An Effective Method for Online Disease Risk Monitoring

Lu You and Peihua Qiu
Department of Biostatistics
University of Florida
Gainesville, FL 32610

Abstract

Many diseases can be prevented or treated if they can be detected early or signaled before their occurrence. Disease early detection and prevention (DEDAP) is thus important for health improvement of our society. Traditionally, people are encouraged to check their health conditions regularly so that readings of relevant medical indices can be compared with certain threshold values and any irregular readings can trigger further medical tests in order to find root causes or diseases. One limitation of such traditional DEDAP methods is that they focus mainly on the data collected at the current time point and historical data are not fully used. Consequently, irregular longitudinal pattern of the medical indices could be neglected and certain diseases could be left undetected. In this paper, we suggest a novel and effective new method for DEDAP. To detect a disease by this method, a patient's risk to the disease is first quantified at each time point, and then the longitudinal pattern of the risk is monitored sequentially over time. A signal will be triggered by a large cumulative difference between the longitudinal risk pattern of the patient under monitoring and the longitudinal risk pattern of a typical person without the disease in concern. Both theoretical arguments and numerical studies show that it works well in practice.

Key Words: Disease early detection; Longitudinal data; Online monitoring; Risk factors; Sequential test; Statistical process control.

1 Introduction

Many diseases (e.g., chronic diseases) can be prevented or treated if they can be detected early. Disease early detection and prevention (DEDAP) is thus a critically important research problem in public health and medical research. To this end, past medical research has discovered major risk factors for many different diseases. For instance, the discovered major risk factors of cardiovascular

diseases (CVDs) include high blood pressure, high cholesterol level, obesity, tobacco use, lack of physical activity, diabetes, unhealthy diet, age, family history, and more (e.g., Mendis et al. 2011). After the major risk factors of a disease are found, medical doctors/researchers can use them for disease prediction and diagnostics. Most current methods for this purpose compare the readings of the risk factors of a given patient collected at a given time point cross-sectionally with those of a properly chosen healthy population. These methods, however, have not made use of all history data of the given patient. This paper aims to develop a novel and effective new statistical method for DEDAP, which effectively combines cross-sectional comparisons between the given patient and the healthy population and sequential monitoring over time of a properly quantified risk to a disease in question of the given patient. Thus, both the data collected at the current time point and all history data are used in the new method; consequently, the disease can be detected effectively.

Because of its importance mentioned above, there has been a tremendous amount of existing research on DEDAP. Much of the existing research is on finding relevant disease risk factors (e.g., Radbill et al. 2008, Ridker 2003). For certain diseases with signs and symptoms (e.g., breast cancer, stroke), some research focuses on increasing the awareness of early signs and symptoms (e.g., Hubbard et al. 2014, Robb et al. 2009). For high-risk populations of a given disease, systematic application of screening tests is especially important. In such cases, more frequent tests would increase success rates of disease diagnoses. But, they would increase the medical cost as well. So, there have been debates in the literature about the real benefits of certain disease early detection measures, e.g., breast cancer detection using mammograms (Kattlove et al. 1995). To balance the diagnostic success rate and medical cost, much research has devoted to the "optimal" design of sampling schemes for the screening tests (e.g., Cohn et al. 2003, Lee et al. 2004, Zelen 1993). After data are collected about certain disease risk factors, almost all existing diagnostic methods compare the observed readings of a given patient with some threshold values. For instance, the resting blood pressure would be diagnosed abnormal if the systolic blood pressure (SBP) is at least 140 or the diastolic blood pressure (DBP) is at least 90. These threshold values are usually obtained from observed data of a healthy population. Therefore, these diagnostic methods actually try to detect diseases by comparing the readings of certain disease risk factors of a given patient cross-sectionally with the readings of a healthy population.

The traditional diagnostic methods mentioned above to compare the observed readings of a given patient with some threshold values are related to the longitudinal data analysis (LDA) in

the statistical literature. By an LDA method, we can construct confidence intervals for the mean functions of the disease risk factors at different time points from an observed longitudinal dataset of a properly chosen healthy population. The constructed confidence intervals can be used for describing the regular longitudinal pattern of the disease risk factors. Some existing methods for constructing such confidence intervals include Chen and Jin (2005), Li (2011), Liang and Zeger (1986), Lin and Carroll (2001), Ma et al. (2012), Wang (2003), Xiang et al. (2013), and Zhao and Wu (2008). *However, these LDA methods are retrospective and cannot be used for prospective monitoring of disease risk factors.* Because they determine the disease status based on the observed data at the current time point only, they are generally ineffective in detecting diseases early. Another statistical research area related to DEDAP is statistical process control (SPC), which is mainly for sequential monitoring of production lines in manufacturing industries. By a SPC chart, we can sequentially monitor disease risk factors of each patient, and a signal can be triggered as soon as the chart detects a shift in the longitudinal pattern of the disease risk factors of the patient from an in-control (IC) status to an out-of-control (OC) status (cf., Hawkins and Olwell 1998, Qiu 2014). However, a conventional SPC chart cannot be applied to the DEDAP problem directly for the following reason. In a typical SPC problem, the distribution of process observations is assumed to be unchanged over time when the process is IC. In the DEDAP problem, however, this distribution can change over time (e.g., the mean total cholesterol level of a healthy person would change with age). To overcome this difficulty, Qiu and Xiang (2014, 2015) suggested modifications of the conventional SPC charts so that the modified charts can be used in cases with time-varying IC process distributions. The major idea behind the modifications is that the regular longitudinal pattern of the disease risk factors can first be estimated from a dataset of a healthy population, and then observations of a person under monitoring can be standardized by the estimated regular longitudinal pattern before a control chart is applied. For different versions of the modifications to address different issues (e.g., serial data correlation), see papers such as Li and Qiu (2016, 2017), Li et al. (2018), Liang et al. (2017), Qiu et al. (2018), and You and Qiu (2019). For a new performance measure of these DEDAP methods, see Qiu et al. (2019).

