Published in final edited form as:

J Am Stat Assoc. 2018 ; 113(522): 582–592. doi:10.1080/01621459.2017.1361329.

Semiparametric Estimation of Longitudinal Medical Cost Trajectory

Liang Li,

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St., Houston, TX, 77030

Chih-Hsien Wu,

Department of Biostatistics, The University of Texas MD Anderson Cancer Center

Jing Ning,

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St., Houston, TX, 77030

Xuelin Huang,

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St., Houston, TX, 77030

Ya-Chen Tina Shih, and

Department of Health Services Research, The University of Texas MD Anderson Cancer Center

Yu Shen

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler St., Houston, TX, 77030

Abstract

Estimating the average monthly medical costs from disease diagnosis to a terminal event such as death for an incident cohort of patients is a topic of immense interest to researchers in health policy and health economics because patterns of average monthly costs over time reveal how medical costs vary across phases of care. The statistical challenges to estimating monthly medical costs longitudinally are multifold; the longitudinal cost trajectory (formed by plotting the average monthly costs from diagnosis to the terminal event) is likely to be nonlinear, with its shape depending on the time of the terminal event, which can be subject to right censoring. The goal of this paper is to tackle this statistically challenging topic by estimating the conditional mean cost at any month t given the time of the terminal event s . The longitudinal cost trajectories with different terminal event times form a bivariate surface of t and s , under the constraint $t \leq s$. We propose to estimate this surface using bivariate penalized splines in an Expectation-Maximization algorithm that treats the censored terminal event times as missing data. We evaluate the proposed model and estimation method in simulations and apply the method to the medical cost data of an incident cohort of stage IV breast cancer patients from the Surveillance, Epidemiology and End Results–Medicare Linked Database.

Keywords

Bivariate Smoothing; Joint Modeling; Lifetime and Survival Analysis; Medical Cost; SEER Medicare

1. INTRODUCTION

Electronic submission of medical billing records enables the automated collection of detailed medical cost data from a large population enrolled in health insurance plans. These data often come in a longitudinal format in the sense that whenever an encounter occurs, the system documents the medical expenditures associated with the encounter with a time stamp. When the medical cost data are structured by following an incident cohort of patients with a specific disease or condition, such data are referred to as the incident cost data (Yabroff et al 2009). A population study of longitudinal incident cost data is of tremendous interest to health policy makers and health economists as patterns of medical costs over time inform healthcare resource use throughout various phases of an illness. For example, the National Cancer Institute (NCI) periodically publishes estimates and projections of the cost of cancer care in the United States based on the analysis of claims data (Mariotto et al 2011; Yabroff et al 2008).

Unlike the total or cumulative medical cost, which summarizes all medical costs of a patient throughout a follow-up period of interest, the longitudinal incident cost provides details on healthcare consumption across different phases of the disease continuum (Yabroff et al 2008). In the motivating application of this paper, for example, the data include the medical costs of an incident cohort of Medicare patients with stage IV breast cancer, from their initial diagnosis until death. Although charges to Medicare as well as Medicare payments may be posted on a daily basis whenever a healthcare encounter occurs, we standardized the unit of analysis to monthly intervals for cost so as to make the data less volatile. The conventional phase-of-care approach of the NCIs cost reporting defines the first 12 months after cancer diagnosis as the initial care phase, the last 12 months before death as the terminal care phase, and the period in between the continuing care phase. During the initial care phase, cancer patients often receive active treatment, including surgery, chemotherapy, and/or radiation therapy to eradicate or inhibit tumor growth. During the terminal care phase, patients may receive intensive salvage treatments and/or palliative care. During the continuing care phase, clinical visits become less intensive as patients have completed active treatment. In general, higher costs are expected during the initial and terminal care phases; whereas lower costs are anticipated for the continuing care phase, which primarily involves surveillance of the patient. The transition time between care phases as defined in the NCIs phase-of-care costing approach appears somewhat arbitrary as the choice of a 12-month duration for each of the initial and terminal care phases makes it difficult to capture the heterogeneity of the cost trajectory among patients with various demographic characteristics and disease progression. It is thus desirable to estimate the population averaged monthly medical costs as a *longitudinal cost trajectory* from diagnosis to death, rather than the average total or cumulative costs for each phase of care as defined by pre-specified cutoff times.

There are a number of statistical challenges in estimating the longitudinal medical cost trajectory. First, while it is of interest for policy makers to study the average costs of a patient population rather than the costs of individuals, simply averaging across all the patients in the study sample may not generate the information required for policy decision making. For example, in the breast cancer patient cohort, not all patients contribute cost data to all three care phases. Aggressive or late-stage tumors may result in patient death shortly after the diagnosis of breast cancer, and the cost trajectories for such patients may consist of the initial and terminal care phases with an absence of the continuing care phase. Conversely, for patients with tumors that progress slowly, the continuing care phase could last many years. Therefore, it would not be appropriate to average, for example, the medical costs 18 months after diagnosis over patients who died at month 22 and patients who died at month 60, because the costs at month 18 would be part of the terminal care costs for the former patients but part of the continuing care costs for the latter patients. A more appropriate aggregation strategy is to calculate the mean cost at any month t conditional on the time of death s . In other words, the mean cost at month t should be expressed as a function of both t and s , forming a mean cost surface of t and s . By definition, this surface can only be defined within a triangular region satisfying $0 < t \leq s$. Second, conditional on the time of death, the longitudinal cost trajectory as a function of t may be nonlinear. Brown et al (2002) demonstrated through descriptive analyses of non-censored patients that the average cost trajectory of a breast cancer cohort is U-shaped with the mouth and bottom of the “U” depending on survival times. Hence, a model for the cost trajectory needs to accommodate nonlinear shapes and dependence between the cost trajectory and survival time. Lastly, the time of death may not be observed for all patients due to the limited duration of the follow-up as a result of administrative censoring commonly observed in claims data. Proper handling of right censoring is necessary in order to estimate the cost trajectories.

