Journal of Data Science 23 (3), 560–574
July 2025

DOI: 10.6339/24-JDS1157
Data Science in Action

Estimating Healthcare Expenditure Using Parametric Change Point Models

Indranil Ghosh[1,2], Qi Zheng[1], Michael E Egger[3,4], and Maiying Kong[1,4,*]

[1]*Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, Kentucky, USA*
[2]*Department of Biostatistics, Apellis Pharmaceuticals, Waltham, Massachusetts, USA*
[3]*The Hiram C. Polk, Jr., MD Department of Surgery, School of Medicine, University of Louisville, Louisville, Kentucky, USA*
[4]*Brown Cancer Center, University of Louisville, Louisville, Kentucky, USA*

Abstract

Estimating healthcare expenditures is important for policymakers and clinicians. The expenditure of patients facing a life-threatening illness can often be segmented into four distinct phases: diagnosis, treatment, stable, and terminal phases. The diagnosis phase encompasses healthcare expenses incurred prior to the disease diagnosis, attributed to frequent healthcare visits and diagnostic tests. The second phase, following diagnosis, typically witnesses high expenditure due to various treatments, gradually tapering off over time and stabilizing into a stable phase, and eventually to a terminal phase. In this project, we introduce a pre-disease phase preceding the diagnosis phase, serving as a baseline for healthcare expenditure, and thus propose a five-phase to evaluate the healthcare expenditures. We use a piecewise linear model with three population-level change points and $4p$ subject-level parameters to capture expenditure trajectories and identify transitions between phases, where p is the number of covariates. To estimate the model's coefficients, we apply generalized estimating equations, while a grid-search approach is used to estimate the change-point parameters by minimizing the residual sum of squares. In our analysis of expenditures for stages I–III pancreatic cancer patients using the SEER-Medicare database, we find that the diagnostic phase begins one month before diagnosis, followed by an initial treatment phase lasting three months. The stable phase continues until eight months before death, at which point the terminal phase begins, marked by a renewed increase in expenditures.

Keywords *changepoint models; healthcare expenditures; pancreatic cancer; phase-based expenditure; SEER-Medicare*

1 Introduction

Estimating healthcare expenditures is crucial in the medical field, particularly in understanding the costs associated with specific diseases. These expenditures can vary based on treatments received, patient characteristics, and comorbidities. Moreover, healthcare expenditure may undergo significant changes due to events such as cancer diagnosis, intensive treatment, and death. Given the increasing costs of healthcare delivery, budgetary constraints, and an aging population, it is essential for policy makers and clinicians to understand the trajectory of expenditures for patients diagnosed with a lethal disease such as cancer, which not only has a relatively high

*Corresponding author. Email: maiying.kong@louisville.edu.

Received July 1, 2024; Accepted October 12, 2024

incidence rate, but also involves substantial treatment costs (see, e.g., Mihaylova et al., 2011; Wijeysundera et al., 2012).

Although much of the literature focuses on lifetime expenditures associated with specific diseases (Lin et al., 1997; Bang and Tsiatis, 2002; Basu et al., 2011; Li et al., 2016), understanding the patterns and trajectories of expenditures as a disease progresses provides even deeper insights. Recent literature suggests that there are multiple phases involved in the expenditure of a patient diagnosed with cancer (see, e.g., Brown et al., 2002; Wijeysundera et al., 2012; Tramontano et al., 2019). For example, Wijeysundera et al. (2012) and Tramontano et al. (2019) proposed that healthcare care expenditure related to cancer can be categorized into four distinct phases. Understanding these expenditure phases and identifying the points at which they occur is crucial for policymakers and health insurers (Tramontano et al., 2019). Simply averaging expenditures over time may overlook the change points and patterns in expenditures that correspond to different disease phases. Recognizing these change points and patterns is also vital from a research perspective, as it allows investigators to estimate the times at which treatments or disease status change based on expenditure fluctuations.

Despite the importance of these phases for expenditure analysis, many existing studies lack rigorous methods for estimating the corresponding change points. For example, Wijeysundera et al. (2012) and Tramontano et al. (2019) defined expenditure phases and estimated costs using sample means from different sub-cohorts. However, these approaches do not model the change points between phases or account for the influence of patient characteristics and treatment choices on expenditures in each phase. In this project, we propose a statistical framework for estimating change points along with expenditure trajectories influenced by patients' characteristics and treatment decisions. While change point detection techniques are widely used in economics and meteorology (see, e.g., Reeves et al., 2007; Paulus et al., 2015), their application in medical expenditure analysis is novel.

Another novel aspect of our method is the introduction of an additional phase—the pre-diagnosis phase—to study healthcare expenditures. In evaluating cancer-related costs, Wijeysundera et al. (2012) and Tramontano et al. (2019) categorized cancer-attributable expenditures into four distinct phases: the diagnosis phase, initial treatment phase, stable phase, and terminal phase, each corresponding to different stages of medical care. However, capturing baseline expenditures prior to the diagnosis phase is critical, as it enables the estimation of baseline costs and helps identify the point where expenditures begin to rise. This, in turn, aids in estimating the transition from the pre-disease phase to the diagnosis phase.

Our proposed model with five healthcare expenditure phases is illustrated in Figure 1. The five different expenditure phases are defined as follows: (P0) Pre-disease phase, this phase occurs before diagnosis (t_0), where patients may not be aware of any disease or may not have any health conditions until a certain time point (say, $t_0 - \tau_{-1}$) before diagnosis. The baseline expenditure parameter in this phase is denoted by β_0, representing the patient's expenditure in good health, serving as a baseline for recognizing subsequent changes. (P1) Diagnosis phase, this phase begins from the change point before diagnosis ($t_0 - \tau_{-1}$) and extends to the time of diagnosis of the disease at t_0. During this period, extensive diagnostic tests and procedures are typically performed, resulting in higher medical expenditure. The parameter associated with this phase is denoted by β_1, capturing the change in monthly expenditure during the diagnosis phase. (P2) Initial treatment phase: this phase starts from the diagnosis time t_0 and continues until the end of intensive treatment at time $t_0 + \tau_1$. Expenditure during this phase tends to decrease over time until it reaches a relatively stable level. The expenditure parameter associated with this phase is denoted by β_2, capturing the change in monthly expenditure during the initial treatment phase.