In this paper, we suggest a novel and effective new method for DEDAP. Existing methods discussed above are either ineffective because they are retrospective (e.g., the LDA methods) or improper to use because of their assumption of time-independent IC process distribution (e.g., the conventional SPC charts). The modified SPC charts, such as the ones in Qiu and Xiang

(2014, 2015), are ineffective because they sequentially monitor all standardized disease risk factors simultaneously, no matter whether some of them cannot actually provide useful information for predicting the occurrence of the disease in question. To effectively monitor the disease risk factors of a person, our suggested new method first quantifies the risk to the disease in question at each time point, based on observations of the original disease risk factors. During the risk quantification process, the original disease risk factors are weighted differently according to their relevance to the disease. Then, the quantified risk to the disease is monitored sequentially over time. Both theoretical arguments and numerical studies show that this new method indeed provides an effective tool for disease early detection.

The remainder of the article is organized as follows. Our proposed new method is described in detail in Section 2. Its numerical performance is evaluated in Section 3. A real-data example for demonstrating its application is discussed in Section 4. Several remarks conclude the article in Section 5. Some technical details are provided in Appendix.

2 Proposed Method

Our proposed method consists of two steps. A survival model is first fitted from a training data, by which the risk to a disease in question can be estimated. Then, the estimated risk to the disease is monitored sequentially and a signal is triggered if the cumulative risk over time exceeds a threshold value. A detailed description of the proposed method is given below.

2.1 Risk estimation by survival modelling

Assume that a training data contains observations of n individuals. For each individual, the following survival and longitudinal data are observed. First, let T_i be the last follow-up time of the ith individual, δ_i be the indicator whether a disease is observed at T_i, and D_i be the true disease occurrence time, for $i = 1, 2, \ldots, n$. Then, define $T_i = \min\{D_i, C_i\}$ and $\delta_i = I(D_i \leq C_i)$, where C_i denotes the censoring time. Second, for the q disease risk factors in $\boldsymbol{x}_i(t)$, they are observed at times $\{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$, where these observation times may not be equally spaced in the study period $[0, \mathcal{T}]$ and $t_{im_i} = T_i$. Finally, let $\boldsymbol{z}_i$ be a vector of p time-independent covariates representing baseline measurements that can explain the heterogeneity of the population. These

observed training data are assumed to follow the Cox proportional hazards model

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t) + \boldsymbol{\gamma}' \boldsymbol{z}_i), \qquad \text{for } t \in [0, \mathcal{T}], \tag{1}$$

where $\lambda_i(t)$ is the hazard rate function of the ith individual, $\lambda_0(t)$ is the baseline hazard function, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are respectively q-dimensional and p-dimensional vectors of coefficients. From (1), it can be seen that the hazard ratio $\lambda_i(t)/\lambda_0(t)$ of the ith person is an increasing function of $\boldsymbol{\beta}' \boldsymbol{x}_i(t) + \boldsymbol{\gamma}' \boldsymbol{z}_i$, in which $\boldsymbol{\beta}' \boldsymbol{x}_i(t)$ represents the contribution of the risk factors in $\boldsymbol{x}_i(t)$ and $\boldsymbol{\gamma}' \boldsymbol{z}_i$ explains the heterogeneity among different individuals. Because our research goal is to detect the disease in question early by monitoring the risk factors in $\boldsymbol{x}_i(t)$, we focus mainly on the part $r_i(t) = \boldsymbol{\beta}' \boldsymbol{x}_i(t)$, which is called the risk function of the ith individual here. For simplicity of presentation, let $R(t) = \{i : T_i \geq t\}$ denote the set of people at risk at time t, $y_i(t) = I(T_i \geq t)$ indicate whether the ith individual is at risk at time t, and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ be the set of all parameters in the model (1).

Next, we discuss estimation of the model (1). Because the observation times $\{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$ of $\boldsymbol{x}_i(t)$ could be unequally spaced and $\boldsymbol{x}_i(t)$ are only observed at $\{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$, if the model (1) is fitted as usual by the conventional partial likelihood estimation, then $\boldsymbol{x}_i(t)$ has to be interpolated or extrapolated at some observed disease occurrence times of other people in the training data. A naive solution is to impute $\boldsymbol{x}_i(t)$ by the last-observation-carried-forward method, by which $\boldsymbol{x}_i(t)$ at an unobserved time is imputed by its last observed value. However, this method will not guarantee an unbiased estimate for $\boldsymbol{\beta}$ when $\boldsymbol{x}_i(t)$ are not piecewise constant or when $\{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$ are sparsely distributed. In the literature, there are some existing discussions about different strategies to extrapolate $\boldsymbol{x}_i(t)$ (cf., Lin and Ying 1993, Paik and Tsai 1997). However, these methods need to assume that $\boldsymbol{x}_i(t)$ are completely observed in an interval, which is invalid in the current problem. When $\boldsymbol{x}_i(t)$ follows a parametric longitudinal model, some people suggested methods for joint modeling of the survival and longitudinal data (e.g., Dupuy et al. 2006), but the assumed parametric longitudinal model is often hard to justify in practice. In this paper, we suggest a kernel smoothing method for a consistent estimation of $\boldsymbol{\beta}$ without imposing a parametric assumption on $\boldsymbol{x}_i(t)$. In the literature, kernel smoothing methods have been found useful in estimating time-varying effects in the Cox proportional hazards modeling framework (e.g., Cai and Sun 2003, Yu and Lin 2010), where time-dependent covariates are usually assumed to be observable at any time point in $[0, \mathcal{T}]$. Here, we extend the use of kernel smoothing methods to cases when the time-dependent covariates $\boldsymbol{x}_i(t)$ are only observed at several irregularly spaced time points. Details of the proposed estimation