As discussed previously, the NCIs current reporting of the cost of cancer care summarizes such costs by pre-defined care phases (Mariotto et al 2011; Yabroff et al 2008; Yabroff et al 2008). However, their analytical methods do not address censoring explicitly and, as a result, cost data from censored phases are not used in the analyses. The existing health sciences literature on cost trajectory has used descriptive statistics to summarize cost data by survival time (Brown et al 2002; Etzioni et al 1996; Yabroff et al 2008; Yabroff et al 2009) or has used parametric functions to model the cost trajectory without allowing the trajectory to vary with the survival time (Akushevich et al 2011). These studies also only used cost data from uncensored subjects. Statistical literature on medical cost research has often focused on estimating the cumulated cost or studying the association between the cumulated cost and baseline covariates (Bang 2005; Bang and Tsiatis 2000; Baser et al 2006; Lin 2000, 2003; Lin et al 1997; O’Hagan and Stevens 2004; Pullenayegum and Willan 2011; Zhao, Cheng and Bang 2011; Zhao and Tian 2001; Zhao et al 2012), adjusting for “induced dependent censoring” in the total cost (Bang 2005; Lin et al 1997). Another related line of research is regression analyses with longitudinal costs modeled as repeated measures outcomes. A comprehensive overview can be found in a special issue of *Medical Care* (vol. 47, no. 7 supplement 1, July 2009). More recent work can be found in Chen et al (2016), Chen et al (2013), Liu et al (2008) and the references therein. These models focus on modeling the

association between covariates and longitudinal medical cost outcomes; the cost trajectories are usually modeled with simple parametric functions (e.g., linear), without allowing them to depend on the possibly censored survival times. Shared parameter models have been used to jointly model longitudinal medical costs and survival times (Liu 2009; Liu, Huang, and O'Quigley 2008; Liu, Wolfe, and Kalbeisch 2007). While these models allow the shape of the trajectory to be indirectly influenced by the survival time through the shared random effect, the intensive computation involved in fitting such models often limits these analyses to using very simple trajectory shapes such as a linear trajectory. However, this restriction does not fit the scenario described in our motivating application. Chan and Wang (2010) studied nonparametric estimation of the mean cost trajectory for a pre-specified time period prior to the terminal events. Their estimated mean trajectory is not conditional on the survival time. The censored survival time data were used to estimate the inverse probability censoring weights, but the cost data from censored patients were not used in the analysis.

In this article, we propose statistical methodology that addresses the three aforementioned challenges in estimating the longitudinal medical cost trajectory. The proposed cost summarization method is completed conditional on the time from diagnosis to the terminal event to properly average the costs for different phases of therapeutic treatment while utilizing all the available data from the patients in the cohort. We believe that this represents a general class of research problems of interest to health economics and health policy studies for which no published statistical methodology is yet available. We present a model and estimation method in Section 2. In Section 3, we provide a data analysis for the motivating application. We present simulation results to assess the empirical performance of the proposed method in Section 4, and a discussion in Section 5.

2. MODEL AND ESTIMATION

2.1 Notation

Suppose we have independent and identically distributed data from n subjects, $\{Y_i, T_i = \min(\tilde{T}_i, C_i), \delta_i = 1\{\tilde{T}_i \leq C_i\}; i = 1, 2, \dots, n\}$, where $\tilde{T}_i$ is the time to the terminal event of interest, such as death, C_i is the censoring time, T_i is the observed time of the terminal event or censoring, whichever is earlier, and δ_i is the event indicator. Let $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ denote the monthly reported medical cost (after proper transformation to reduce the skewness of the data; see Section 3) up to T_i , and let $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{in_i})^T$ be the corresponding months, starting from the time of the initial diagnosis of the disease. For notational convenience, we may also express Y_{ij} as a function of t_{ij} as $Y_{ij} = Y_i(t_{ij})$, and $\mathbf{t}_i$ is deterministic given T_i . When every month's cost data are available, $t_{ij} = j, j = 1, 2, \dots, n_i$ and $t_{in_i} - T_i < t_{in_i} + 1$. Since the cost for a fraction of a month is considered undefined, t_{ij} can only be integers, but T_i is allowed to be continuous. The notations and methodology in this paper extend naturally to applications with other time units, such as quarterly or yearly incident cost data. Throughout this paper, we assume that the censoring time C_i is independent of $\tilde{T}_i$ and Y_i . This assumption is likely to hold in the insurance claims database research similar to the motivating application in this paper, where the interest is in the medical cost till death and the censoring is caused by administrative end of the follow-up.

The goal of the research is to estimate

$$E\{Y(t)|\tilde{T}\} \doteq \mu(t, \tilde{T}), \quad \forall 0 < t \leq \tilde{T}.$$

Although t is an integer in this application, for ease of statistical modeling and estimation, we treat both t and $\tilde{T}$ as continuous variables operationally and make the working assumption that $\mu(t, \tilde{T})$ is a smooth surface over the continuous upper triangular area $\{(t, \tilde{T}) | 0 < t \leq \tilde{T}, 0 < \tilde{T}\}$. This assumption is reasonable given that the longitudinal cost trajectory is expected to be a visually smooth U-shaped curve that depends on the survival times (Section 1).

2.2 Model

The general idea is to model the marginal distribution of survival and the conditional distribution of the incident costs given survival. In the following presentation, we use the notation $f(Y_i|\tilde{T}_i)$ to denote the conditional density function of Y_i given the time of the terminal event $\tilde{T}_i$ and $g(\tilde{T}_i)$ to denote the marginal density function of $\tilde{T}_i$. We use the notation Y_i, T_i, δ_i to denote the random variables or their observed values when there is no ambiguity. The log-likelihood of the observed data is

$$\sum_{i=1}^n \left\{ \delta_i \log \left[f(Y_i|\tilde{T}_i = T_i)g(\tilde{T}_i = T_i) \right] + (1 - \delta_i) \log \int_{T_i}^{\infty} f(Y_i|\tilde{T}_i = s)g(\tilde{T}_i = s)ds \right\}. \quad (1)$$

The marginal distribution of the time to the terminal event $\tilde{T}$ can be estimated nonparametrically (*e.g.*, Kaplan-Meier method) or semiparametrically (*e.g.*, penalized splines; Cai, Hyndman and Wand 2002). Note that when the Kaplan-Meier method is used, we need a regularity condition that the support of the true survival time $\tilde{T}$ lies inside the support of the censoring time C . This condition is needed for calculating the integrand in (1) over (T_i, ∞) .

For any subject i , the model for the conditional distribution of Y_i given the time of the terminal event $\tilde{T}_i$ ($\tilde{T}_i > T_i$) follows a multivariate normal distribution. The j -th ($j = 1, 2, \dots, n_j$) element of Y_i , Y_{ij} , has the distribution

$$Y_{ij}|\tilde{T}_i = s \sim N\left(\mu(t = t_{ij}, \tilde{T}_i = s), \sigma^2\right),$$

where σ^2 is the variance parameter. We use α to denote the correlation parameters for Y_i . The modeling assumption above assumes that the data Y_{ij} are normally distributed with a mean function that varies with t_{ij} , the month of the incident cost, and true survival time s . This mean function is modeled by flexible bivariate splines. Since the medical costs may be highly skewed, it is often necessary to apply a monotone transformation prior to the data analysis to reduce skewness. In such a situation, Y_{ij} denotes the transformed cost data. In subsequent development, we will show that the multivariate normal model not only

simplifies the computation, but also is robust against moderate model violation due to either skewness or zero monthly costs.