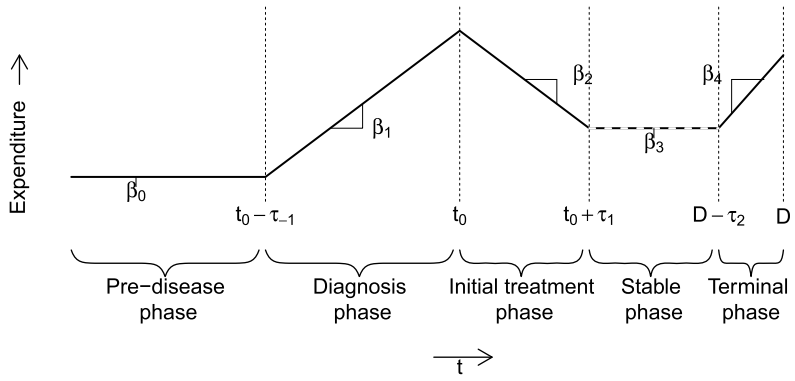

Figure 1: Illustration of expenditure phases and change points.

(P3) Stable phase: this phase lies between the end of the initial treatment phase at time $t_0 + \tau_1$ and the beginning of the terminal phase at time $(D - \tau_2)$, where patients become severely ill again, that is τ_2 months before the end of life at time D. The expenditure parameter associated with this phase is denoted by β_3, representing the average monthly expenditure for the patient during the stable phase. (P4) Terminal phase: this phase spans from becoming severely ill after the stable phase at $(D - \tau_2)$ to the end of life at time D. The expenditure parameter associated with this phase is denoted as β_4, capturing the change in monthly expenditure during the terminal phase.

We use a piecewise linear model with three population-level change points $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)^T$ and $4p$ subject-level parameters to capture expenditure trajectories and identify transitions between phases, where p is the number of covariates. It is widely recognized that medical expenditures are often influenced by the treatment received and patient comorbidities (Austin, 2011). To account for patient-level characteristics and treatment effects, we allow the expenditure parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$ to be patient-specific and model them based on individual patient's characteristics, enabling a more comprehensive understanding of how these characteristics and treatment choices influence healthcare expenditure trajectories across different phases of disease progression and treatment.

The remainder of the paper is structured as follows. In Section 2, we present detailed information on the proposed method. Section 3 is dedicated to applying the proposed method to estimate the expenditure trajectory for pancreatic cancer patients with stages I–III in the SEER-Medicare 2005–2014 database. The final section, Section 4, is reserved for discussion.

2 The Proposed Model for Change Point Detection and Expenditure Trajectory Estimation

2.1 A Simple Change Point Model

Without loss of generality, we set t_0 as time 0, since we can always align and standardize the patients' expenditures from the time of diagnosis. Let $C(t)$ denote the expenditure during month t, and let D denote the time of death since diagnosis. We assume that everyone in the cohort has survived the diagnosis, with $D > 0$. Within the five-phase framework as illustrated in Figure 1 and based on established change points models in the literature (see, e.g., Reeves et al., 2007; Paulus et al., 2015), the expected expenditure trajectory during the pre-disease phase and

diagnosis phase can be captured by a 3-parameter model $E[C(t)] = [\beta_0+\beta_1(t-\tau_{-1})^+]I_{(t<0)}$, where $A^+ = A$ if $A > 0$ and 0 otherwise for a generic quantity A, and $I_{(\cdot)}$ is the indicator function such that $I_{(A)} = 1$ if a generic event A is true and 0 otherwise. The expected expenditure profile from diagnosis to death can be captured by a 5-parameter model: $E[C(t)] = [\beta_3 + \beta_2(\tau_1 - t)^+]I_{(t\geqslant 0)} + \beta_4(t - (D - \tau_2))^+$. Thus, the proposed piece-wise linear model with three change points can be written as:

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

where $\epsilon(t)$ is a random variable with Gaussian process of autoregressive model of order one (AR(1)) and zero means. Since the expenditure function in practice is often continuous, we impose the constraint $\beta_0 + \beta_1(0 - \tau_{-1}) = \beta_3 + \beta_2(\tau_1 - 0)$ to ensure that $E[C(t)]$ is continuous at 0, where the 3-parameter model for the first two phases and the 5-parameter model for the last three phases meet. Thus, by replacing $\beta_3 = \beta_0 - \beta_1\tau_{-1} - \beta_2\tau_1$ in $C(t)$, the expectation of the expenditure function in equation (1) can be written as:

$$\begin{aligned} &E\big[C(t)\big] \\ =&\left[\beta_0 + \beta_1(t - \tau_{-1})^+\right]I_{(t<0)} \\ &+ \left[\beta_0 - \beta_1\tau_{-1} - \beta_2\tau_1 + \beta_2(\tau_1 - t)^+\right]I_{(t\geqslant 0)} + \beta_4\big(t - (D - \tau_2)\big)^+ \\ =&\beta_0 + \beta_1\left[(t - \tau_{-1})^+I_{(t<0)} - \tau_{-1}I_{(t\geqslant 0)}\right] \\ &+ \beta_2\left[\big(-\tau_1 + (\tau_1 - t)^+\big)I_{(t\geqslant 0)}\right] + \beta_4\big(t - (D - \tau_2)\big)^+ = \mathbf{Z}^\top\boldsymbol{\beta}, \end{aligned} \tag{2}$$

where $\mathbf{Z} = (1, (t-\tau_{-1})^+I_{(t<0)} - \tau_{-1}I_{(t\geqslant 0)}, (-\tau_1+(\tau_1-t)^+)I_{(t\geqslant 0)}, (t-(D-\tau_2))^+)^\top$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_4)^\top$, and $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)^\top$. Both $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are unknown and must be estimated. Additionally, β_3 can be estimated from the relationship $\beta_3 = \beta_0 - \beta_1\tau_{-1} - \beta_2\tau_1$. Here β_0 and β_3 represent the expenditure per time unit in the pre-disease phase and stable phase, respectively; β_1 and β_4 capture the rate of expenditure increase per time unit in the diagnosis phase and terminal phase, respectively; and β_2 captures the rate of the expenditure decrease per time unit in the initial treatment phase. τ_{-1} represents the time units prior to diagnosis marking the entry into the diagnosis phase, τ_1 represents the time units post-diagnosis to the end of the initial treatment phase, and τ_2 represents the time units before death indicating the start of the terminal phase.

