 1$. Since the cost related to death is usually effective within 12 months prior to death, we consider knots ranged between $[-12, 0]$ months prior to death. Let us take L knots in the range of time points aligned to the time of death, say, $-12 = \tau_1^* < \tau_2^* < \cdots < \tau_L^* = 0$ with the interior knots based on the equally-spaced time-points of t^*_{ij} $(i \in \{i : D_i = 1\};\ j \in [-12, 0])$. Let us denote $I^3_\iota(t^*)$ $(\iota = 1, \cdots, L-2)$ as the cubic I-spline basis functions based on the L knots, where $I^3_\iota(t^*)$ is a smooth monotonic function ranged between $[0, 1]$ with support interval on $[\tau^*_\iota, \tau^*_{\iota+3}]$ (Wan et al. 2017). The cost trajectory for the patients who died during the study period involves a terminal phase modelled by the penalized splines with knots as $(\tau_1^* < \tau_2^* < \cdots < \tau_L^*)$.

To incorporate each patients' study duration in the model we use a block diagonal matrix S_i which captures the patient's study duration. The updated cost trajectory over time for the i^{th} patient $(i = 1, 2, \cdots, N)$ is

$$C_i = \mathcal{I}(\mathcal{T}_i)^\top S_i \boldsymbol{\beta}_i + D_i \mathcal{I}_{\mathcal{D}}(\mathcal{T}_i^*)^\top \boldsymbol{\alpha}_i + \epsilon_i \tag{4.7}$$

where $S_i = \begin{bmatrix} \mathbf{I}_{(a_i+b_i)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(K-2)\times(K-2)}$ captures the study duration of the i^{th} patient, $\mathbf{I} \in \mathbb{R}^{(a_i+b_i)\times(a_i+b_i)}$ is an identity matrix,

$$\mathcal{I}_{\mathcal{D}}(\mathcal{T}_i^*) = \begin{bmatrix} I_1^3(t^*_{i(-12)}) & I_2^3(t^*_{i(-12)}) & \cdots & I_{L-2}^3(t^*_{i(-12)}) \\ \vdots & \vdots & \vdots & \vdots \\ I_1^3(t^*_{ij}) & I_2^3(t^*_{ij}) & \cdots & I_{L-2}^3(t^*_{ij}) \\ \vdots & \vdots & \vdots & \vdots \\ I_1^3(t^*_{i0}) & I_2^3(t^*_{i0}) & \cdots & I_{L-2}^3(t^*_{i0}) \end{bmatrix}^\top$$ is known once the knots and t^*_{ij} are specified, the regression parameter $\boldsymbol{\alpha}_i = \left(\alpha_{1i}, \alpha_{2i}, \cdots \alpha_{L-2,i}\right)^\top$ are patient specific which are dependent on patient's variables and D_i is the indicator variable for death. That is, we model both $\boldsymbol{\beta}_i$ and $\boldsymbol{\alpha}_i$ as linear functions of the patient covariates $X_i = \left(X_{i1}, X_{i2}, \cdots, X_{ip}\right)^\top$:

$$\boldsymbol{\beta}_i = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{(K-2)1} & \sigma_{(K-2)2} & \cdots & \sigma_{(K-2)p} \end{bmatrix} X_i \triangleq \Sigma\, X_i$$

and

$$\boldsymbol{\alpha}_i = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1p} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{(L-2)1} & \delta_{(L-2)2} & \cdots & \delta_{(L-2)p} \end{bmatrix} X_i \triangleq \Delta\, X_i$$

where $\Sigma \in \mathbb{R}^{(K-2)\times p}$ and $\Delta \in \mathbb{R}^{(L-2)\times p}$ are the parameter matrices to be estimated, and $X_i = \left(X_{i1}, X_{i2}, \cdots, X_{ip}\right)^\top$ is the observed patient's variables. Hence, we rewrite equation (4.7) as,

$$C_i = \mathcal{I}(\mathcal{T}_i)^\top S_i \boldsymbol{\beta}_i + D_i\, \mathcal{I}_{\mathcal{D}}(\mathcal{T}_i^*)^\top \boldsymbol{\alpha}_i + \epsilon_i = \mathcal{I}(\mathcal{T}_i)^\top\, S_i\, \Sigma\, X_i + D_i\, \mathcal{I}_{\mathcal{D}}(\mathcal{T}_i^*)^\top\, \Delta\, X_i\, + \epsilon_i \quad (4.8)$$