For the motivating application in particular, it is expected that $\mu(t, s)$ is a smooth surface over the upper triangular area and $\mu(t, s)$ as a function of t most likely takes a U-shape, which may be bent and twisted by s in a flexible manner. While the proposed methodology below works with any flexible parametric function for $\mu(t, s)$, we prefer to express $\mu(t, s)$ as a linear combination of unknown parameters for ease of computation. This motivates us to consider approximating $\mu(t, s)$ with bivariate penalized splines with a truncated quadratic basis (Ruppert, Wand and Carroll 2003). However, conventional penalized splines for a bivariate surface are defined on a rectangular area and cannot be used directly on surfaces defined only on the upper triangular area. Thus, we propose the following shrinkage-expansion transformation to solve this problem:

$$\mu(t, s) = \tilde{\mu}(u = t(\tau/s), s).$$

The expanded surface $(\tilde{\mu}u, s)$ is defined on the rectangular area $0 < u < \tau, 0 < s < \tau$, where τ is the maximum allowable survival time. We model $\tilde{\mu}(u, s)$ by

$$\tilde{\mu}(u, s) = \mathbf{B}(u, s)^T \boldsymbol{\theta}$$

where $\mathbf{B}(u, s)$ is a vector of a bivariate truncated quadratic basis such that $\mathbf{B}(u, s)$ consists of $(M_1 + 3) \times (M_2 + 3)$ elements of the outer product matrix of vector $(1, s, s^2, (s - c_1)_+^2, \dots, (s - c_{M_1})_+^2)$ and vector $(1, u, u^2, (u - v_1)_+^2, \dots, (u - v_{M_2})_+^2)$, where $c_1, \dots, c_{M_1}$ are pre-specified internal knots on the scale of s , and $v_1, \dots, v_{M_2}$ are internal knots on the scale of u . The number of internal knots M_1 and M_2 must be specified prior to data analysis. $\boldsymbol{\theta}$ is a vector of unknown parameters to be estimated.

Let $\mathbf{D}_i(s)$ be the matrix with the j -th row being $\mathbf{B}(t_{ij}(\tau/s), s)^T$. Under this parameterization, the model can be written as

$$\mathbf{Y}_i | \tilde{T}_i = s \sim \text{MVN}(\mathbf{D}_i(s)\boldsymbol{\theta}, \sum_i(\boldsymbol{\alpha}, \sigma^2)) \quad , \quad s \geq T_i \quad (2)$$

where $\boldsymbol{\Sigma}_i(\boldsymbol{\alpha}, \sigma^2)$ is the variance covariance matrix of $\mathbf{Y}_i$ with dimension $n_i \times n_i$.

2.3 Estimation

Even when the marginal distribution of $\tilde{T}$ is known, direct maximization of the log-likelihood (1) is difficult for two reasons. First, the integrals inside each subject's log-likelihood generally do not have a closed expression, and numerical integration must be used, which increases the complexity and reduces the stability of computation. Second, the marginal log-likelihood (1) is nonlinear with respect to $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, and σ^2 , and maximization must be done numerically. To make the cost trajectory model flexible, a large number of parameters must be used in $\boldsymbol{\theta}$, which further increases the computational difficulty.

We propose to make the computation feasible by treating $\tilde{T}_i$ as missing data and applying the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin 1977). For this approach to work, the continuous survival time $\tilde{T}_i$ must first be “coarsened” into a discrete scale within the interval (T_i, ∞) (Heitjan and Rubin 1991). One advantage of coarsening is that the calculation of the expected complete data log-likelihood in the E-step and its maximization in the M-step can both be done analytically through a closed form formula, as shown below.

If patient i is censored at T_i , we assume that the true time to death $\tilde{T}_i$ falls within one of the K_i pre-specified intervals, denoted by I_{ik} ($k = 1, 2, \dots, K_i$). The marginal survival probability is estimated up to τ . Operationally, we can set $\tau = \max \{T_i | i = 1, 2, \dots, n\}$. The coarsening above must ensure that $\cup_{k=1}^{K_i} I_{ik} = (T_i, \tau)$. Let l_{ik} denote a representative point of interval I_{ik} ; we use the middle point throughout this paper. The coarsening approximation implies that if $\tilde{T}_i \in I_{ik}$, then $\tilde{T}_i = l_{ik}$. The intervals $\{I_{ik}\}$ and their middle points $\{l_{ik}\}$ are specified prior to applying the EM algorithm. We define the intervals of each censored subject to have equal lengths of at least one month, and impose a cap $K = 10$ to be the maximum number of intervals that a censored subject can have. In other words, when both the censoring time T_i and the maximum survival time τ are expressed in month, $K_i = \min\{\lfloor \tau - T_i \rfloor, K\}$, where $\lfloor b \rfloor$ denotes the largest integer that is smaller than b . Therefore, for a censored subject whose T_i is very close to τ , K_i will be smaller than K ; for a censored subject whose T_i is far away from τ , K_i may reach the cap K and the intervals are longer. Imposing the cap K is a simple and effective way to control the computational complexity. Smaller K results in quicker computation.

The coarsening of the survival time implies that if a subject is censored at T_i , $\tilde{T}_i$ can take on one of the values in $\{l_{ik}; k = 1, 2, \dots, K_i\}$. Hence, the indicators $1\{\tilde{T}_i = l_{ik}\}$ are viewed as missing data for censored subjects. Let $\Theta = (\theta, \alpha, \sigma^2)$. The complete data log-likelihood is proportional to

$$\sum_{i=1}^n \left\{ \delta_i \log f(Y_i | \tilde{T}_i = T_i; \Theta) + (1 - \delta_i) \sum_{k=1}^{K_i} 1\{\tilde{T}_i = l_{ik}\} \log f(Y_i | \tilde{T}_i = l_{ik}; \Theta) \right\}. \quad (3)$$

In the E-step, we calculate the conditional expectation of $1\{\tilde{T}_i = l_{ik}\}$ given Y_i , T_i , $\delta_i = 0$ and $\hat{\Theta}_{(m)}$, the estimated Θ in the current iteration (m), as follows:

$$p_{ik}(\hat{\Theta}^{(m)}) = E \left(1\{\tilde{T}_i = l_{ik}\} | Y_i, T_i, \delta_i = 0; \hat{\Theta}^{(m)} \right) \quad (4)$$

$$= \frac{f(Y_i | \tilde{T}_i = l_{ik}; \hat{\Theta}^{(m)}) \hat{g}(\tilde{T}_i = l_{ik} | T_i, \delta_i = 0)}{\sum_{j=1}^{K_i} f(Y_i | \tilde{T}_i = l_{ij}; \hat{\Theta}^{(m)}) \hat{g}(\tilde{T}_i = l_{ij} | T_i, \delta_i = 0)}.$$

Note that before initiating the EM algorithm, the marginal survival function of $\tilde{T}$ can be estimated consistently from $\{T_i, \delta_i; i = 1, 2, \dots, n\}$ using the Kaplan-Meier method. The conditional probability mass functions $g(\tilde{T}_i = l_{ij} | T_i, \delta_i = 0)$ in (4) have been replaced by their estimates calculated prior to the EM iteration.