Note that the third term in equations (1) and (2), $\beta_4(t - (D - \tau_2))^+$, does not incorporate a time indicator variable, indicating that the terminal phase is consistently present and lasts for τ_2 months. In situations where the death time D is shorter than the combined durations of the initial treatment phase and the terminal phase, i.e., $D < \tau_1 + \tau_2$, equation (2) still holds, but there is no stable phase in between. If $D < \tau_2$, the expenditures prior to diagnosis are a mixture of the diagnostic and terminal phases, while expenditures post-diagnosis are a combination of the initial treatment and terminal phases. If $\tau_2 \leqslant D < \tau_1 + \tau_2$, the post-diagnosis expenditures include the initial treatment phase, followed by a combination of the initial treatment phase and the terminal phase.

The expenditure profile often depends on patient characteristics such as age, comorbid conditions, and treatment received (Austin, 2011). We propose a model in which the regression parameters $\boldsymbol{\beta}$ are patient-specific and depend on individual patient variables, while the change point parameters remain population-specific. This approach allows us to understand how individual patient characteristics influence healthcare expenditure trajectories over time, while still accounting for common trends observed across the population.

2.2 The Proposed Patient-Level Change Point Expenditure Models

Let $(\mathbf{X}, \mathcal{T}, C, \delta)$ denote the random variable observed for a patient. $\mathbf{X}$ represents a vector of p time-invariant covariates of the patient, including patients' characteristics, medical history, and treatment information. $\mathcal{T} = (t_{-a}, \ldots, t_0, t_1, \ldots, t_b)^\top$ denotes the vector of time points at which medical expenditures $C = (C_{-a}, \ldots, C_0, C_1, \ldots, C_b)^\top$ are recorded, where we set $t_0 = 0$. δ is an indicator variable that specifies whether the patient died at the last observed time point t_b, with $\delta = 1$ indicating death at time t_b.

The proposed model (2) incorporates each patient's survival time, D. For patients who died during the study period, the observed survival time is used in equation (2). For censored patients, the median of the residual lifetime is estimated based on a working model such as the accelerated failure time (AFT) model. That is, if $\delta = 1$, the survival time D is equal to the last observed time point t_b. If $\delta = 0$, the survival time D is estimated as the sum of the predicted median of the residual lifetime and the censoring time. A terminal expenditure phase, as described in equation (2), occurs for a patient with censored death (i.e., $\delta = 0$) if the last observed time point t_b satisfies the condition $t_b > \hat{D} - \tau_2$, where $\hat{D}$ is the sum of the predicted median of the residual lifetime and the censoring time.

Let $(\mathbf{X}_i, \mathcal{T}_i, C_i, \delta_i)_{i=1}^N$ represent the observed data for N patients in the study. The expected expenditure trajectory for each patient is assumed to follow the pattern outlined in Figure 1. We propose that the parameters $\boldsymbol{\beta}$ in equation (2) are patient-specific, denoted as $\boldsymbol{\beta}_i$ for the ith patient ($i = 1, \ldots, N$), while the change point parameters $\boldsymbol{\tau}$ remain population-specific. The expenditure for the ith patient at time t_{ij} can be written as: $C_{ij} = \mathbf{Z}_{ij}^\top \boldsymbol{\beta}_i + \epsilon_{ij}$, where $C_{ij} = C(t_{ij})$, $\mathbf{Z}_{ij} = (1, (t_{ij} - \tau_{-1})^+ I_{(t_{ij}<0)} - \tau_{-1} I_{(t_{ij}\geqslant 0)}, (-\tau_1 + (\tau_1 - t_{ij})^+) I_{(t_{ij}\geqslant 0)}, (t_{ij} - (D_i - \tau_2))^+)^\top$, and $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \beta_{i2}, \beta_{i4})^\top$. We model $\boldsymbol{\beta}_i$ as a linear function of the patient covariates $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$. That is

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

where $\Gamma \in \mathbb{R}^{4\times p}$ is the parameter matrix to be estimated. Once we have Γ estimated (denoted as $\hat{\Gamma}$), we can gauge the contribution of each covariate on the expenditure profile in each phase. Further, we can also predict the patient-level expenditure profile for the ith patient with covariate $\mathbf{X}_i$ using the relationship $\hat{\boldsymbol{\beta}}_i = \hat{\Gamma} \mathbf{X}_i$. Note that

$$C_{ij} = \mathbf{Z}_{ij}^\top \boldsymbol{\beta}_i + \epsilon_{ij} = \mathbf{Z}_{ij}^\top \Gamma \mathbf{X}_i + \epsilon_{ij}. \tag{3}$$

To estimate Γ, we apply Roth's Columns Lemma (Roth, 1934), which is popularly known as the "vec trick", to equation (3):

$$C_{ij} = \left[\mathbf{X}_i{}^\top \otimes \mathbf{Z}_{ij}^\top\right] \mathbf{vec}(\Gamma) + \epsilon_{ij},$$

where

$$\left[\mathbf{X}_i{}^\top \otimes \mathbf{Z}_{ij}^\top\right] = \left[X_{i1}\mathbf{Z}_{ij}^\top,\ X_{i2}\mathbf{Z}_{ij}^\top,\ \ldots,\ X_{ip}\mathbf{Z}_{ij}^\top\right] \in \mathbb{R}^{4p}$$

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

$$C_i = \left[\mathbf{X}_i{}^\top \otimes \mathbf{Z}_i^\top\right] \mathbf{vec}(\Gamma) + \epsilon_i, \tag{4}$$

where

$$\mathbf{Z}_i^\top = \begin{bmatrix} \mathbf{Z}_{i(-a_i)}^\top \\ \vdots \\ \mathbf{Z}_{i(0)}^\top \\ \vdots \\ \mathbf{Z}_{ib_i}^\top \end{bmatrix}, \quad \mathbf{X}_i{}^\top \otimes \mathbf{Z}_i^\top = \begin{bmatrix} X_{i1}\mathbf{Z}_{i(-a_i)}^\top, X_{i2}\mathbf{Z}_{i(-a_i)}^\top, \ldots, X_{ip}\mathbf{Z}_{i(-a_i)}^\top \\ \vdots \\ X_{i1}\mathbf{Z}_{i(0)}^\top, X_{i2}\mathbf{Z}_{i(0)}^\top, \ldots, X_{ip}\mathbf{Z}_{i(0)}^\top \\ \vdots \\ X_{i1}\mathbf{Z}_{ib_i}^\top, X_{i2}\mathbf{Z}_{ib_i}^\top, \ldots, X_{ip}\mathbf{Z}_{ib_i}^\top \end{bmatrix},$$