To estimate Σ and Δ we apply Roth's Columns Lemma (Roth 1934), also popularly known as the ”vec trick”, and rewrite equation (4.8) as follows:

$$C_i = \Big[X_i{}^\top \bigotimes \mathcal{I}(\mathcal{T}_i)^\top\, S_i\Big]\,\mathbf{vec}\big(\Sigma\big) + \Big[X_i{}^\top \bigotimes D_i\,\mathcal{I}_{\mathcal{D}}(\mathcal{T}_i^*)^\top\Big]\,\mathbf{vec}\big(\Delta\big) + \epsilon_i \qquad (4.9)$$

where

$$\Big[X_i{}^\top \bigotimes \mathcal{I}(\mathcal{T}_i)^\top S_i\Big] \in \mathbb{R}^{(a_i+b_i+1)\times(K-2)p},$$

$$\mathbf{vec}\big(\Sigma\big) = \big(\sigma_{11},\cdots,\sigma_{(K-2)1},\cdots,\sigma_{(K-2)p}\big) \in \mathbb{R}^{(K-2)p},$$

$$\Big[X_i{}^\top \bigotimes D_i\,\mathcal{I}_{\mathcal{D}}(t_{ij}^*)^\top\Big] \in \mathbb{R}^{(a_i+b_i+1)\times(L-2)p} \text{ and}$$

$$\mathbf{vec}\big(\Delta\big) = \big(\delta_{11},\cdots,\delta_{(L-2)1},\cdots,\delta_{(L-2)p}\big) \in \mathbb{R}^{(L-2)p}.$$

To estimate Σ and Δ, we expand equation (4.9) for the entire sample as,

$$C = \begin{bmatrix} X_1{}^\top \bigotimes \mathcal{I}(\mathcal{T}_1)^\top\, S_1 \\ \vdots \\ X_2{}^\top \bigotimes \mathcal{I}(\mathcal{T}_2)^\top\, S_2 \\ \vdots \\ X_N{}^\top \bigotimes \mathcal{I}(\mathcal{T}_N)^\top\, S_N \end{bmatrix} \mathbf{vec}\big(\Sigma\big) + \begin{bmatrix} X_1{}^\top \bigotimes D_1\,\mathcal{I}_{\mathcal{D}}(\mathcal{T}_1^*)^\top \\ \vdots \\ X_2{}^\top \bigotimes D_2\,\mathcal{I}_{\mathcal{D}}(\mathcal{T}_2^*)^\top \\ \vdots \\ X_N{}^\top \bigotimes D_N\,\mathcal{I}_{\mathcal{D}}(\mathcal{T}_N^*)^\top \end{bmatrix} \mathbf{vec}\big(\Delta\big) + \epsilon \quad (4.10)$$

where $C = \big(C_1,\cdots,C_i,\cdots,C_N\big)^\top \in \mathbb{R}^{\sum_{i=1}^{N}(a_i+b_i+1)}$ and ϵ is the vector of random noises, and $\epsilon_i \sim MVN(0,\sigma^2 R)$.

4.2.2 Estimation procedure

We use penalized regression models to estimate Σ and Δ from equation (4.10). It can be performed by the function *gam* in R-package *mgcv*. Once we obtain the estimates for the parameters Σ and Δ, we can estimate the cost curve and determine the change points by studying the cost trajectory and its first order derivative. Note that the first order derivative of I-spline is a M-spline. We thus can obtain the first derivative of the cost function using the same regression parameters Σ and Δ but with M-spline basis functions instead of I-spline functions. That is,

$$C' = \begin{bmatrix} {X_1}^\top \bigotimes \mathcal{M}(\mathcal{T}_1)^\top S_1 \\ \vdots \\ {X_2}^\top \bigotimes \mathcal{M}(\mathcal{T}_2)^\top S_2 \\ \vdots \\ {X_N}^\top \bigotimes \mathcal{M}(\mathcal{T}_N)^\top S_N \end{bmatrix} \mathbf{vec}\big(\Sigma\big) + \begin{bmatrix} {X_1}^\top \bigotimes D_1\, \mathcal{M}_{\mathcal{D}}(\mathcal{T}_1^*)^\top \\ \vdots \\ {X_2}^\top \bigotimes D_2\, \mathcal{M}_{\mathcal{D}}(\mathcal{T}_2^*)^\top \\ \vdots \\ {X_N}^\top \bigotimes D_N\, \mathcal{M}_{\mathcal{D}}(\mathcal{T}_N^*)^\top \end{bmatrix} \mathbf{vec}\big(\Delta\big) \quad (4.11)$$

where $\mathcal{M}(\mathcal{T}_i) == \begin{bmatrix} M_1^3(t_{i(-a_i)}) & M_2^3(t_{i(-a_i)}) & \cdots & M_{K-2}^3(t_{i(-a_i)}) \\ \vdots & \vdots & \vdots & \vdots \\ M_1^3(t_{ij}) & M_2^3(t_{ij}) & \cdots & M_{K-2}^3(t_{ij}) \\ \vdots & \vdots & \vdots & \vdots \\ M_1^3(t_{ib_i}) & M_2^3(t_{ib_i}) & \cdots & M_{K-2}^3(t_{ib_i}) \end{bmatrix}^\top$ and

$$\mathcal{M}_{\mathcal{D}}(\mathcal{T}_i^*) = \begin{bmatrix} M_1^3(t^*_{i(-12)}) & M_2^3(t^*_{i(-12)}) & \cdots & M_{L-2}^3(t^*_{i(-12)}) \\ \vdots & \vdots & \vdots & \vdots \\ M_1^3(t^*_{ij}) & M_2^3(t^*_{ij}) & \cdots & M_{L-2}^3(t^*_{ij}) \\ \vdots & \vdots & \vdots & \vdots \\ M_1^3(t^*_{i0}) & M_2^3(t^*_{i0}) & \cdots & M_{L-2}^3(t^*_{i0}) \end{bmatrix}^\top .$$

Here $M_\kappa^3(t)$ and $M_\iota^3(t^*)$ are the cubic M-spline basis functions based on the K and L knots specified above.