The expectation of the complete data log-likelihood is

$$\sum_{i=1}^n \left\{ \delta_i \log f(Y_i | \tilde{T}_i = T_i; \Theta) + (1 - \delta_i) \sum_{k=1}^{K_i} p_{ik}(\hat{\Theta}^{(m)}) \log f(Y_i | \tilde{T}_i = l_{ik}; \Theta) \right\}. \quad (5)$$

We call the failure time of the k -th ($k = 1, 2, \dots, K_i$) term of a censored subject i the k -th pseudo failure time of subject i . In (5), we replace the failure time of each censored subject with K_i pseudo failure times at l_{ik} ($k = 1, 2, \dots, K_i$), which are properly weighted, with the total weights summing up to 1, i.e., $\sum_{k=1}^{K_i} p_{ik}(\hat{\Theta}^{(m)}) = 1$.

The M-step can be accomplished in two steps. In the first step, given the estimated $\mathbf{a}^{(m)}$ and $\sigma^{(m)2}$ in the current iteration, we estimate θ by maximizing a penalized version of the expected log complete data likelihood above:

$$Q_p(\theta; \hat{\theta}^{(m)}, \hat{\alpha}^{(m)}, \hat{\sigma}^{(m)}) = Q(\theta; \hat{\theta}^{(m)}, \hat{\alpha}^{(m)}, \hat{\sigma}^{(m)}) - \frac{\lambda}{2} \theta^T \Omega \theta. \quad (6)$$

In this expression, λ is a non-negative penalty parameter that controls the smoothness of the surface fit. The matrix Ω is a penalty matrix pre-specified according to the basis functions. When the truncated quadratic basis is used, Ω is a block diagonal matrix with the first block being a 3×3 matrix of 0's and the second block being an identity matrix (Ruppert, Wand and Carroll 2003). Under the proposed model, (6) is a quadratic function of θ , hence, the maximizer can be solved analytically as a weighted least squares estimator. The penalty term will balance the tradeoff between overfitting and underfitting the model and controls the smoothness of the fit: a very large λ leads to a quadratic fit and a small λ leads to a more flexible fit.

In the second step, given the estimated $\boldsymbol{\theta}^{(m)}$, we estimate σ^2 and $\boldsymbol{\alpha}$ by the method of moments similar to a generalized estimating equations model (Liang and Zeger 1986; Fitzmaurice et al 2009; SAS Institute 2008). Let $\hat{e}_{ij}^{(m)} = Y_{ij} - \mathbf{B}(t_{ij}(\tau/T_i), T_i)^T \hat{\boldsymbol{\theta}}^{(m)}$ be the Pearson residual. The variance σ^2 can be estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \delta_i \{\hat{e}_{ij}^{(m)}\}^2}{\sum_{i=1}^n n_i \delta_i}. \quad (7)$$

For estimating $\boldsymbol{\alpha}$, the parameters associated with the correlation structure, we define $\mathbf{U}_i$ as the $n_i(n_i - 1)/2$ vector of the distinct pairwise products of $\hat{e}_{ij}^{(m)}$, and $\boldsymbol{\rho}_i(\boldsymbol{\alpha})$, the corresponding vector of expectations according to the chosen correlation structure; $\boldsymbol{\alpha}$ can be estimated as the solution to $\sum_{i=1}^n \delta_i \{\partial \boldsymbol{\rho}_i(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}\}^T \{\mathbf{U}_i - \boldsymbol{\rho}_i(\boldsymbol{\alpha})\} = 0$ (Liang and Zeger 1986; Fitzmaurice et al 2009).

The penalty parameter λ is selected by M -fold cross-validation, where $M = 10$ is commonly used (McLachlan, Do and Ambrose 2004), in the following steps.

1. Pre-specify a value for λ . Divide the data set in m subsets with an equal number of subjects in each.
2. For $m = 1, 2, \dots, M$, repeat steps (2.1) and (2.2):
 - 2.1 Use the m -th subset as the test data set, and the other $m - 1$ subsets as the training data set.
 - 2.2 Let $\{\delta_j, T_j, Y_j, \mathbf{t}_j | j = 1, 2, \dots, J_m\}$ denote the test data set. The cross-validated score for this test data set is defined by

$$S_m = \frac{\sum_{j=1}^{J_m} \delta_j \|Y_j - \hat{\mu}(t = \mathbf{t}_j, \tilde{T}_j = T_j)\|^2}{\sum_{j=1}^{J_m} \delta_j n_j},$$

where the bivariate surface $\hat{\mu}(\cdot, \cdot)$ is estimated from the training data.

3. The cross-validation score for λ is $CV(\lambda) = M^{-1} \sum_{m=1}^M S_m$.
4. Calculate $CV(\lambda)$ for a series of candidate values of λ and select the one with the smallest cross-validation score.

Remark 1. While the mean structure of the cost surface is estimated by using all the data, only the non-censored subjects are used to estimate σ^2 and $\boldsymbol{\alpha}$. In our preliminary investigation, we found that including the censored subjects in this step slowed the convergence of the EM algorithm and hence is undesirable in computationally demanding tasks such as bootstrapping or cross-validation. A plausible explanation is that the residuals

from the pseudo subjects are more variable than those from the non-censored subjects due to the uncertainty in $\tilde{T}_i$.

Remark 2. A degree of freedom correction may be used for the denominator in (7). Let $\mathbf{D}$ be the matrix consisting of $\mathbf{D}_i(T_i)$ ($i = 1, 2, \dots, n$; $\delta_i = 1$) stacked one on top of the other, $\mathbf{\Sigma}$ be the block diagonal matrix with the diagonal blocks being $\mathbf{\Sigma}_i(\boldsymbol{\alpha}, \sigma^2)$, and $\mathbf{Y}$ be the concatenated vector of the $\mathbf{Y}_i$'s, with $\delta_i = 1$. The denominator of (7) can be replaced by $\sum_{i=1}^n n_i \delta_i - 2\text{tr}(\mathbf{S}_\lambda) + \text{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda^T)$ (Ruppert, Wand, and Carroll 2003). In our application, the number of subjects and the average number of months per subject are both large. In such a situation, (7) provides a nearly unbiased estimation, while $\mathbf{S}_\lambda$ is costly to compute due to its large size. Therefore, we used (7) in the numerical studies.