procedure are given below.

In the conventional partial likelihood estimation procedure, the partial likelihood function is defined as

$$\prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i(T_i)+\boldsymbol{\gamma}'\boldsymbol{z}_i)}{\sum_{l\in R(T_i)}\exp(\boldsymbol{\beta}'\boldsymbol{x}_l(T_i)+\boldsymbol{\gamma}'\boldsymbol{z}_i)}.$$

This function cannot be used here because many terms in $\{r_l(T_i)=\boldsymbol{\beta}'\boldsymbol{x}_l(T_i)\}$ are not observed, as discussed above. To overcome this difficulty, we suggest using the following local smoothing partial likelihood function:

$$L(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i(T_i)+\boldsymbol{\gamma}'\boldsymbol{z}_i)}{\sum_{l\in R(T_i)}\sum_{j=1}^{m_l} K_{h_\theta}(T_i-t_{lj})\exp(\boldsymbol{\beta}'\boldsymbol{x}_l(t_{lj})+\boldsymbol{\gamma}'\boldsymbol{z}_l)}, \tag{2}$$

where $K_{h_\theta}(s)=K(s/h_\theta)/h_\theta$, $K(s)$ is a density kernel function, and $h_\theta>0$ is a bandwidth. In (2), $\{r_l(t_{lj}), j=1,2,\ldots,m_l\}$ are weighted averaged for estimating $r_l(T_i)$ and the weights are determined by the kernel function and the bandwidth. To estimate $\boldsymbol{\theta}$, we can work with the following logarithm of $L(\boldsymbol{\theta})$:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log(L(\boldsymbol{\theta})) \\ &= \sum_{i:\delta_i=1}\left\{\boldsymbol{\beta}'\boldsymbol{x}_i(T_i)+\boldsymbol{\gamma}'\boldsymbol{z}_i - \log\left[\sum_{l\in R(T_i)}\sum_{j=1}^{m_l} K_{h_\theta}(T_i-t_{lj})\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_l(t_{lj})+\boldsymbol{\gamma}'\boldsymbol{z}_l\right)\right]\right\}. \end{aligned}$$

Then, the estimates of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{3}$$

To compute $\widehat{\boldsymbol{\theta}}$ by (3), the following Newton-Raphson iterative algorithm can be used: for $j\geq 0$, let

$$\widehat{\boldsymbol{\theta}}^{(j+1)} = \widehat{\boldsymbol{\theta}}^{(j)} - \left[\nabla^2 l(\widehat{\boldsymbol{\theta}}^{(j)})\right]^{-1}\nabla l(\widehat{\boldsymbol{\theta}}^{(j)}),$$

where $\widehat{\boldsymbol{\theta}}^{(0)} = \left(\boldsymbol{\beta}^{(0)\prime}, \boldsymbol{\gamma}^{(0)\prime}\right)' = \mathbf{0}_{(q+p)\times 1}$, and

$$\nabla l(\boldsymbol{\theta}) = \sum_{i:\delta_i=1} \left[\begin{pmatrix} \boldsymbol{x}_i(T_i) \\ \boldsymbol{z}_i \end{pmatrix} - \frac{\mathbf{S}_1(\boldsymbol{\theta}; T_i)}{S_0(\boldsymbol{\theta}; T_i)} \right],$$

$$\nabla^2 l(\boldsymbol{\theta}) = -\sum_{i:\delta_i=1} \left[\frac{S_0(\boldsymbol{\theta}; T_i)\mathbf{S}_2(\boldsymbol{\theta}; T_i) - \mathbf{S}_1(\boldsymbol{\theta}; T_i)\mathbf{S}_1'(\boldsymbol{\theta}; T_i)}{[S_0(\boldsymbol{\theta}; T_i)]^2} \right],$$

$$S_0(\boldsymbol{\theta}; t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\theta}(t - t_{ij}) y_i(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}_i(t_{ij}) + \boldsymbol{\gamma}'\boldsymbol{z}_i),$$

$$\mathbf{S}_1(\boldsymbol{\theta}; t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\theta}(t - t_{ij}) y_i(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}_i(t_{ij}) + \boldsymbol{\gamma}'\boldsymbol{z}_i) \begin{pmatrix} \boldsymbol{x}_i(t_{ij}) \\ \boldsymbol{z}_i \end{pmatrix},$$

$$\mathbf{S}_2(\boldsymbol{\theta}; t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\theta}(t - t_{ij}) y_i(t) \exp(\boldsymbol{\beta}'\boldsymbol{x}_i(t_{ij}) + \boldsymbol{\gamma}'\boldsymbol{z}_i) \begin{pmatrix} \boldsymbol{x}_i(t_{ij}) \\ \boldsymbol{z}_i \end{pmatrix}^{\otimes 2}.$$

In the last expression above, the outer product of a vector $\boldsymbol{x}$ with itself is denoted as $\boldsymbol{x}^{\otimes 2}$. The algorithm stops at the $(j+1)$th iteration when $\|\widehat{\boldsymbol{\theta}}^{(j+1)} - \widehat{\boldsymbol{\theta}}^{(j)}\| \le \epsilon$, where $\epsilon > 0$ is a small number. After $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\gamma}}')'$ is obtained by (3), the estimated risk associated with $\boldsymbol{x}(t)$ is $\widehat{r}_i(t) = \widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i(t)$.