$\mathbf{Z}_i^\top \in \mathbb{R}^{(a_i+b_i+1)\times 4}$, $\mathbf{X}_i{}^\top \otimes \mathbf{Z}_i^\top \in \mathbb{R}^{(a_i+b_i+1)\times 4p}$, and each ϵ_i follows a multivariate normal distribution $\text{MVN}(0, \sigma^2 R_i)$ with R_i being an auto-regressive correlation matrix with a common first-order correlation coefficient ρ to capture the possible correlations among observed expenditures over time for the ith patient. To evaluate Γ and the change points, we expand equation (4) for the entire sample as:

$$C = \begin{bmatrix} X_1{}^\top \otimes \mathbf{Z}_1^\top \\ \vdots \\ X_i{}^\top \otimes \mathbf{Z}_i^\top \\ \vdots \\ X_N{}^\top \otimes \mathbf{Z}_N^\top \end{bmatrix} \mathbf{vec}(\Gamma) + \epsilon, \tag{5}$$

where $C = (C_1, \ldots, C_i, \ldots, C_N)^\top \in \mathbb{R}^{\sum_{i=1}^N (a_i+b_i+1)\times 1}$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)^\top$ is the vector of random noises with ϵ_i and $\epsilon_{i'}$ being mutually independent. Given the potentially large number of parameters (i.e., Γ) to be estimated in the model, there is a risk of overfitting. To address this, we apply penalized generalized estimating equations (PGEE) (Wang et al., 2012) to estimate the parameters Γ. PGEE is effective for simultaneous variable selection and estimation, particularly when the number of covariates in the model is large.

2.3 The Estimation Procedure

The key parameters in the proposed change point model are the parameter matrix

$$\Gamma = \begin{bmatrix} \gamma_{01} & \gamma_{02} & \cdots & \gamma_{0p} \\ \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \gamma_{41} & \gamma_{42} & \cdots & \gamma_{4p} \end{bmatrix}$$

and the change points $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)$.

Subsequently, we can obtain the patient-level expenditure profile using the relationship $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \beta_{i2}, \beta_{i4})^\top = \Gamma \mathbf{X}_i$ $(i = 1, \ldots, N)$ and $\beta_{i3} = \beta_{i0} - \beta_{i1}\tau_{-1} - \beta_{i2}\tau_1$. To estimate Γ and $\boldsymbol{\tau}$, we first need to estimate the time of death for censored patients using a survival model. We propose using the AFT model to predict the median of the residual lifetime, after which the survival time is estimated as the sum of the predicted residual lifetime and the censoring time for patients with $\delta_i = 0$. The median of the residual lifetime for patients with $\delta_i = 0$ is estimated using the AFT model and the covariates $\mathbf{X}_i$ $(i = 1, \ldots, N)$. Next, we specify how to estimate the change points, $\boldsymbol{\tau}$, which are population-specific parameters in our proposed model. According to existing literature for cancer patients (see, e.g. Tramontano et al., 2019), τ_{-1}, τ_1, and τ_2 are often considered as 2 months prior to 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

$(1, 2, 3, 4, 5, 6)$ and τ_2 is within $(5, 6, 7, 8, 9, 10)$. 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 estimate Γ by using the PGEE method (Wang et al., 2012; Inan and Wang, 2017). We then calculate the residual mean square errors (RMSE) for model (5) based on the estimated $\hat{\Gamma}$ for a given $\boldsymbol{\tau}$. The final estimated $\hat{\boldsymbol{\tau}}$ is the one that minimizes the RMSE. The $\hat{\Gamma}$ is the estimate corresponding to the selected $\hat{\boldsymbol{\tau}}$.

To make inferences on the change points' parameters $\boldsymbol{\tau} = (\tau_{-1}, \tau_1, \tau_2)^T$, we use the non-parametric bootstrap resampling scheme to evaluate their accuracy and distribution. 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 obtain the distribution of $\hat{\boldsymbol{\tau}}$, which provides insight into the accuracy of the optimal selection for $\boldsymbol{\tau}$. Additionally, we provide the estimated $\hat{\Gamma}$ matrix along with its standard errors to evaluate the impact of different covariates on healthcare expenditures in different phases. We further estimate the expenditure profile for each patient through the parameter $\boldsymbol{\beta}_i$. These expenditure profiles can be summarized at the population level or within subgroups of interest by calculating the average of the estimated subject-level parameters in each subgroup.

3 Case Study

We applied our proposed method to SEER-Medicare 2005–2014 pancreatic cancer data to investigate healthcare expenditure patterns over time and their association with various covariates. Our primary objective was to estimate population-specific change points and the parameter matrix Γ, along with patient-level expenditure trajectories. This case study delved into the expenditure patterns related to patients with stage I–III pancreatic cancers throughout their diagnoses and treatments across different phases.

This study utilized the 2005–2014 SEER cancer file along with Parts A and B claims files. The study cohort comprised patients diagnosed with pancreatic cancer at the primary site, with specified histology and behavioral codes in the SEER cancer file between March 2006 and December 2013. Eligibility required patients to receive at least one treatment within six months of diagnosis. All cohort members were continuously enrolled in Medicare Parts A and B, without HMO coverage, from 14 months prior to diagnosis until December 2014 or death, whichever occurred first. In cases of multiple diagnoses, the initial occurrence was considered. Comorbidities were assessed using the NCI comorbidity index, based on data from the year preceding the pancreatic cancer diagnosis. This index, developed by Klabunde et al. (2000), is tailored specifically for cancer and excludes solid tumors, leukemia, and lymphomas as comorbid conditions. The NCI comorbidity index was calculated using the 2014 NCI SAS Macro (NCI, 2014), leveraging data from the SEER cancer file, inpatient file (Medpar), outpatient file (Outpat) and carrier claims file (NCH). Covariates obtained from the SEER cancer file included demographic variables (race, age, and sex), geographical variables (Metro vs. non-Metro), and cancer-specific variables (grade 1–4; stages I & II vs III). Treatment assignments were categorized as chemotherapy or surgery, based on the first intervention administered post-diagnosis. Our study specifically targeted patients aged 65 and older, diagnosed with stage I–III pancreatic cancer, who received either surgery or chemotherapy. After applying these inclusion and exclusion criteria, our sample consisted of 2,899 patients. Of these, 2,277 patients passed away during the study period, while 622 survived until the end of the study in December 2014.