In literature, medical cost are often divided into different cost phases, and the change points of the phases could provide important information (Wijeysundera et al. 2012, Tramontano et al. 2019). We are particularly interested in the 3 change points: τ_{-1}, the change point from pre-diagnosis phase to diagnosis phase; τ_1, the change point from treatment phase to stabilized phase; τ_2, the change point from stabilized phase to terminal phase. We proposed the following rules to detect the three change points, $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)$:

Step 1. Calculate the 95% confidence interval of the first derivative of the cost function from 12 months to 6 months prior diagnosis. 6 months prior diagnosis is considered because it is a reasonable window during which cost related to the disease diagnosis is unlikely to occur. Set the two boundary centered around zero for the first derivative, which are used to evaluate the change points.

Step 2. The change point from pre-cancer phase to diagnosis phase (τ_{-1}) is estimated as the earliest time point prior diagnosis where the value of $C'(j)$ exceeds the upper confidence interval boundary.

Step 3. The change point (τ_1) from treatment phase to stabilized phase is assessed as the latest time point after diagnosis where the value of $C'(j)$ traveled from negative value into the lower boundary of the 95% confidence interval calculated in Step 1.

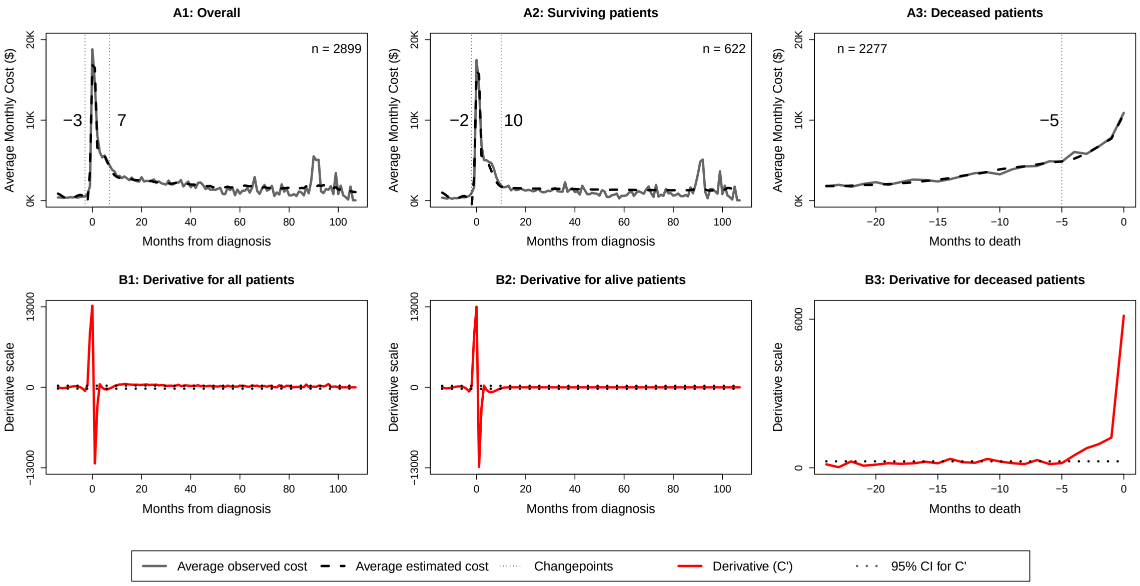
Step 4. The change point (τ_2) from stabilized phase to terminal phase is assessed as the earliest time point before death where the value of $C'(j)$ exceeds the upper boundary of the 95% confidence interval calculated in Step 1.

4.3 Case Study

We applied our proposed method to study the healthcare cost pattern for patients with pancreatic cancer using the SEER Medicare $2005-2014$ database. Our main goal here was to estimate the population-specific change points along the patient-level cost trajectory. The study cohort included patients with pancreatic cancer specific diagnosis at primary site, histology, and behavioral code in the SEER database within year $2006 - 2013$ and having at least one treatment after diagnosis. If multiple combinations of diagnosis were found, the first such occurrence was considered. The comorbidity for a patient were obtained by using the NCI comorbidity index obtained based on one year data prior to pancreatic cancer diagnosis. NCI comorbidity index is cancer specific and excludes solid tumors, leukemia, and lymphomas as comorbid conditions (Klabunde et al. 2000). The NCI comorbidity index was calculated using the 2014 NCI SAS Macro from the NCI website (NCI 2014) using the SEER-Medicare enrollment file (Pedsf) and the diagnostic codes in the inpatient file (Medpar), the outpatient file (Outpat) and the carrier claims file (NCH). The other covariates obtained from Pedsf included demographic variables (i.e. race, age, sex) and geographical variable (such as urban/rural and state). Treatment assigned for each patient was categorized as chemotherapy or surgery whenever it came first after diagnosis. After applying the inclusion/exclusion criteria, we had 2899 patients in the sample, of whom 2277 patients died during the study period and 622 patients survived the study period.