The proposed estimation procedure is essentially a two-step approach. The first step is to use the Kaplan-Meier method to estimate the marginal distribution of survival. The asymptotic behavior of the Kaplan-Meier estimate is well established. The second step is to obtain the maximum likelihood estimator of the parameters that characterize the longitudinal cost distribution given the consistent survival distribution estimator from the first step. For the second step, we used bivariate penalized spline. The penalized splines are widely used in statistical methodological research (Ruppert, Wand and Carroll, 2003, 2009), but their asymptotic theory has only been developed in the context of univariate scatter plot smoothers where both the x- and y- variables are continuous (Wang, Shen, Ruppert 2011; Li and Ruppert 2008; Hall and Opsomer 2005). The theory is not directly applicable to the proposed model because one of the two x- variables is the month in which the incident cost is incurred, which is discrete by definition, and the number of distinct months within the follow-up period does not increase with the sample size. Further work is needed to study the asymptotic theory in the context of the proposed model.

3. DATA APPLICATION

The proposed methodology was motivated by the analysis of medical cost data of a study cohort constructed from the Surveillance, Epidemiology and End Results (SEER)–Medicare linked database. The study cohort consists of an incident cohort of 4,950 patients over the age of 65 who were diagnosed with stage IV breast cancer between January 1, 2002 and December 31, 2009. The Medicare payment amounts for inpatient and outpatient services are available for each patient from the date of diagnosis until death or the end of the follow-up, which was December 31, 2010. The overall censoring proportion was 21% and no patients were censored within the first 12 months following diagnosis. The maximum recorded survival time is 104 months, and the median survival time is estimated to be 16 months. The goal of the research was to estimate the population averaged medical costs among these patients so as to learn how the cost varies after diagnosis. We summarized Medicare payments as monthly costs.

We used a $\log(1 + x)$ transformation to reduce the skewness of the medical cost data in the initial model fitting. Figure 1 is a descriptive analysis of the monthly medical costs of individual patients and their averages, stratified by the duration of survival since diagnosis,

among the non-censored subjects. While the individual cost trajectories exhibited large variations, the average transformed costs appeared to follow a U-shape, with higher costs at both tails of the distribution. In Figure 1, for each patient subgroup defined by the year of death, we estimated the standard deviations of the residuals as 2.90, 2.92, 3.11, 3.02, 3.24, and 3.10, respectively. These standard deviations are similar across these patient subgroups, supporting the modeling assumption of constant variance. The month-to-month cost fluctuation was affected by many factors. Since the intra-subject correlation is a nuisance parameter, we opted for an exchangeable autocorrelation structure, which can be interpreted as arising from a latent, patient-specific factor that increases or decreases each patient's monthly costs. The trajectory surface $\mu(t, s)$ is modeled by using a truncated quadratic basis with pre-specified internal knots for survival time s at months 15, 35, 55, and 75, and internal knots for time t at months 10, 25, 45, 65, and 85, leading to a total of 56 basis functions and their corresponding coefficients to model the surface. These knots were chosen a priori to be equally spaced and cover the time scale of s and t .

Figure 2 (A) shows estimated average cost trajectories for patients whose survival time is 24, 36, 48, 60 or 72 months after diagnosis, together with pointwise 95% Wald-type confidence intervals obtained from variances estimated from 200 bootstrap samples. After transforming the estimators back to the dollar cost, we used Figure 2 (B) to illustrate the distribution of the monthly costs for each survival time on the scale of the dollar amount through the 25-th, 50-th, and 75-th quantiles at each month. These quantiles can be estimated with the proposed methodology due to the multivariate normal assumption in (2). Considerable heterogeneity in Medicare reimbursement is observed among patients in this cohort. For example, for patients at the 75-th quantile, their medical costs within the first few months after diagnosis could be as high as \$ 60,000 per month, while for patients at the 25-th quantile, the monthly costs were around \$ 1,000. The curves in Figure 2 also show that the monthly Medicare payment for patients with stage IV breast cancer generally follows a U-shaped curve from the initial diagnosis to death, regardless of the length of survival. The estimated cost trajectory suggests that a plausible cutoff for the initial active treatment phase is 12 months, which coincides with the duration that the NCI used to define the initial care phase. During the first 12 months after diagnosis, the patients receive intensive chemotherapy, radiation therapy, and/or surgical treatments. Interestingly, the costs for the active treatment phase are similar, regardless of the patients' eventual survival times.

Following the active treatment phase is the surveillance phase, during which patients usually receive less intensive treatments, and the main purpose is to monitor tumor recurrence and manage long-term complications. The length of the surveillance phase varies by patient. We observed that patients who lived longer had lower monthly costs in the surveillance phase. However, this observation should not be interpreted as medical costs being predictive of survival. Rather, this pattern is likely driven by the fact that patients who had fewer complications or comorbidities had lower risk of death.

It is important to note that the U-shaped cost trajectories generally are not symmetric. The decline in monthly costs in the active treatment phase appeared to be quicker than the monthly cost increase toward the tail of the surveillance phase, when entering the end-of-life phase. Although almost all patients underwent costly treatments in the active treatment

phase, their cost trajectories started to diverge as they entered the surveillance phase. This resulted in lower average monthly costs in the surveillance phase, with cost starting to climb at various time points, depending on the patients survival. For this reason, the transition between the surveillance phase and the end-of-life phase happened gradually and was heterogeneous, making it difficult to identify a clear cutoff separating the two phases, as was done in the NCIs cost approach. The rich information offered by longitudinal trajectories of monthly cost data cannot be obtained from the analysis of cumulative or total costs, nor can it be obtained by summarizing costs into pre-defined phases of care. This demonstrates the significance of our research question and the important implication of the proposed methodology in health policy as well as clinical research.

The estimated intra-subject correlation and variance were 0.62 and 8.53, respectively. The moderate positive intra-subject correlation suggested considerable variation in costs between patients. For comparison, we also show a “crude” estimate in Figure 2 (A), obtained by averaging the monthly costs over non-censored subjects whose survival times were within the year of the selected survival times. Since the crude averages were wiggly, we smoothed them with penalized splines (R `SemiPar` package with the default settings) for better visualization. Even after smoothing, the crude average trajectory estimates showed some local curvy shapes that were difficult to interpret (e.g., the strata corresponding to patients who died at 48 or 60 months after diagnosis).