Let $\mu(t) = \mathrm{E}[r_i(t)|T_i \ge t]$ be the mean risk at t among all people at risk in the training dataset, and $\sigma^2(t) = \mathrm{Var}(r_i(t)|T_i \ge t)$ be the corresponding variance. We can use the following local linear kernel smoothing procedure to estimate $\mu(t)$ and $\sigma^2(t)$ (cf., Qiu and Xiang 2014):

$$\widehat{\mu}(t) = \frac{R_0(t)W_{\mu,2}(t) - R_1(t)W_{\mu,1}(t)}{W_{\mu,0}(t)W_{\mu,2}(t) - W_{\mu,1}(t)^2}, \tag{4}$$

$$\widehat{\sigma}^2(t) = \frac{Q_0(t)W_{\sigma,2}(t) - Q_1(t)W_{\sigma,1}(t)}{W_{\sigma,0}(t)W_{\sigma,2}(t) - W_{\sigma,1}(t)^2}, \tag{5}$$

where $\widehat{\epsilon}_i(t_{ij}) = \widehat{r}_i(t_{ij}) - \widehat{\mu}(t_{ij})$, and for $l = 0, 1, 2$,

$$W_{\mu,l}(t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\mu}(t_{ij} - t) \left(\frac{t_{ij} - t}{h_\mu} \right)^l,$$

$$R_l(t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\mu}(t_{ij} - t) \left(\frac{t_{ij} - t}{h_\mu} \right)^l \widehat{r}_i(t_{ij}),$$

$$W_{\sigma,l}(t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\sigma}(t_{ij} - t) \left(\frac{t_{ij} - t}{h_\sigma} \right)^l,$$

$$Q_l(t) = \frac{1}{n} \sum_{i\in R(t)} \sum_{j=1}^{m_i} K_{h_\sigma}(t_{ij} - t) \left(\frac{t_{ij} - t}{h_\sigma} \right)^l \widehat{\epsilon}_i^2(t_{ij}),$$

and $h_\mu, h_\sigma > 0$ are two bandwidths.

When computing $\widehat{\boldsymbol{\theta}}$, $\widehat{\mu}(t)$ and $\widehat{\sigma}^2(t)$ by (3)-(5), the kernel function $K(s)$ can be chosen to be the Epanechnikov kernel function $K(s) = 0.75(1 - s^2)I(|s| \le 1)$, because it has some good properties

(Epanechnikov 1969). The bandwidths h_θ, h_μ and h_σ can be chosen by the cross-validation (CV) procedures described below. To choose h_θ, we suggest using the following leave-one-out CV score modified from the one suggested by Tian et al. (2005) that is based on the martingale residuals:

$$CV_\theta(h_\theta) = \sum_{i=1}^{n} PE_i^\theta(h_\theta),$$

where $PE_i^\theta(h_\theta)$ is the leave-one-out prediction error for the ith subject, defined by

$$PE_i^\theta(h_\theta) = \left(\delta_i - \sum_{\substack{k \neq i,\, \delta_k = 1 \\ T_k \leq T_i}} \frac{\sum_{j=1}^{m_i} K_{h_\theta}(T_k - t_{ij}) \exp\{\widehat{\boldsymbol{\beta}}'_{-i} \boldsymbol{x}_i(t_{ij}) + \widehat{\boldsymbol{\gamma}}'_{-i} \boldsymbol{z}_i\}}{\sum_{d \neq k, d \in R(T_k)} \sum_{j=1}^{m_d} K_{h_\theta}(T_k - t_{dj}) \exp\{\widehat{\boldsymbol{\beta}}'_{-i} \boldsymbol{x}_d(t_{dj}) + \widehat{\boldsymbol{\gamma}}'_{-i} \boldsymbol{z}_d\}} \right)^2, \quad (6)$$

and $\widehat{\boldsymbol{\beta}}_{-i}$ and $\widehat{\boldsymbol{\gamma}}_{-i}$ are the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ when the observed data of the ith subject are excluded from estimation. A detailed derivation of (6) will be given in an appendix. Following Qiu and Xiang (2014), the CV scores for choosing h_μ and h_σ are defined by

$$CV_\mu(h_\mu) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(\widehat{\epsilon}_i(t_{ij}) - \widehat{\mu}_{-i}(t_{ij}) \right)^2,$$
$$CV_\sigma(h_\sigma) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(\widehat{\epsilon}_i^2(t_{ij}) - \widehat{\sigma}^2_{-i}(t_{ij}) \right)^2,$$

where $\widehat{\mu}_{-i}(t)$ and $\widehat{\sigma}^2_{-i}(t)$ denote the leave-one-out estimates of $\mu(t)$ and $\sigma^2(t)$ when the observations of the ith subject are excluded. Then, h_θ, h_μ and h_σ are chosen by minimizing $CV_\theta(h_\theta)$, $CV_\mu(h_\mu)$ and $CV_\sigma(h_\sigma)$, respectively.