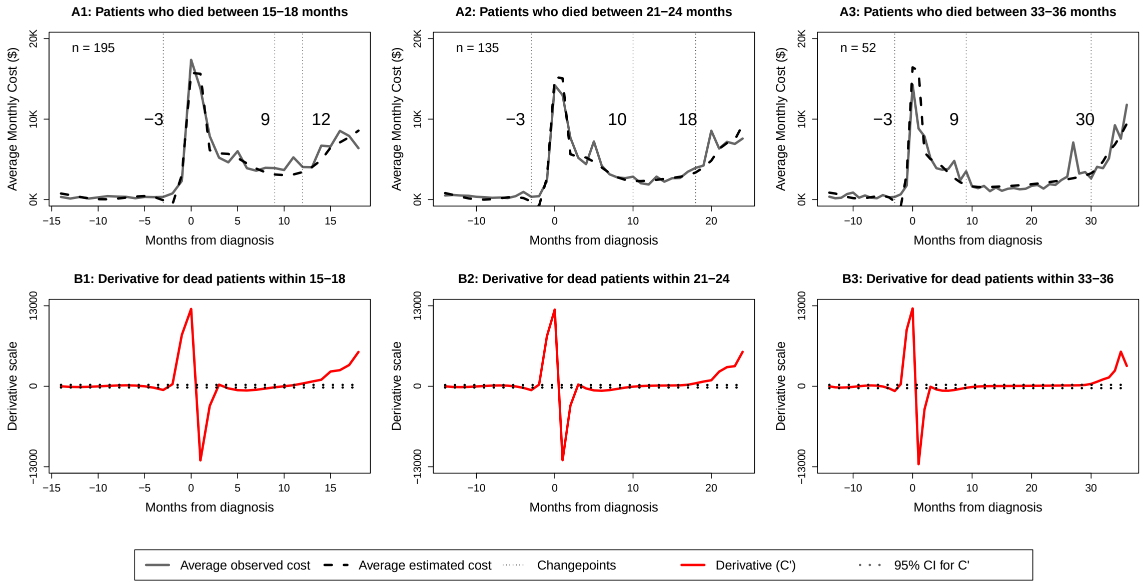
All medical cost from Medpar, Outpat and NCH files were obtained as the primary outcome in the analysis. The costs were standardized to 2014 cost rates using the Consumer Price Index (CPI-U) data (U.S. Department of Labor Bureau of Labor Statistic 2021). For each patient, we set the time of diagnosis as origin $(t_0 = 0)$ and calculate the monthly cost as the sum of all costs occurred during that month. We restrict our time period of observation before diagnosis to 12 months to have enough data to capture the pre-disease phase cost. We fit our proposed model to the data to estimate the cost trajectory. We perform a penalized regression and predict the cost trajectory and estimate the change points from the first derivative and the cost patterns. Figure 4.1 and 4.2 Panels A1-A3 provided the average observed monthly cost (solid line) and average predicted monthly cost (dashed lines) for different selected cohorts of the sample, along with the estimated change points (dotted lines). Each panel represented a different sub-cohort. In Figure 4.1 Panel A1 represented the entire sample, A2 represented the patients who survived during the study period, and A3 represented all those who died during the study period. Similarly, Figure 4.2 Panel A1 showed the results from patients who died between $15-18$ months after diagnosis, Panel A2 showed the results from patients who died between $21-24$ months after diagnosis, and Panel A3 showed the results from patients who died between $33-36$ months after diagnosis. We then obtained the first order derivative for the curve at each time point. The derivatives were displayed below each panel of A1-A3 for ease of visualization and were displayed as solid lines in both Figures 4.1 and 4.2 in Panels B1-B3. In Figure 4.1, Panel B1 represented the derivative for the entire sample, B2 represented the derivative for patients who survived during the study period and B3 represented the derivative for all those who died during the study period. The same followed in Figure 4.2 Panels B1-B3. We applied our proposed change point detection technique to estimate the change points. The change points were provided in the plots. Note that, in Figure 4.1 Panels A1 and A2 were aligned to the time of diagnosis as origin on the x-axis. Hence, we cannot provide the terminal change point in the plot in Panel A1 and A2 as it is related to the time of death. We provided the terminal change point results in Figure 4.2 Panels A1-A3 but with relative narrow range of survival time, with x-axis ranged to the longest survival time in their respective cohort. However, these representations did not accurately illustrate the change point from stable phase to terminal phase. All these panels are standardized to the time of diagnosis, while the time of death varies for each patient. The best representation of the change point from stable phase to terminal phase would be in a panel where the survival time on the x-axis is aligned with the time of death. Figure 4.1 Panel A3 represented the most accurate illustration of the terminal change point dependent upon death which describes the cost change time from death.