The proposed methodology can provide an estimated average cost trajectory for each distinct survival time. Putting all of these trajectories together, we obtained a “smooth” surface over the triangular region (Figure 3 (A)). This plot is complementary to Figure 2 because Figure 2 only shows the surface at a few cross-sections. We placed a quotation mark on “smooth” because the “month” variable cannot be continuous in the algebraic sense when we deal with monthly cost data. Since the month of the reported cost must be no greater than the total number of months from diagnosis to death or censoring, the surface is defined only on an upper triangular region. This is different from the conventional bivariate smoothing problems, and we solved this problem by using the shrinkage-expansion transformation in Section 2. We illustrate the estimated surface in both a contour plot and a 3-D plot in Figure 3.

The penalty parameter λ was selected by the proposed cross-validation procedure. We calculated the cross-validation score $CV(\lambda)$ for $\lambda = 5 \times 10^x$ with $x = -6, -5, -4, -3, -2, -1, 0, 1$. The corresponding cross-validation scores are equal to 9.14, 9.13, 9.13, 9.15, 9.23, 9.27, 9.28, and 9.28, respectively. This result suggests that the desired smoothness can be obtained when λ is below 5×10^{-3} ; the cross-validation score increases quickly as λ increases beyond this threshold. Note that the above grid for λ covers a wide range as adjacent values represent 10-fold increase. As λ decreases from 5×10^{-3} to 5×10^{-6} , the run time for obtaining the point estimator increases from 27 minutes to 81 minutes on a computer with 3.30 GHz CPU and 16 GB memory, suggesting that a strong penalty restricts the movement of the coefficients of the basis functions and helps convergence. We also observe more local oscillation when λ is small. Given that a smooth U-shaped trajectory with no local oscillation is supported by the previous literature and clinical practice, we

chose the penalty parameter at 0.002, the middle point of the interval $(5 \times 10^{-4}, 5 \times 10^{-3})$, for the final analysis in this section. The corresponding cross-validation score is 9.14.

4. SIMULATION

We conducted simulations to assess the finite sample performance of the proposed method. We generated the true survival time $\tilde{T}_i$ from an exponential distribution with mean 45, right truncated at 100 so that the maximum survival τ is taken to be 100 months. The log medical cost Y_{ij} of a subject with survival time $\tilde{T}_i$ at month t_{ij} is generated according to the model in Section 2 that ensures a U-shaped trajectory for any survival time. The residual variance was 4 and the intra-subject correlation was 0.5. The knots of survival are placed at months 15, 35, 55, and 75, and the knots for monthly costs are placed at months 10, 25, 45, 65, and 85. We considered four scenarios with a sample size of 1000 or 5000, and censoring rate of 20% or 50%. In each scenario, we conducted 200 Monte Carlo simulations and averaged the results. The convergence criterion is that for each element θ in Θ_1 , $|\theta^{(m)} - \theta^{(m-1)}|$ should be less than 10^{-4} . The penalty parameter is chosen as $\lambda_1 = 0.01$ in preliminary runs.

Figure 4 shows that the proposed trajectory estimator has little bias when estimating the true trajectory (the red and black curves almost overlap). The crude method has some finite-sample bias, mostly because of its large variation: it does not use the cost data from the censored subjects and the estimation of one trajectory does not borrow information from the estimation of other trajectories. As in the simulation as well as in the data application, the subgroup sample size may be relatively small for subjects with long survival times, and this makes the crude method inefficient. Figure 5 shows that the proposed method has a favorable mean squared error (MSE) compared with the crude method in all scenarios. The MSE improvement is more prominent at the right tails of the curves where the data are sparser due to censoring.

We conducted additional simulations to study the robustness of the proposed method in the context of the data application. The simulation design and results are included in Section S1 of the Online Supplementary Materials. Specifically, we studied the performance of the proposed method when the distribution of the data deviates from multivariate normality. We simulated data from either a skew normal distribution or a mixture of two normal distributions. The simulation results show that under these two types of misspecified modeling assumptions, the proposed model produces trajectory estimates that nearly overlap with the true trajectories. Similar robustness of the normal theory models have been widely recognized in other statistical contexts (Carroll and Ruppert, 1988).

5. DISCUSSION

Motivated by the challenges in summarizing and modeling Medicare costs for patients with stage IV breast cancer, this paper studied longitudinal medical cost data from a new perspective, i.e., estimating the average incident cost trajectory conditional on survival. The survival time and the mean cost trajectory form a bivariate surface on a triangular region, which is estimated by a semiparametric model that includes a bivariate penalized spline model for the surface, and nonparametric estimation for survival. The estimation properly

accounts for right censoring of survival times. The rich information offered by longitudinal trajectories of monthly cost data cannot be obtained from the analysis of cumulative or total costs, as is the focus of statistical literature reviewed in Section 1, nor can it be obtained by summarizing costs into pre-defined phases of care, as done by the NCI's cost summary. This highlights the significance of our research question and proposed methodology. Using the proposed methods, we found that almost all stage IV breast cancer patients underwent costly treatments during the first 6–12 months after diagnosis, though the amount varied substantially. Their cost trajectories started to diverge as they entered the surveillance phase after completing active treatment, and the treatment cost started to climb at various time points, depending on the patients survival time. Because the transition between the surveillance phase and the end-of-life phase happened gradually and is heterogeneous across patients, there is no clear-cut time separating the two phases, as used in the NCI's cost summary. The R code for implementing the proposed methodology is available upon request.

The assumption that the transformed monthly costs of a patient follow a multivariate normal distribution is both a limitation and advantage of the proposed method. On the one hand, this assumption may be violated if an excessive proportion of the monthly medical costs are zero. In our data set, 10% of the 103,973 monthly incident costs from the 4,950 patients were zero. The proportion of zero cost months was small in this data set because the study cohort consisted of patients with stage IV breast cancer. Alternatively, we can aggregate the monthly costs into quarterly costs, which resulted in just 3% zero quarterly costs in our study cohort. In populations with less disease burden, there may be more zero costs in the data and it may be necessary to introduce a two-part model (Liu et al 2008) feature to our proposed methodology for adjustment. That will be studied in our future work. However, an advantage of building this parametric distribution assumption into the proposed semiparametric modeling framework is that once the estimation is complete on the transformed cost data, we can easily obtain the distribution of monthly costs in dollar amounts given any survival time, including medians and various other quantiles, as illustrated in Figure 2 (B). Without this assumption, one would need to jointly model survival and multiple nonlinear quantile functions of skewed longitudinal cost data, for which statistical methodology is yet to be developed. In Section S1 of the Online Supplementary Materials, we present a simulation study with data simulated from either a skew normal distribution or a mixture of two normal distributions. The proposed method is shown to produce robust estimated trajectories against those two types of model violation.