All expenditures from Medpar, Outpat, and NCH files were recorded as observed expenditures, reflecting payments made by Medicare. These expenditures were adjusted to 2014 rates

using the Consumer Price Index (CPI) data (US Department of Labor Bureau of Labor Statistic, 2021). For each patient, the time of diagnosis was set as the origin ($t_0 = 0$), and monthly expenditures were calculated by summing all expenditures incurred during that month. To adequately capture the pre-disease phase expenditure, we limited our observation period to 14 months prior to diagnosis. This ensured sufficient data to delineate the expenditure trajectory during this phase. For a comprehensive analysis of expenditures, we followed patients over time until death or the end of the study period in December 2014, whichever occurred first. The diagnostic window was set from March 2006 to December 2013 to ensure (1) at least 14 months to investigate the expenditure patterns prior to diagnosis, encompassing the pre-disease and diagnostic phases, and (2) a minimum of one year of follow-up to assess post-diagnosis Medicare expenditure patterns for the cohort.

As healthcare expenditure data are often highly skewed, we transformed the expenditures 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 parameter: τ_{-1}, τ_1 and τ_2 range from -6 to -1, 1 to 6, and 5 to 10, respectively. In this case study, the optimal choice of $\boldsymbol{\tau}$, $\hat{\boldsymbol{\tau}}$, which minimized the RMSE of our proposed model, was obtained as $(-1, 3, 8)$. Figure 2 Panel A1 illustrates the contour plot of RMSE for different choices of τ_{-1} and τ_1, with τ_2 fixed at the optimal value 8. Figure 2 Panel A2 presents the contour plot of RMSE for different choices of τ_1 and τ_2, with the optimal choice of τ_{-1} set at -1. From Figure 2, it is evident that our optimal selection of $\hat{\boldsymbol{\tau}} = (-1, 3, 8)$ minimized the RMSE of the model among all 216 possible values of $\boldsymbol{\tau}$. We further estimated the distribution of $\hat{\boldsymbol{\tau}}$ by performing 1000 bootstrap samplings. The relative frequencies of the selected change points $\boldsymbol{\tau}$ are shown in Table 1. It is clear that each component of the selected $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{-1}, \hat{\tau}_1, \hat{\tau}_2) = (-1, 3, 8)$ was the mode of its distribution.

With the change point parameters fixed at $\hat{\tau}$ and the estimated time of death, the design matrix X_i in the model (5) was determined. We then estimated the regression parameter matrix $\hat{\Gamma}$ using the PGEE method. The estimated parameter matrix $\hat{\Gamma}$ and their standard errors are shown in Table 2. From the $\hat{\Gamma}$ matrix, we calculated the expenditure profile parameters $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_4)^T = \hat{\Gamma} X$, and computed $\hat{\beta}_3 = \hat{\beta}_0 - \hat{\beta}_1 \hat{\tau}_{-1} - \hat{\beta}_2 \hat{\tau}_1$, which captures the expenditure during the stable phase. The estimated $\hat{\Gamma}$, along with their standard errors (SE) and p-values, are presented in Table 2.

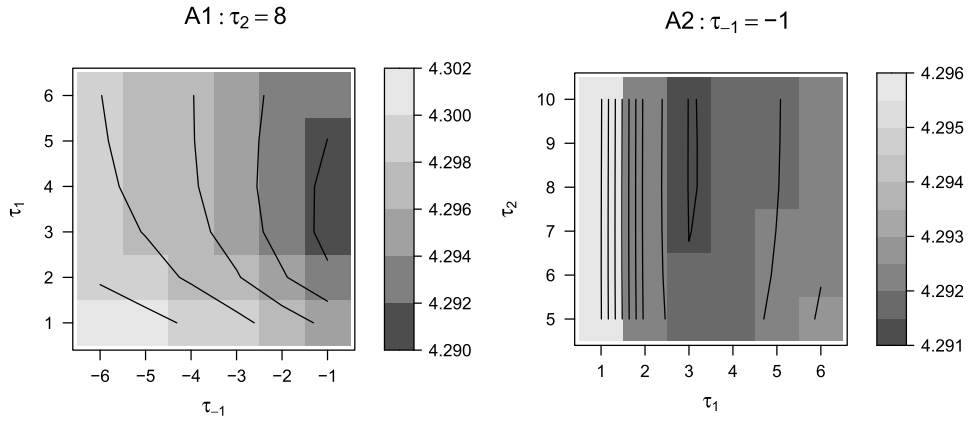

Figure 2: The contour plot of RMSE for $\hat{\tau}_1$ versus $\hat{\tau}_{-1}$ with $\hat{\tau}_2$ fixed at the optimal value of 8 (Panel A1), and the contour plot of RMSE for $\hat{\tau}_2$ versus $\hat{\tau}_1$ with $\hat{\tau}_{-1}$ fixed at the optimal value of -1.

Table 1: Distribution of $\hat{\boldsymbol{\tau}}$ based on 1000 bootstrap samples.

$\hat{\boldsymbol{\tau}}_{-1}$						
Choices	−6	−5	−4	−3	−2	−1
Occurrence %	0%	0%	0%	0%	12%	88%
$\hat{\boldsymbol{\tau}}_{1}$						
Choices	1	2	3	4	5	6
Occurrence %	0%	2%	82%	16%	0%	0%
$\hat{\boldsymbol{\tau}}_{2}$						
Choices	5	6	7	8	9	10
Occurrence %	0%	0%	7%	42%	32%	19%

Table 2: Estimated values (Est) and their standard errors (SE) of the Γ matrix along with their level of significance (p-value).

$\hat{\beta}$		$\hat{\Gamma}$, SE and p-value of Γ for different covariates								
		Intercept	NCI	Metro/Non-Metro	Grade	Age	Race	Sex	Stage	Treatment
	Ref:			Metro			White	Male	I & II	Chemotherapy
$\hat{\beta}_0$	Est	385.7	262.1	38.8	3.0	47.4	−201.2	113.1	−52.7	−152.2
	SE	156.8	25.2	46.4	8.5	36.5	50.2	36.6	44.1	78.7
	(p-value)	(0.014)	(< 0.001)	(0.403)	(0.724)	(0.194)	(< 0.001)	(0.002)	(0.231)	(0.053)
$\hat{\beta}_1$	Est	6403.8	2283.3	1597.6	−334.4	1058.5	2304.1	−1913.0	55.9	9720.8
	SE	3614.1	345.1	1197.3	147.7	813.9	2448.8	975.6	756.0	1013.3
	(p-value)	(0.076)	(< 0.001)	(0.182)	(0.024)	(0.193)	(0.347)	(0.005)	(0.941)	(< 0.001)
$\hat{\beta}_2$	Est	1426.9	753.7	551.8	−152.9	476.3	790.0	−571.0	−320.4	3740.8
	SE	1240.2	118.9	407.9	50.8	279.2	847.8	336.5	258.4	352.9
	(p-value)	(0.25)	(< 0.001)	(0.176)	(< 0.003)	(0.088)	(0.351)	(0.090)	(0.215)	(< 0.001)
$\hat{\beta}_3$	Est	2508.8	284.3	−19.0	127.3	−323.0	−267.1	−86.9	964.4	−1653.8
	SE	14932.7	1440.4	4914.8	613.4	3363.2	10129.2	4040.7	3125.7	4268.1
	(p-value)	(0.867)	(0.843)	(0.997)	(0.836)	(0.924)	(0.979)	(0.983)	(0.758)	(0.693)
$\hat{\beta}_4$	Est	920.8	5.1	−208.4	20.6	−152.1	154.8	−93.3	−90.8	247.9
	SE	269.1	22.1	72.2	13.0	57.7	121.4	63.6	66.3	91.2
	(p-value)	(0.001)	(0.816)	(0.004)	(0.113)	(0.008)	(0.202)	(0.142)	(0.171)	(0.007)