Figure 4.1: Observed and estimated average cost trajectory along with change points for different cohorts in Panels A1-A3 and the estimated first order derivative of the cost trajectory in Panels B1-B3 for pancreatic cancer patients as estimated by the proposed non-parametric approach



It can be clearly seen that the estimated change points matched the change of cost in different phases. That is, the diagnosis phase started $2-3$ month before diagnosis, followed by an initial treatment phase of $7-10$ months after diagnosis and then a stabilized phase. For the deceased population, the cost dramatically increased in the last $5-6$ months of their

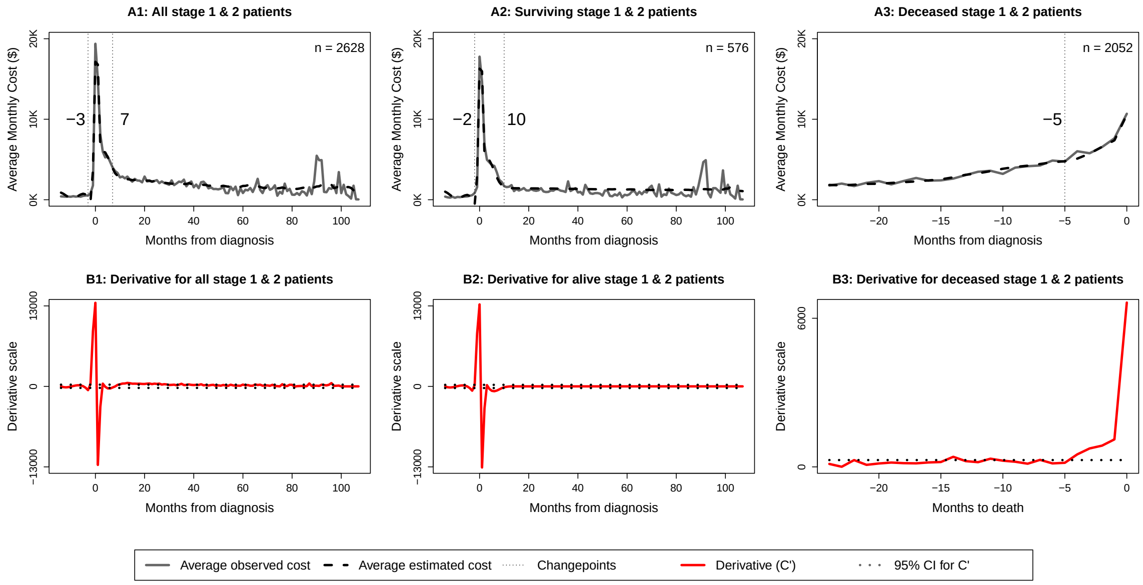
Figure 4.2: Observed and estimated average cost trajectory along with change points for different death cohorts in Panels A1-A3 and the estimated first order derivative of the cost trajectory in Panels B1-B3 for pancreatic cancer patients as estimated by the proposed non-parametric approach



lives.

In literature, we often see that the major contributing factor to the change in cost phases are due to cancer stages (Wijeysundera et al. 2012; Tramontano et al. 2019). We have already included the cancer stage as a factor variable in our model. To further examine the impact of cancer stages on the cost, we stratify the 2899 pancreatic cancer patients into two groups: the first group consisting of patients with stage 1 and 2 pancreatic cancer and the second group consisting of patients with stage 3 pancreatic cancer. We used our proposed model to predict the costs for both groups and estimated the change points based on the derivatives. The observed (solid line) and estimated (dashed lined) cost profiles for the stratified groups are provided in Figure 4.3 for stage 1 and 2 cancer patients and Figure 4.4 for stage 3 cancer patients. We first presented the cost pattern for the patients in each stage group (Panels A1) then each group is further divided into cohorts of surviving and deceased patients in Panels A2 and A3. The derivatives were displayed below each panel of A1-A3 for ease of visualization and were displayed as solid lines in both Figure 4.3 and 4.3 in Panels B1-B3. The change points derived from our proposed method is plotted in the figures as well. It is clear that the cancer stage did have a significant impact on the cost. In particular, the cost after treatment for patients with stage 1 and 2 cancer stabilizes after an initial high treatment phase as compared to the patients with stage 3 cancer. This is likely due to surgical treatment of stage 1 and 2 pancreatic cancer at the time of diagnosis, whereas stage 3 pancreatic cancer is more likely treated with chemotherapy only and hence have larger cost variation after initial treatment period, indicating more costs due to medical needs and possibly early deaths. Our model is able to capture the first and second change points as shown in Figure 4.3 Panels A1, A2, B1 and B2 and the final change point before death (see Panel A3 and B3).