We used coarsening approximation with the maximum number of intervals $K \equiv 10$ in the numerical studies. The proposed coarsening approximation is equivalent to using a discrete distribution to approximate the conditional distribution of $\tilde{T}_i$ given the observed data and $\tilde{T}_i > T_i$. We compared the coarsening approximation with $K = 10$ with another, much finer approximation that sets all the intervals of all subjects to be about one month. Since the proposed coarsening rule (Section 2.3) requests each interval to be at least one month in length, the second approach is equivalent to not setting the cap K , and K_i , the number of intervals per subject, roughly equals to the number of months from the censoring time to τ . In the data analysis of Section 3, if we use the monthly coarsening approximation, the EM

algorithm converges in 232 minutes on a computer with 3.30 GHz CPU and 16 GB memory. For a coarsening approximation with $K = 10$, the convergence takes 30 minutes. The estimated surfaces with or without the coarsening approximation are nearly the same. Hence, in this example the coarsening approximation reduced 87% of the computing time without notable change in the estimates. In Section S2 of the Online Supplementary Materials, we present a simulation study to compare the coarsening approximation with $K = 10$ and the monthly coarsening approximation without the cap K . The simulation shows that two approximations produced very similar estimated trajectories with no noticeable differences. Both are very close to the true trajectories. This robustness and time-saving feature can be important in some research studies. In studies of health economics, medical cost data are usually obtained from large patient registries. Therefore, we believe that developing computationally tractable approaches for large data sets is very important for this kind of problem. In a shared parameter joint model of longitudinal and survival data, Barrett et al (2015) demonstrated that proper discretization of the time scale can make the computation tractable without notable bias or efficiency loss.

The proposed methodology can be extended to terminal events that are subject to competing risks (e.g., breast cancer death vs. death due to other causes) by replacing the Kaplan-Meier estimate with the cause-specific survival distribution. This will lead to estimated cost trajectories specific to the designated terminal event type. The goal of this paper is to estimate the population averaged longitudinal cost trajectory with proper adjustment for survival and censoring. This type of marginal estimation is useful for health services research and health policy making, to provide a basis for cost projection of a defined population. For future research, it is of interest to study how the shapes of the possibly nonlinear longitudinal cost trajectories are influenced by both covariates and survival.

In Section 2.2, a regularity condition needed for estimating the entire marginal distribution of $\tilde{T}$ by the Kaplan-Meier method is that the support of $\tilde{T}$ lies inside the support of the censoring time C . This regularity condition holds approximately for the motivating data application, where the terminal event is death and only about 6.5% of the subjects survived beyond the largest observed follow-up time at 104 months. When a substantial proportion of subjects are censored beyond the largest observed failure time, this regularity assumption does not hold, making $g(\tilde{T}_i = s)$ in (1) not everywhere identifiable in the interval (T_i, ∞) for the integral. The proposed methodology can be extended to accommodate this situation and estimate the cost trajectory up to the maximum follow-up time. Let τ_C denote the maximum follow-up time, and subjects are censored beyond τ_C . The complete data log-likelihood (3) can be modified to be

$$\sum_{i=1}^n \left\{ \delta_i \log f(Y_i | \tilde{T}_i = T_i; \Theta) + (1 - \delta_i) \sum_{k=1}^{K_i} 1\{\tilde{T}_i = l_{ik}\} \log f(Y_i | \tilde{T}_i = l_{ik}; \Theta) + (1 - \delta_i) 1\{\tilde{T}_i > \tau_C\} \log f(Y_i | \tilde{T}_i > \tau_C; \Theta_2) \right\}.$$

For each censored subject, the coarsening approximation uses $K_i + 1$ intervals: the K_i intervals that satisfy $\cup_{k=1}^{K_i} I_{ik} = (T_i, \tau_C)$ plus one extra interval to account for the possibility that $\tilde{T}_i \in (\tau_C, \infty)$. Additional parameters Θ_2 are introduced to characterize the conditional distribution of Y_i when $\tilde{T}_i$ falls in this extra interval (τ_C, ∞) . The designation of the middle point $I_{K(K_i+1)}$ is not needed for this interval. The rest of the EM algorithm is similar to that in Section 2.3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Duncan Family Institute (Shih), the Agency for Healthcare Research and Quality (Shih, R01HS020263) and the Cancer Center Support Grant from the NIH (Li, Ning, Huang, Shih, Shen, P30CA016672). The authors would like to thank Ying Xu, Department of Health Services Research at The University of Texas MD Anderson Cancer Center, for the creation of the analytical sample, and thank Lee Ann Chastain, Department of Biostatistics, for editing assistance. This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the Applied Research Program, NCI; the Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database. Dr. Shih had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

- Akushevich I, Kravchenko J, Akushevich L, Ukraintseva S, Arbeev K, Yashin AI. 2011 Medical cost trajectories and onsets of cancer and noncancer diseases in US elderly population. *Comput Math Methods Med.* :857–892.
- Bang H. 2005; Medical cost analysis: application to colorectal cancer data from the SEER Medicare database. *Contemp Clin Trials.* 26(5):586–597. [PubMed: 16084777]
- Bang H, Tsiatis AA. 2000; Estimating medical costs with censored data. *Biometrika.* 87(2):329–343.
- Barrett J, Diggle P, Henderson R, Taylor-Robinson D. 2015; Joint modelling of repeated measurements and time-to-event outcomes: exible model specication and exact likelihood inference. *Journal of the Royal Statistical Society, Series B.* 77(1):131–148.
- Baser O, Gardiner JC, Bradley CJ, Yuce H, Given C. 2006; Longitudinal analysis of censored medical cost data. *Health Economics.* 15(5):513–525. [PubMed: 16389670]
- 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 Suppl):IV-104–17.
- Cai T, Hyndman RJ, Wand MP. 2002; Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics.* 11(4):784–798.
- Carroll, RJ, Ruppert, D. *Transformation and Weighting in Regression.* Chapman & Hall; 1988.
- Chan KCG, Wang M-C. 2010; Backward estimation of stochastic processes with failure events as time origins. *Annals of Applied Statistics.* 4(3):1602–1620. [PubMed: 21359167]
- Chen J, Liu L, Shih YC, Zhang D, Severini TA. 2016; A flexible model for correlated medical costs, with application to medical expenditure panel survey data. *Statistics in Medicine.* 35(6):883–894. [PubMed: 26403805]
- Chen J, Liu L, Zhang D, Shih YC. 2013; A flexible model for the mean and variance functions, with application to medical cost data. *Statistics in Medicine.* 32(24):4306–4318. [PubMed: 23670952]
- Dempster AP, Laird NM, Rubin DB. 1977; Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B.* 39:1–38.