Based on the estimated regression coefficients $\hat{\Gamma}$ for β_0 (i.e., the first row block) in Table 2, it is evident that for each unit increase in the NCI comorbidity index, the baseline monthly expenditure rises by \$262.1. The second row of $\hat{\Gamma}$ for β_1 indicates both the NCI comorbidity index and surgery (compared to Chemotherapy) significantly increase diagnosis expenditure. Factors such as the NCI index, tumour grade, sex, stage of cancer, and the type of treatment provided significantly impact expenditure profiles. For example, females incur higher expenditure during pre-disease phase but lower expenditure in the diagnosis phase and initial treatment phase. Surgery emerges as a major driver of expenditures in the diagnosis and treatment phases. Early-stage cancers are more likely to be treated surgically, resulting in higher initial expenditures. In contrast, patients with stage 3 or 4 pancreatic cancer are typically deemed unresectable and are treated with chemotherapy alone in the initial treatment phase, thereby incurring no surgical expenses. Thus, we can use the $\hat{\Gamma}$ values along with their SEs in Table 2 to infer the effects of the covariates on expenditure profiles across different phases.

Table 3: Observed and estimated monthly expenditures during the five different phases.

Expenditure phases	Average observed expenditure	Average estimated expenditure
	Mean (SD)	Mean (SD)
Pre-disease phase	450 (993)	523 (431)
Diagnosis phase (1 month)	20702 (36803)	20089 (6273)
Initial treatment phase (3 months)	9835 (13719)	8839 (3067)
Stable phase (varied months)	3312 (5225)	2446 (1061)
Terminal phase (8 months)	5573 (7228)	5577 (2502)

We further summarized the observed and estimated monthly expenditures in terms of mean expenditure and standard deviation (SD) for the study cohort during different expenditure phases in Table 3. It is evident that the monthly expenditure during the diagnosis phase was the highest, followed by the initial treatment phase, and then the terminal phase. The expenditure during the stable phase was higher than in the pre-disease phase. From Table 3, we also conclude that the estimated expenditures closely align with the observed expenditures in each phase.

Finally, Figure 3 provides the average observed monthly expenditure (solid line) and average predicted monthly expenditure (dashed lines) for different selected sub-cohorts, along with the estimated change points at $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{-1}, \hat{\tau}_1, \hat{\tau}_2) = (-1, 3, 8)$ (dotted lines). Figure 3 Panel A1 shows the expenditure and its estimation for the entire cohort, while Panel A2 focuses on patients who survived during the study period. Panels B1–B3 present similar results but for different sub-cohorts based on their survival time since diagnosis. Panel B1 includes patients who died

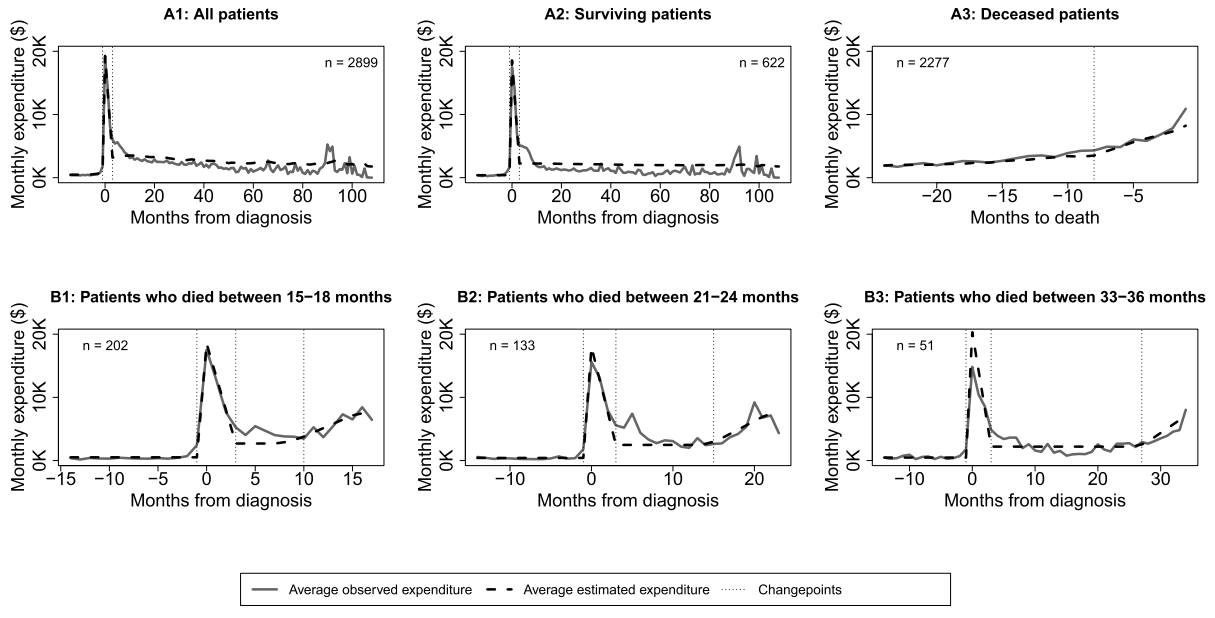

Figure 3: Average observed and estimated monthly expenditure along with change points for different sub-cohorts of pancreatic cancer patients using the proposed parametric change point approach.