Figure 4.3: Observed and estimated average cost trajectory along with change points for stage 1 and 2 patients in Panels A1-A3 and the estimated first order derivative of the cost trajectory in Panels B1-B3 for pancreatic cancer patients as estimated by the proposed non-parametric approach


A1: All stage 1 & 2 patients

A2: Surviving stage 1 & 2 patients

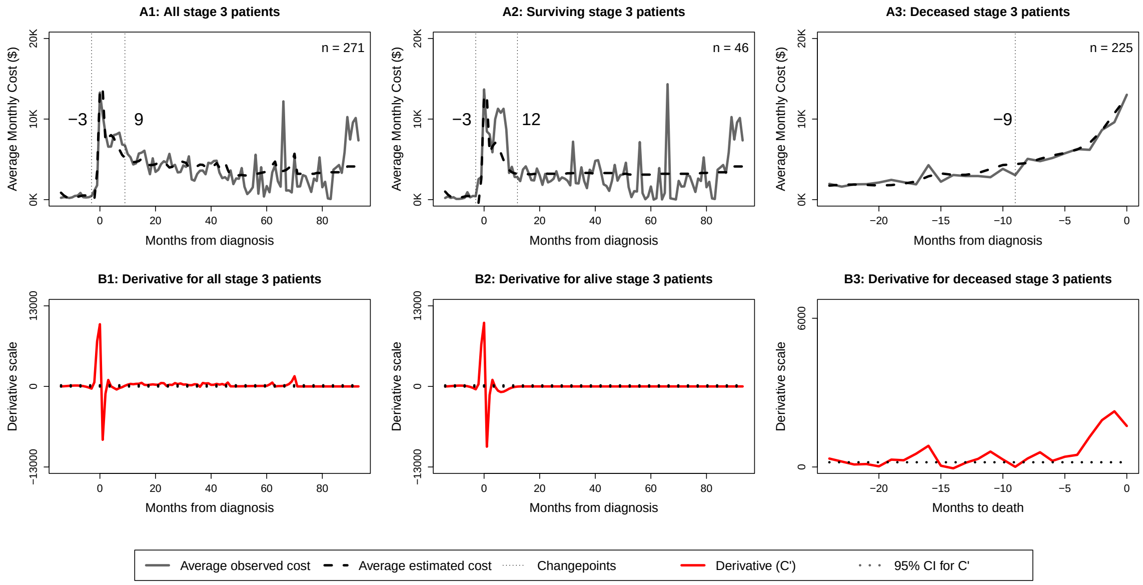
A3: Deceased stage 1 & 2 patients

B1: Derivative for all stage 1 & 2 patients

B2: Derivative for alive stage 1 & 2 patients

B3: Derivative for deceased stage 1 & 2 patients

Figure 4.4: Observed and estimated average cost trajectory along with change points for stage 3 patients in Panels A1-A3 and the estimated first order derivative of the cost trajectory in Panels B1-B3 for pancreatic cancer patients as estimated by the proposed non-parametric approach


A1: All stage 3 patients

A2: Surviving stage 3 patients

A3: Deceased stage 3 patients

B1: Derivative for all stage 3 patients

B2: Derivative for alive stage 3 patients

B3: Derivative for deceased stage 3 patients

4.4 Discussion

In this study, we proposed a flexible non-parametric change point models to estimate cohort-level change points and patient level cost trajectory. The patient level cost trajectory could be associated with different patient's demographics, comorbidity, and treatment choices. The change point and cost trajectory is also dependent on the patients end of life information. To incorporate these we proposed a model with basis functions based on both the time of diagnosis and the time of death for the deceased patients. We used I-spline basis functions to model our cost trajectory. Once we had the cost curve estimated, we then used the first derivative of the spline function, particularly the M-spline basis function to estimate the change points on the cost curve. The results from the case study showed that our proposed method provided an appropriate estimates of the change points in the cost pattern. The accurate estimation of the change points can help infer about the cost patterns due to a disease diagnosis. It further helped to understand the change points in cost pattern, which enable us to make more informed decisions regarding the funds for treatment over a period of time. Further, an upward trend in cost pattern after a stabilized cost period could also act as a warning sign of deteriorating health of the patient or recurrences of the disease.

REFERENCES

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

Bang, H., & Tsiatis, A. A. (2002). Median regression with censored cost data. *Biometrics*, *58*(3), 643–649.

Başer, O., Gardiner, J. C., Bradley, C. J., & Given, C. W. (2004). Estimation from censored medical cost data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *46*(3), 351–363.

Basu, A., Polsky, D., & Manning, W. G. (2011). Estimating treatment effects on healthcare costs under exogeneity: Is there a 'magic bullet'? *Health Services and Outcomes Research Methodology*, *11*(1-2), 1–26.

Brown, M. L., Riley, G. F., Schussler, N., & Etzioni, R. (2002). Estimating health care costs related to cancer treatment from seer-medicare data. *Medical Care*, IV104–IV117.

Correa, P. (2013). Gastric cancer: Overview. *Gastroenterology Clinics of North America*, *42*(2), 211.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Hernan, M., & Robins, J. (2019). *Causal inference*. Taylor & Francis.

Klabunde, C. N., Potosky, A. L., Legler, J. M., & Warren, J. L. (2000). Development of a comorbidity index using physician claims data. *Journal of Clinical Epidemiology*, *53*(12), 1258–1267.