- Dumont S, Jacobs P, Turcotte V, Anderson D, Harel F. 2010; The trajectory of palliative care costs over the last 5 months of life: a Canadian longitudinal study. *Palliative Medicine*. 24(6):630–640. [PubMed: 20501512]
- Etzioni R, Urban N, Baker M. 1996; Estimating the costs attributable to a disease with application to ovarian cancer. *Journal of Clinical Epidemiology*. 49(1):95–103. [PubMed: 8598518]
- Fitzmaurice, G, Davidian, M, Verbeke, G, Molenberghs, G. *Longitudinal Data Analysis*. Chapman & Hall/CRC; 2009.
- Hall P, Opsomer JD. 2005; Theory for penalized spline regression. *Biometrika*. 92:10518.
- Heitjan DF, Rubin DB. 1991; Ignorability and Coarse Data. *Annals of Statistics*. 19(4):2244–2253.
- Li Y, Ruppert D. 2008; On the asymptotics of penalized splines. *Biometrika*. 95(2):415–436.
- Liang KY, Zeger SL. 1986; Longitudinal data analysis using generalized linear models. *Biometrika*. 73:13–22.
- Lin DY. 2000; Linear regression analysis of censored medical costs. *Biostatistics*. 1(1):35–47. [PubMed: 12933524]
- Lin DY. 2003; Regression analysis of incomplete medical cost data. *Statistics in Medicine*. 22(7): 1181–1200. [PubMed: 12652561]
- Lin DY, Feuer EJ, Etzioni R, Wax Y. 1997; Estimating medical costs from incomplete follow-up data. *Biometrics*. 53:419434.
- Liu L. 2009; Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*. 28(6):972–986. [PubMed: 19040210]
- Liu L, Conaway MR, Knaus WA, Bergin JD. 2008; A random effects four-part model, with application to correlated medical cost. *Computational Statistics and Data Analysis*. 52:44584473.
- Liu L, Huang X, O’Quigley J. 2008; Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*. 64(3):950–958. [PubMed: 18162110]
- Liu L, Wolfe RA, Kalbeisch JD. 2007; A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*. 26:139155.
- Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. 2011; Projections of the cost of cancer care in the United States, 2010–2020. *Journal of the National Cancer Institute*. 103(2):117–128. [PubMed: 21228314]
- McLachlan, GJ, Do, Kim-Anh; Ambroise, C. *Analyzing microarray gene expression data*. John Wiley & Sons, Inc.; 2004.
- O’Hagan A, Stevens JW. 2004; On estimators of medical costs with censored data. *Journal of Health Economics*. 23:615625.
- Pullenayegum EM, Willan AR. 2011; Marginal models for censored longitudinal cost data: appropriate working variance matrices in inverse-probability-weighted GEEs can improve precision. *The International Journal of Biostatistics*. 7(1)
- Ruppert, D, Wand, MP, Carroll, RJ. *Semiparametric Regression*. Cambridge University Press; 2003.
- Ruppert D, Wand MP, Carroll RJ. 2009; *Semiparametric Regression during 2003–2007*. *Electronic Journal of Statistics*. 3:11931256.
- SAS Institute. *SAS/STAT 9.2 User’s Guide: The GENMOD Procedure*. SAS Institute Inc.; Cary, NC, USA: 2008.
- Wang X, Shen J, Ruppert D. 2011; On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*. 5:1–17.
- Yabroff KR, Mariotto AB, Feuer E, Brown ML. 2008; Projections of the costs associated with colorectal cancer care in the United States, 2000–2020. *Health Economics*. 17(8):947–959. [PubMed: 17910108]
- Yabroff KR, Lamont EB, Mariotto A, Warren JL, Topor M, Meekins A, Brown ML. 2008; Cost of care for elderly cancer patients in the United States. *Journal of National Cancer Institute*. 100(9):630–641.
- Yabroff KR, Warren JL, Schrag D, Mariotto A, Meekins A, Topor M, Brown ML. 2009; Comparison of approaches for estimating incidence costs of care for colorectal cancer patients. *Medical Care*. 47(7 Suppl 1):S56–563. [PubMed: 19536010]

- Zhao H, Cheng Y, Bang H. 2011; Some insight on censored cost estimators. *Statistics in Medicine*. 30(19):2381–2388. [PubMed: 21748774]
- Zhao H, Tian L. 2001; On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics*. 57:10021008.
- Zhao H, Zuo C, Chen S, Bang H. 2012; Nonparametric inference for median costs with censored data. *Biometrics*. 68(3):717–725. [PubMed: 22364557]

Figure 1.

Individual monthly medical cost data after $\log(1 + x)$ transformation, stratified by the time gap from diagnosis to death. The black lines are the average monthly costs among all the non-censored patients in each stratum. A random sample of 50 non-censored patients was selected from each stratum and their individual monthly cost trajectories are illustrated by the gray curves in the background.

Figure 2.

Cost trajectories corresponding to patients with survival times of 24, 36, 48, 60, and 72 months after diagnosis from the breast cancer data set. (A) The red solid curves are the estimated average cost trajectories; the red dashed curves are the 95% pointwise confidence intervals. The crude estimate is illustrated with green curves for comparison. (B) The 25-th (pink), 50-th (red), and 75-th (purple) quantiles of monthly costs estimated by the proposed method, which correspond to different survival times.

Figure 3.

The estimated average trajectory surface in (A) contour plot and (B) 3-D plot from the SEERMedicare data set of patients with stage IV breast cancer.

Figure 4.

Simulation results for the estimation of the average cost trajectory corresponding to patients whose survival times are 10, 20, 30, ... 90 months respectively when the censoring rate is 20% or 50%. The sample size is 1000. The true trajectories are plotted in black curves. The mean estimated trajectories, averaged over 200 repetitions, are plotted in red curves. The black and red curves almost overlap, suggesting little bias. The green lines denote the results from the crude method, as introduced in Section 3.

Figure 5.

Simulation results for the mean squared error (MSE) of the proposed method (red lines) in comparison with the crude method (green dashed lines) for different combinations of sample size ($n=1000, 5000$) and censoring rate (20%, 50%). The MSE was evaluated for trajectories up to 20, 40, 60 and 80 months.