Next, we give some statistical properties of the estimators $\widehat{\boldsymbol{\theta}}$, $\widehat{\mu}(t)$ and $\widehat{\sigma}^2(t)$. To this end, let $M_i(t) = \sum_{j=1}^{m_i} I(t_{ij} \leq t)$ denote the point process for observation times $t_{i1}, t_{i2}, \ldots, t_{im_i}$ of the ith subject. Then, we have the following results with the proof given in an appendix.

Theorem 1. Assume that (a) $\{\boldsymbol{x}_i(t), \boldsymbol{z}_i, D_i, C_i, M_i(t)\}_{i=1}^n$ are independent and identically distributed among n people in the training data, (b) within each person, $(\boldsymbol{x}_i(t), \boldsymbol{z}_i, D_i)$, C_i and $M_i(t)$ are independent, (c) $\boldsymbol{x}_i(t)$ are bounded, predictable and left-adapted q-dimensional random processes in $[0, \mathcal{T}]$, $\boldsymbol{z}_i$ are bounded random variables, and $\mathrm{P}(C_i \geq \mathcal{T}) \geq 0$, (d) $\lambda_0(t), \mu(t), \sigma^2(t) \in C^2[0, \mathcal{T}]$, (e) there exists some positive function $\phi(t) \in C^2[0, \mathcal{T}]$ such that for any $1 \leq i \leq n$ and any $0 \leq t_0 < t_1 \leq T_i$, $\int_{t_0}^{t_1} \phi(s)\, ds = \mathrm{E}[\int_{t_0}^{t_1} dM_i(t)] = \mathrm{E}[\#\{t_{ij} : t_{ij} \in [t_0, t_1], 1 \leq j \leq m_i\}]$, (f) let $f_T(t)$ be the marginal probability density of T_i, and $f_x(\boldsymbol{x}, \boldsymbol{z}, s|t)$ be the conditional probability density of $(\boldsymbol{x}_i(s), \boldsymbol{z}_i)$ among the at-risk population at t, then $\frac{\partial}{\partial t} f_T(t)$ is continuous in t, and $\frac{\partial^2}{\partial s^2} f_x(\boldsymbol{x}, \boldsymbol{z}, s|t)$ is continuous in $\boldsymbol{x}$, $\boldsymbol{z}$, s and t, (g) the kernel function $K(s) \in C^2[-1, 1]$ is continuous and symmetric

about 0 with support $[-1, 1]$, and (h) the bandwidths h_θ, h_μ and h_σ satisfy the conditions that $h_\theta, h_\mu, h_\sigma \to 0$, and $\sqrt{n}h_\theta, \sqrt{n}h_\mu, \sqrt{n}h_\sigma \to \infty$, as $n \to \infty$. Then, we have

(i) $\lim_{n\to\infty} \widehat{\boldsymbol{\theta}} \stackrel{p}{=} \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0')'$ denotes the true value of $\boldsymbol{\theta}$,

(ii) $\lim_{n\to\infty} \sup_{t\in[h_\mu, \mathcal{T}-h_\mu]} |\widehat{\mu}(t) - \mu(t)| \stackrel{p}{=} 0$, and

(iii) $\lim_{n\to\infty} \sup_{t\in[h_\sigma, \mathcal{T}-h_\sigma]} |\widehat{\sigma}^2(t) - \sigma^2(t)| \stackrel{p}{=} 0$.

At the end of this part, we would like to point out that model (1) assumes that the impact of the risk factors in $\boldsymbol{x}_i(t)$ and the covariates in $\boldsymbol{z}_i$ does not change over time. If it is believed that such impact may change over time, then the following time-varying coefficient model might be more appropriate (cf. Zucker and Karr 1990, Hastie and Tibshirani 1993):

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}'(t)\boldsymbol{x}_i(t) + \boldsymbol{\gamma}'(t)\boldsymbol{z}_i), \qquad \text{for } t \in [0, \mathcal{T}], \tag{7}$$

where $\boldsymbol{\beta}(t)$ and $\boldsymbol{\gamma}(t)$ are time-varying coefficients. Let $\boldsymbol{\theta}(t) = (\boldsymbol{\beta}'(t), \boldsymbol{\gamma}'(t))'$. There exists some discussion about estimation of model (7) using the generalized additive modelling (e.g. Hastie and Tibshirani 1993), smoothing splines (e.g. Zucker and Karr 1990), and local likelihood estimation (e.g. Cai and Sun 2003) in cases when both the covariates and the risk factors do not change over time or when they are continuously observed. To follow the idea in estimating the model (1) above, we propose a kernel estimation procedure to deal with irregularly spaced observation times, which can also accommodate time-varying coefficients. First, define the following local log-likelihood function

$$\begin{aligned} l(\boldsymbol{\theta}, t) &= l(\boldsymbol{\beta}, \boldsymbol{\gamma}, t) \\ &= \sum_{i:\delta_i=1} K_{h_\theta}(T_i - t) \left\{ \boldsymbol{\beta}'\boldsymbol{x}_i(T_i) + \boldsymbol{\gamma}'\boldsymbol{z}_i - \log \left[\sum_{l\in R(T_i)} \sum_{j=1}^{m_l} K_{h_\theta}(T_i - t_{lj}) \exp\left(\boldsymbol{\beta}'\boldsymbol{x}_l(t_{lj}) + \boldsymbol{\gamma}'\boldsymbol{z}_l\right) \right] \right\}, \end{aligned}$$