Li, J., Handorf, E., Bekelman, J., & Mitra, N. (2016). Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in Medicine*, *35*(12), 1985–1999.

Lin, D., Feuer, E., Etzioni, R., & Wax, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics*, 419–434.

Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics*, *20*(4), 461–494.

Mihaylova, B., Briggs, A., O'Hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, *20*(8), 897–916.

NCI. (2014). *Seer-medicare: Selecting the appropriate comorbidity sas macro.* https://healthcaredelivery.cancer.gov/seermedicare/considerations/calculation.html

Paulus, M. T., Claridge, D. E., & Culp, C. (2015). Algorithm for automating the selection of a temperature dependent change point model. *Energy and Buildings*, *87*, 95–104.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical science*, 425–441.

Reeves, J., Chen, J., Wang, X. L., Lund, R., & Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, *46*(6), 900–915.

Robins, J., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Roth, W. E. (1934). On direct product matrices. *Bulletin of the American Mathematical Society*, *40*(6), 461–468.

Tramontano, A. C., Chen, Y., Watson, T. R., Eckel, A., Sheehan, D. F., Peters, M. L. B., Pandharipande, P. V., Hur, C., & Kong, C. Y. (2019). Pancreatic cancer treatment costs, including patient liability, by phase of care and treatment modality, $2000-2013$. *Medicine*, *98*(49).

U.S. Department of Labor Bureau of Labor Statistic. (2021). *Consumer price index data*. https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008

Wan, Y., Datta, S., Lee, J. J., & Kong, M. (2017). Monotonic single-index models to assess drug interactions. *Statistics in Medicine*, *36*(4), 655–670.

Wijeysundera, H. C., Wang, X., Tomlinson, G., Ko, D. T., & Krahn, M. D. (2012). Techniques for estimating health care costs with censored data: An overview for the health services researcher. *ClinicoEconomics and Outcomes Research: CEOR*, *4*, 145.

Wu, C., Zhou, H., Wang, T., Zhang, I., Chen, Y., & Zhao, D. (2019). Impact of the time from the completion of neoadjuvant chemotherapy to surgery on the outcomes of patients with gastric cancer. *Translational Cancer Research*, *8*(5), 1853–1862.

Yan, X., Zheng, Q., & Kong, M. (2021). Estimating medical costs from incomplete follow-up data. *Journal of Statistical Computation and Simulation (To appear)*.

Yeh, J. M., Tramontano, A. C., Hur, C., & Schrag, D. (2017). Comparative effectiveness of adjuvant chemoradiotherapy after gastrectomy among older patients with gastric adenocarcinoma: A seer–medicare study. *Gastric Cancer*, *20*(5), 811–824.

APPENDIX

I-spline calculation:

We define I-splines matrix as (Wan et al. 2017):

$$I_\kappa^k(T) = \begin{cases} 0 & , \kappa > l; \\ \sum\limits_{m=\kappa}^{l} \frac{\tau_{m+K-1}-\tau_m}{k+1} M_m^{k+1}(T) & , l-k+1 \leq \kappa \leq l; \\ 1 & , \kappa < l-k+1; \end{cases}$$ where

for $k > 1,\ M_\kappa^k(T) = \frac{k[(T-\tau_\kappa)M_\kappa^{K-1}(T)+(\tau_{\kappa+k}-T)M_{\kappa+1}^{K-1}(T)]}{(K-1)(\tau_{\kappa+k}-\tau_\kappa)}$,

for $k = 1,\ M_\kappa^1(T) = \begin{cases} \frac{1}{\tau_{\kappa+1}-\tau_\kappa} & , \tau_\kappa \leq t < \tau_{\kappa+1}; \\ 0 & , \text{otherwise}; \end{cases}$,

$\min\limits_{1 \leq i \leq N}(T_i) = \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K = \max\limits_{1 \leq i \leq N}(T_i)$

and $l = 0, \cdots, K$.

Figure A0.1: Flowchart for creating the study cohort (Panel A) and different treatment combinations (Panel B) in case study of Chapter 2