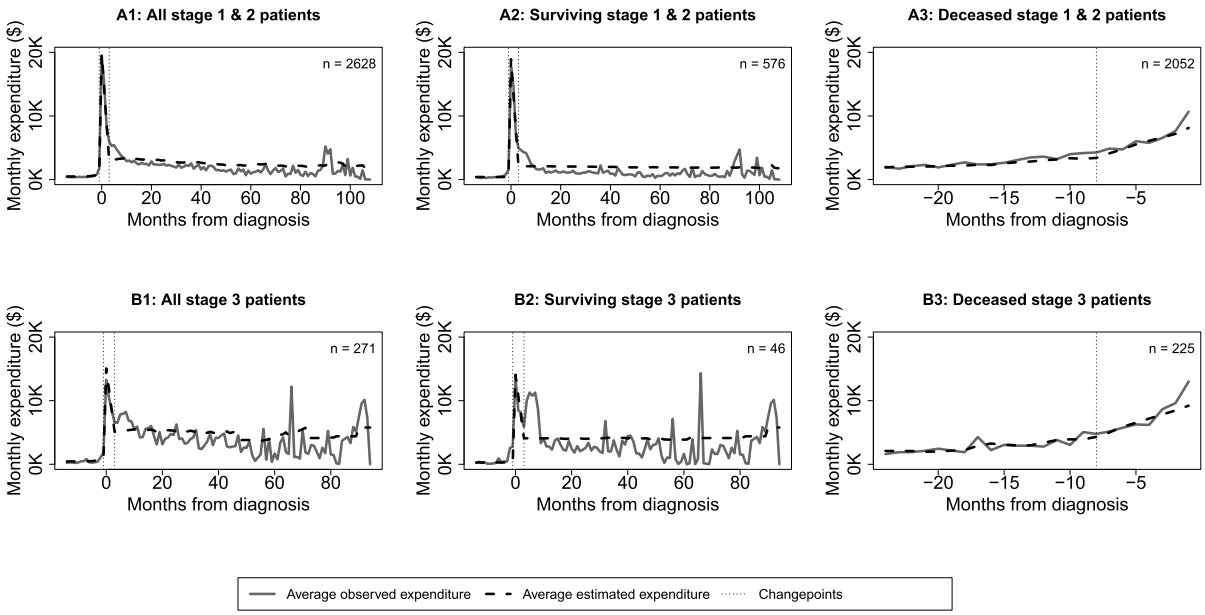

Figure 4: Average observed and estimated monthly expenditure along with change points for different stages of pancreatic cancer patients using the proposed parametric change point approach.

between 15–18 months after diagnosis, Panel B2 includes those who died between 21–24 months after diagnosis, and Panel B3 features patients who died between 33–36 months after diagnosis. Note that all plots are aligned to the time of diagnosis as the origin (at time 0) on the x-axis. Therefore, we cannot display τ_2 in Panels A1 and A2, as the time varies across patients. Panels B1–B3 provide a similar representation but with a narrower range of survival times, with the x-axis extending to the longest survival time in each subcohort. However, these representations again do not accurately depict τ_2, which captures the beginning of the terminal phase prior to death. Since the time of death varied for each patient, the best representation of τ_2 would align the x-axis with respect to the time of death. Panel A3 of Figure 3 summarizes the expenditure aligned with death, illustrating the role of τ_2. Here, the x-axis is set to the time of death as the origin and represents months prior to the time of death. The optimal value τ_2 is plotted at -8 months, clearly indicating an upward trend in expenditure starting 8 months before death, thus confirming that the estimated τ_2 effectively captured the change point.

Existing literature suggests that cancer stages are major contributors to healthcare expenditures (see, e.g., Wijeysundera et al., 2012; Tramontano et al., 2019). To further investigate this, we incorporated cancer stage as a categorical variable in our model and stratified the 2,899 pancreatic cancer patients into two groups: the first group consisted of patients with stage I and II pancreatic cancer, and the second group included those with stage III cancer. We then predicted the expenditures for both groups using our model. The expenditure profiles for the stratified groups are shown in Figure 4, Panels A1 and B1. Each group was further divided into cohorts of surviving and deceased patients, as depicted in Panels A2, A3, B2, and B3. The results demonstrate that cancer stage significantly affects expenditure. Specifically, expenditures for stage I and II patients stabilized after the initial treatment phase, whereas stage III patients showed greater variability in costs. This difference is likely attributable to the fact that stage I

and II patients often undergo surgical treatment at diagnosis, whereas stage III patients are more commonly treated with chemotherapy, leading to higher medical costs and potentially earlier deaths. Our model effectively captured the first and second change points, as shown in Figure 4 Panels A1, A2, B1 and B2, as well as the final change point before death (see Panel A3 and B3).

The results presented in Table 1 and Figures 3 and 4 demonstrate that our estimation of the change point parameters $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{-1}, \hat{\tau}_1, \hat{\tau}_2) = (-1, 3, 8)$ aligns with the change patterns observed in expenditure profiles. Our method effectively captured the upward and downward trends in these expenditure profiles, with the peaks occurring at diagnosis and stabilizing occurring around three months post-diagnosis. Additionally, the model accurately reflected the increase in expenditures prior to death. Thus, the proposed model has the potential to enhance understanding of expenditure patterns, facilitate planning for costly expenditures, and raise awareness of significant events such as disease diagnosis or impending mortality.

4 Discussion

In this project, we introduce a pre-disease phase preceding the diagnosis phase, serving as a baseline for healthcare expenditure, and propose a five-phase piecewise linear model to capture expenditure trajectories and identify transition points between phases. The model uses three population-level parameters to model these transition points, while patient-level characteristics such as demographics, comorbidities, and treatments are incorporated to estimate expenditure amounts and phase durations. Given the change points, we employ penalized generalized estimating equations to estimate the regression coefficients in the proposed model. A grid-search approach is used to estimate the change point parameters by minimizing the residual sum of squares. The innovative aspect of our approach lies in modeling expenditure trajectories using change point detection models, while enhancing accuracy by making the regression parameters dependent on patient characteristics.

We applied this method to estimate expenditure trajectories for stages I–III pancreatic cancer patients using the SEER-Medicare 2005–2014 database, we found that the diagnosis phase initiates one month prior to diagnosis, followed by a three-month initial treatment phase. The stable phase persists until eight months before death, marking the onset of the terminal phase, characterized by increased expenditure once again. The estimation of the change points facilitated precise inference about expenditure patterns for patients with specific diseases. It enhanced our understanding of expenditure pattern dynamics, aiding in more informed decisions regarding treatment funds allocation over time. Moreover, an upward trend in expenditure following a stabilized expenditure period could serve as an early warning sign of deteriorating patient health or disease recurrence.