where the related quantities are the same as those in (2). Then, a Newton-Raphson algorithm can be used to minimize the above function. The resulting estimate of $\boldsymbol{\theta}(t)$ is defined as

$$\widehat{\boldsymbol{\theta}}(t) = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, t),$$

where $\widehat{\boldsymbol{\theta}}(t) = (\widehat{\boldsymbol{\beta}}'(t), \widehat{\boldsymbol{\gamma}}'(t))'$. The risk function of interest becomes $r_i(t) = \boldsymbol{\beta}'(t)\boldsymbol{x}_i(t)$, and the estimated risk function is thus $\widehat{r}_i(t) = \widehat{\boldsymbol{\beta}}'(t)\boldsymbol{x}_i(t)$. After these modifications, the estimates of $\mu(t)$ and $\sigma^2(t)$ can still be computed by Equations (4) and (5), and the the online risk monitoring scheme discussed in Section 2.2 below can still apply.

2.2 Online risk monitoring

Next, we discuss how to online monitor the disease risk of a new person so that the disease in question can be detected early. Assume that the disease risk factors in $\boldsymbol{x}(t)$ of the person are observed at times $\{t_1^*, t_2^*, \ldots\}$. Then, the disease risks at these time points can be estimated by $\widehat{r}(t_j^*) = \widehat{\boldsymbol{\beta}}'\boldsymbol{x}(t_j^*)$, for $j \geq 1$, where $\widehat{\boldsymbol{\beta}}$ is computed in advance by (3) from the training data. The standardized risks are defined as

$$\widehat{e}(t_j^*) = \frac{\widehat{r}(t_j^*) - \widehat{\mu}(t_j^*)}{\widehat{\sigma}(t_j^*)}, \qquad \text{for } j \geq 1,$$

where $\widehat{\mu}(t)$ and $\widehat{\sigma}^2(t)$ are computed in advance from the training data by (4) and (5), respectively. In the above expression, $\widehat{\mu}(t)$ and $\widehat{\sigma}(t)$ describe the estimated longitudinal pattern of the disease risk over time for the population that the training dataset represents, and $\widehat{e}(t)$ is the standardized version of the estimated risk of the new person under monitoring, after comparing his/her disease risk with the mean disease risk of people in the population that is represented by the training data. So, a larger value of $\widehat{e}(t)$ implies a larger chance of the disease.

To online monitor the standardized risks $\{\widehat{e}(t_j^*), j \geq 1\}$, we need to take into account the fact that the observation times $\{t_1^*, t_2^*, \ldots\}$ are often unequally spaced. The conventional control charts in the literature are usually for cases with equally spaced observation times only (cf., Qiu 2014). To overcome that difficulty and accommodate unequally spaced observation times, we suggest using a modified exponentially weighted moving average (EWMA) chart described below. Let $\omega > 0$ be a basic time unit such that all unequally spaced observation times are its integer multiples. Then, $\{n_j^* = t_j^*/\omega, j = 1, 2, \ldots\}$ are the observation times in the basic time unit. The modified EWMA chart is defined by

$$\begin{aligned} E_1 &= V(t_1^*)\widehat{e}(t_1^*) \\ E_j &= (1 - V(t_j^*))E_{j-1} + V(t_j^*)\widehat{e}(t_j^*), \qquad \text{for } j \geq 2, \end{aligned} \tag{8}$$

where $V(t_1^*) = 1 - (1-\lambda)^{\bar{\Delta}}$, $V(t_j^*) = V(t_{j-1}^*)/[(1-\lambda)^{\Delta_j} + V(t_{j-1}^*)]$, for $j \geq 2$, $\bar{\Delta}$ is the mean of $\Delta_j = n_j^* - n_{j-1}^*$ which can be estimated from the training data, and $\lambda \in [0, 1)$ is a weighting parameter. It can be checked that E_j in (8) is a weighted average of $\widehat{e}(t_j^*), \widehat{e}(t_{j-1}^*), \ldots, \widehat{e}(t_2^*), \widehat{e}(t_1^*)$ and the weights are proportional to $1, (1-\lambda)^{n_j^* - n_{j-1}^*}, \ldots, (1-\lambda)^{n_j^* - n_2^*}, (1-\lambda)^{n_j^* - n_1^*}$, respectively. Therefore, the weights are controlled by λ and by how far away a previous observation time is from the current observation time in the basic time unit, which is recommended by Qiu et al. (2018)

and Wright (1986). So, unequal observation times have been accommodated properly in E_j. In cases when observation times are equally spaced, the chart (8) becomes the conventional EWMA chart (i.e., $V(t_j^*)$ is a constant in (8)). The chart (8) gives a signal when

$$E_j > \rho, \tag{9}$$

where $\rho > 0$ is a control limit.