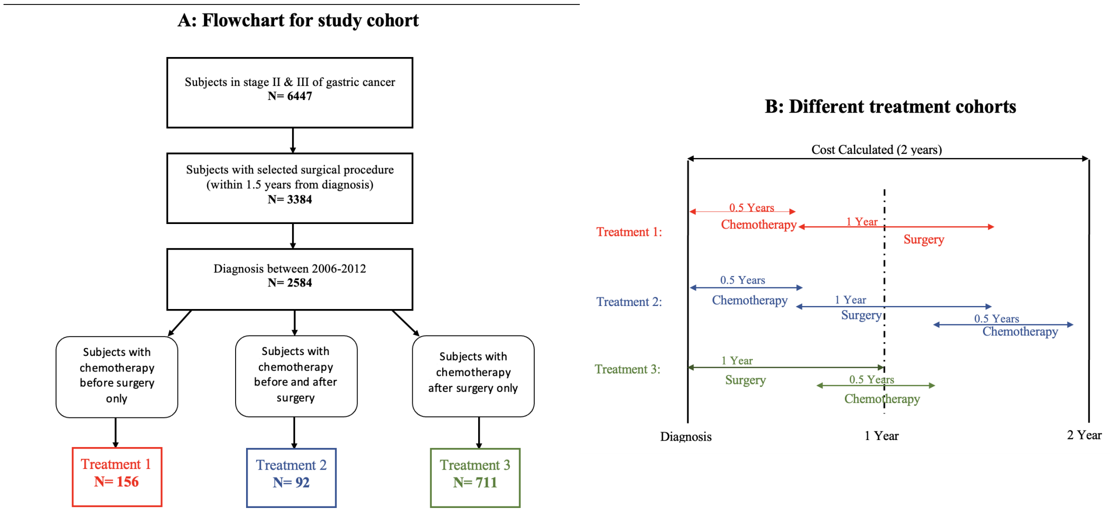
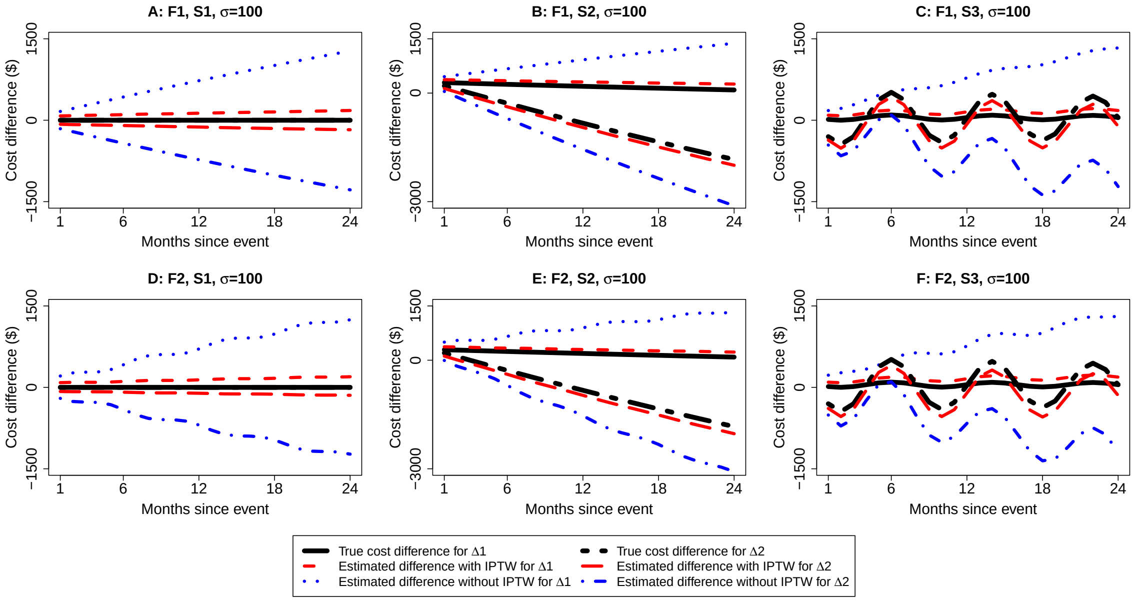

A: Flowchart for study cohort
Subjects in stage II & III of gastric cancer
N= 6447
Subjects with selected surgical procedure
(within 1.5 years from diagnosis)
N= 3384
Diagnosis between 2006-2012
N= 2584
Subjects with chemotherapy before surgery only
Subjects with chemotherapy before and after surgery
Subjects with chemotherapy after surgery only
Treatment 1
N= 156
Treatment 2
N= 92
Treatment 3
N= 711
B: Different treatment cohorts
Cost Calculated (2 years)
Treatment 1:
0.5 Years
Chemotherapy
1 Year
Surgery
Treatment 2:
0.5 Years
Chemotherapy
1 Year
Surgery
0.5 Years
Chemotherapy
Treatment 3:
1 Year
Surgery
0.5 Years
Chemotherapy
Diagnosis
1 Year
2 Year

Figure A0.2: Additional simulation results from Chapter 2 for the average treatment effect in terms of cost difference over time under different scenarios for $\delta = 100$ in Chapter 2



CURRICULUM VITA

NAME: Indranil Ghosh

ADDRESS: Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, KY 40292

DOB: Kolkata, India - May 2, 1991

EDUCATION: B.Sc. Statistics,
University of Calcutta, 2012
M.Sc. Statistics,
Banaras Hindu University, 2014

PRESENTATIONS: JSM Conference (Virtual), Philadelphia, PA, August 2020.
Estimating Treatment Effect on Medical Cost Trajectory
using Propensity Scores and I-Splines.
ASA KY Chapter Spring Conference, Louisville, KY, April 2021.
Estimating Treatment Effect on Medical Cost Trajectory
using Propensity Scores and I-Splines.

AWARDS NSF funded Harshbarger Travel award for the SRCOS
Summer Research Conference, Jekyll Island, GA, October 2021.