However, it is important to acknowledge the limitations of the SEER registry data. SEER-Medicare data only captures claims billed to fee-for-service (FFS) Medicare, so individuals enrolled in Part C (Medicare HMO) or without Part B enrollment are likely receiving healthcare that is not recorded in the dataset. In our study, we restricted the analysis to individuals enrolled in both Parts A and B of Medicare, without HMO enrollment during the study period. Otherwise, Medicare expenditures could be misattributed to other insurance and misclassified in the study. For instance, an individual enrolled only in Part C throughout the entire study period would have minimal fee-for-service (FFS) claims, potentially leading to a false perception of zero Medicare expenditure on their care. Such cases must be excluded from the analysis to avoid bias. Therefore, if the analysis were not limited to those with continuous enrollment in Parts A and

B but not C during the study, the results would be significantly biased. Since SEER-Medicare users often restrict their analysis to individuals considered "likely to have complete claims" (Enewold et al., 2020), such limitations may affect the generalizability of the findings. More information on the limitations of using SEER registry data can be found at "https://seer.cancer.gov/data-software/documentation/seerstat/nov2022/treatment-limitations-nov2022.html".

Another limitation of this study is the assumption that the change points are population-specific. This assumption may not hold in heterogeneous populations. For example, patients may receive different treatments, and these treatments could affect the timing of transitions between phases. Additionally, patient characteristics such as age and comorbidities can influence disease progression and, consequently, the change points in expenditure patterns. To address this limitation, a subgroup analysis could be conducted by dividing the data into groups based on treatment types or specific patient characteristics. Applying the proposed model to each subgroup would provide a more tailored understanding of how different factors influence healthcare expenditure trajectories. This approach would offer more nuanced insights into the dynamics of expenditure over time and provide a more comprehensive understanding of how expenditures evolve across different patient subsets, potentially leading to more personalized healthcare strategies.

Despite these limitations, the current article still provides valuable insights into the change points of expenditure trajectories over time.

Supplementary Material

R Codes for Key Steps of the Case Study

Acknowledgement

We greatly appreciate the insightful and constructive comments from the Medicare reviewers, as well as the reviewer and editor from this journal, whose feedback has significantly improved our paper.

The collection of cancer incidence data used in this study was supported by several entities. The California Department of Public Health provided support pursuant to California Health and Safety Code Section 103885; the Centers for Disease Control and Prevention's National Program of Cancer Registries, under cooperative agreement 1NU58DP007156; and the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contracts HHSN261201800032I to the University of California, San Francisco, HHSN261201800015I to the University of Southern California, and HHSN261201800009I to the Public Health Institute. The views expressed here are solely those of the authors and do not necessarily reflect the views of the State of California, the Department of Public Health, the National Cancer Institute, or the Centers for Disease Control and Prevention or their contractors and subcontractors.

Funding

M.E. Egger and M. Kong thank the American Cancer Society for their generous support of this study (CSDG-22-125-01-HOPS). M. Kong also acknowledges the support from the Wendell Cherry Chair in Clinical Trial Research endowment funds at the University of Louisville, along with funding from the National Institute of Health (P30ES030283, R01HL158779, and P20GM155899). Q. Zheng appreciates the support from the National Institute of Health (R21AG070659) and the National Science Foundation (DMS-1952486).

References

Austin PC (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3): 399–424. https://doi.org/10.1080/00273171.2011.568786

Bang H, Tsiatis AA (2002). Median regression with censored cost data. *Biometrics*, 58(3): 643–649. https://doi.org/10.1111/j.0006-341X.2002.00643.x

Başer O, Gardiner JC, Bradley CJ, Given CW (2004). Estimation from censored medical cost data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 46(3): 351–363. https://doi.org/10.1002/bimj.200210036

Basu A, Polsky D, Manning WG (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. https://doi.org/10.1007/s10742-011-0072-8

Brown ML, Riley GF, Schussler N, Etzioni R (2002). Estimating health care costs related to cancer treatment from SEER-Medicare data. *Medical Care*, 40(8): IV104–IV117.

Enewold L, Parsons H, Zhao L, Bott D, Rivera DR, Barrett MJ, et al. (2020). Updated overview of the SEER-Medicare data: enhanced content and applications. *JNCI Monographs*, 2020(55): 3–13.

Inan G, Wang L (2017). PGEE: an R package for analysis of longitudinal data with high-dimensional covariates. *R Journal*, 9(1): 393. https://doi.org/10.32614/RJ-2017-030

Klabunde CN, Potosky AL, Legler JM, Warren JL (2000). Development of a comorbidity index using physician claims data. *Journal of Clinical Epidemiology*, 53(12): 1258–1267. https://doi.org/10.1016/S0895-4356(00)00256-0

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. https://doi.org/10.1002/sim.6842

Lin D, Feuer E, Etzioni R, Wax Y (1997). Estimating medical costs from incomplete follow-up data. *Biometrics*, 53(2): 419–434. https://doi.org/10.2307/2533947

Manning WG, Mullahy J (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4): 461–494. https://doi.org/10.1016/S0167-6296(01)00086-8

Mihaylova B, Briggs A, O'Hagan A, Thompson SG (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8): 897–916. https://doi.org/10.1002/hec.1653

NCI (2014). SEER-medicare: Selecting the appropriate comorbidity SAS macro.

Paulus MT, Claridge DE, Culp C (2015). Algorithm for automating the selection of a temperature dependent change point model. *Energy and Buildings*, 87: 95–104. https://doi.org/10.1016/j.enbuild.2014.11.033

Reeves J, Chen J, Wang XL, Lund R, Lu QQ (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6): 900–915. https://doi.org/10.1175/JAM2493.1

Roth WE (1934). On direct product matrices. *Bulletin of the American Mathematical Society*, 40(6): 461–468. https://doi.org/10.1090/S0002-9904-1934-05899-3

Tramontano AC, Chen Y, Watson TR, Eckel A, Sheehan DF, Peters MLB, et al. (2019). Pancreatic cancer treatment costs, including patient liability, by phase of care and treatment modality, 2000–2013. *Medicine*, 98(49): e18082.

US Department of Labor Bureau of Labor Statistic (2021). Consumer price index data.

Wang L, Zhou J, Qu A (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2): 353–360. https://doi.org/10.1111/j.1541-0420.2011.01678.x

Wijeysundera HC, Wang X, Tomlinson G, Ko DT, Krahn MD (2012). Techniques for estimating health care costs with censored data: an overview for the health services researcher. *ClinicoEconomics and Outcomes Research: CEOR*, 4: 145. https://doi.org/10.2147/CEOR.S31552