The performance of a control chart is usually measured by the IC average run length (ARL), defined as the average number of observations from the start of online monitoring to the signal of the chart, and the OC ARL, defined as the average number of observations from the occurrence of a shift to the signal of the chart. However, in cases when observation times are unequally spaced, the IC and OC ARLs would not be good performance measures any more. In such cases, more reasonable performance measures are the IC average time to signal (ATS), denoted as ATS_0, and the OC ATS, denoted as ATS_1. Namely, the average number of observations is replaced by the average length of time in the related performance measures. Then, ATS_0 is usually fixed in advance, and the chart performs better for detecting a given shift if its ATS_1 value is smaller. In the chart (8)-(9), the weighting parameter λ and the control limit ρ need to be chosen in advance. To this end, λ is usually pre-specified and ρ is chosen to achieve a given value of ATS_0. It has been demonstrated in the literature that a large value of λ is good for detecting a large shift and a small value of λ is good for detecting a small shift (e.g., Qiu 2014, Chapter 5). To choose ρ, we can use the block bootstrap procedure described below, where the block bootstrap is considered because the within-subject observations are usually correlated. First, the training data with n subjects are divided into two halves, with all diseased and non-diseased people evenly splitted, respectively. Second, the first half of the training data is used for estimating $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\mu(t)$ and $\sigma^2(t)$, as discussed in Section 2.1 (cf., (3)-(5)). The second half of the training data is used for determining the value of ρ in the following several steps. i) A non-diseased person is randomly selected from the second half of the training data, and the chart (8)-(9) with a given value of ρ is applied to this person for online monitoring. If there is a signal, then the time from the beginning of online monitoring to the signal time is recorded. ii) This process is then repeated for B times, and all times to signals are averaged. The averaged time to signals is used to approximate $ATS_0(\rho)$ value. If $ATS_0(\rho)$ is smaller than the pre-specified ATS_0 value, then increase the value of ρ. Otherwise, decrease the value of ρ. iii) The above two steps are executed iteratively until the pre-specified ATS_0 value is reached. The searched value of ρ is then defined as the one used when the iterative procedure

stops. In this searching process, a bisection search or other numerical algorithms can be used to speed up the search. See a related discussion in Section 4.2 of Qiu (2014).

3 Simulation Study

In this section, we will present some numerical results about the proposed risk estimation and online monitoring methods discussed in the previous section. The results are presented after a brief description about the simulation setup.

The study time period is assumed to be $[0, 1]$, which is discretized into 1000 basic time units $\mathbb{T} = \{i\omega, i = 1, 2, \ldots, 1000\}$, where $\omega = 0.001$. We first considered cases when there are 3 risk factors in $\boldsymbol{x}_i(t)$ and no time-independent covariates $\boldsymbol{z}_i$ in the model (1). The risk vector $\boldsymbol{x}_i(t)$ is assumed to follow five different models described below, and the true disease occurrence times D_i are generated from the Cox proportional hazards model (1) with the regression coefficients being $\boldsymbol{\beta} = (0.5, 0.3, 0.2)'$. Here, the method in Hendry (2014) is used for generating survival outcomes from a Cox proportional hazards model with time-dependent covariates. The risk factors in $\boldsymbol{x}_i(t)$ are assumed to be observed at $t_{i1}, t_{i2}, \ldots, t_{im_i}$, where $t_{i(j+1)} = t_{ij} + (N_{ij} + 1)\omega$ and $\{N_{ij}, j = 1, 2, \ldots, m_i, i = 1, 2, \ldots, n\}$ are i.i.d. and Poisson(20) distributed (note: the quantity $N_{ij}+1$ is used here so that $t_{i(j+1)} - t_{ij}$ cannot be 0). The censoring times C_i are generated from the Weibull(5,1) distribution, where 5 is its shape parameter value and 1 is the scale parameter value.

In the simulation study, we consider three different types of longitudinal models for the risk vector $\boldsymbol{x}_i(t)$ to study the performance of the proposed method in different longitudinal setups. First, we consider the following mixed-effects model for generating the three IC risk factors in $\boldsymbol{x}_i(t) = (x_{i1}(t), x_{i2}(t), x_{i3}(t))'$:

$$\begin{aligned} x_{i1}(t) &= 2t + \zeta_{i1} + (1 - 0.5t^2)\epsilon_{i1}(t) \\ x_{i2}(t) &= -\sin(2\pi t) + \zeta_{i2} + (1 + 0.2\sin(3\pi t))\epsilon_{i2}(t) \\ x_{i3}(t) &= \log(t + 1) + \zeta_{i3} + (1 + 0.2\cos(3\pi t))\epsilon_{i3}(t), \end{aligned}$$

where the random-effect terms $(\zeta_{i1}, \zeta_{i2}, \zeta_{i3})'$ follow the distribution $\mathrm{N}(\mathbf{0}, \tau^2\,\Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & 0.25 & 0 \\ 0.25 & 1 & 0.25 \\ 0 & 0.25 & 1 \end{pmatrix},$$

τ^2 is a parameter, and $\{\epsilon_{ik}(t) : i = 1, \ldots, n,\ k = 1, 2, 3,\ t \in \mathbb{T}\}$ are i.i.d. and generated from the standard normal distribution. In the above model, the random-effect terms are considered for accommodating the within-subject data correlation. Different functions are included as multipliers of the error terms for accommodating time-varying variability of the pure measurement error. To allow different magnitudes of within-subject variability, we set τ to be 0.2, 0.3, or 0.4. A constant baseline hazard $\lambda_0(t) = 0.2$ is assumed. The corresponding mixed-effects models are labeled as models Ia, Ib and Ic, respectively, when $\tau = 0.2$, 0.3 and 0.4. By considering these three models together, we can study the impact of the within-subject variability on the performance of the proposed method.

Next, we consider the following time-series model for generating $\boldsymbol{x}_i(t)$:

$$\begin{aligned}
x_{i1}(t) = &\quad \sin(\pi t)/2 + \zeta_{i1}(t) + (1 - 0.25\cos(3\pi t))\xi_{i1}(t) \\
x_{i2}(t) = &\quad \cos(\pi t)/2 + \zeta_{i2}(t) + (1 + 0.25\sin(3\pi t))\xi_{i2}(t) \\
x_{i3}(t) = &-\sin(\pi t)/2 + \zeta_{i3}(t) + (1 + 0.25\cos(3\pi t))\xi_{i3}(t),
\end{aligned}$$

where for each $k = 1, 2,$ or 3, $\{\zeta_{ik}(t), i = 1, 2, \ldots, n\}$ are i.i.d. and from the Brownian motions with the starting values being $\{\zeta_{ik}(0), i = 1, 2, \ldots, n\}$ that all follow the N(0,0.3^2) distribution and $\zeta_{ik}(s) - \zeta_{ik}(t)$ follow the N(0,$|t - s|$) distributions, for any $s, t \in \mathbb{T}$, and $\{\xi_{ik}(t), i = 1, 2, \ldots, n\}$ are generated from the AR(1) model

$$\xi_{ik}(t) = 0.9\xi_{ik}(t - \omega) + 0.5\epsilon_{ik}(t), \text{ for } t \in \mathbb{T},\ i = 1, \ldots, n,\ k = 1, 2, 3,$$

in which $\{\epsilon_{ik}(t), i = 1, \ldots, n,\ k = 1, 2, 3,\ t \in \mathbb{T}\}$ are i.i.d. and standard normally distributed. In the above model, the three components of $\boldsymbol{x}_i(t)$ are assumed independent. Besides the temporal trends specified in the first terms of the three equations, the Brownian motions are considered to add randomness to the temporal paths of the three components of $\boldsymbol{x}_i(t)$. A time-varying baseline hazard $\lambda_0(t) = t/2$ is assumed. This model is labeled as model II.

Finally, we consider the following non-Gaussian model for generating $\boldsymbol{x}_i(t)$:

$$\begin{aligned}
x_{i1}(t) = &\quad 0.3\zeta_{i1} + (1 - 0.5t^2)\epsilon_{i1}(t) \\
x_{i2}(t) = &-\sin(2\pi t)/2 + 0.3\zeta_{i2} + (1 + 0.25\sin(3\pi t))\epsilon_{i2}(t) \\
x_{i3}(t) = &-\cos(2\pi t)/2 + 0.2\eta_{i3}(t) + (1 + 0.25\cos(3\pi t))\epsilon_{i3}(t),
\end{aligned}$$

where $\{\epsilon_{i1}(t), \epsilon_{i2}(t), i = 1, \ldots, n, t \in \mathbb{T}\}$ and ζ_{i1} are i.i.d. and t_{10} distributed, $\{\epsilon_{i3}(t), t \in \mathbb{T}\}$ are i.i.d. and standard normal, $\{\zeta_{i2}\}$ are i.i.d. and Gamma(2,4) distributed, and $\eta_{i3}(t)$ are i.i.d. and

Poisson(1) distributed. A constant baseline hazard $\lambda_0(t) = 0.2$ is assumed. This model is labeled as model III.

In all simulation studies in this section, the sample size of the training data is $n = 1000$. For online monitoring, the ATS_0 and ATS_1 values are computed based on 10,000 replicated simulations. When evaluating the OC performance of the proposed online monitoring procedure, the OC risk factors are assumed to be

$$\boldsymbol{x}^{\delta}(t) = \boldsymbol{x}(t) + \delta\boldsymbol{\beta},$$

where $\boldsymbol{x}(t)$ is the IC risk factors specified in the above five models, and $\delta\boldsymbol{\beta}$ is the vector of step shift sizes in the three components.

We first report some numerical results about the risk estimation in model (1). Under the five different models described above, results about estimates of $\boldsymbol{\beta}$ based on 1,000 replicated simulations are presented in Table 1. In the model estimation procedure, bandwidths h_θ, h_μ, and h_σ are chosen by CV, as discussed in Section 2.1. From the table, it can be seen that estimates of $\boldsymbol{\beta}$ are close to the true values in all cases considered. The selected bandwidths by CV are presented in Table 2, together with the optimal bandwidths obtained by minimizing the MSE (for h_θ) and MISE (for h_μ and h_σ) values. From the table, it can be seen that the selected bandwidths by CV are generally smaller than their optimal values, but the corresponding MSE or MISE values when the bandwidths are chosen by CV are close to those when they are chosen optimally. To study this phenomenon, Figure 1 presents the MSE or MISE values when their related bandwidths change around their optimal values in all cases considered. It can be seen that the MSE or MISE curves are quite flat around the optimal values of the bandwidths (denoted by the vertical lines and dark dots in the plots), and thus the MSE or MISE values at the selected bandwidths by CV (denoted by dark crosses) are close to their values at the optimal bandwidths. Therefore, the selected bandwidths by CV perform well in that sense.

Next, we evaluate the numerical performance of the proposed online monitoring chart (7)-(8). To this end, the DySS method suggested in Qiu and Xiang (2015) is used as a standard. As discussed in Section 1, the DySS method in Qiu and Xiang (2015), denoted as DySS, monitors all risk factors in $\boldsymbol{x}(t)$ directly for disease early detection. As a comparison, the proposed method in the current paper, denoted as New, monitors the estimated risk $\widehat{r}(t) = \widehat{\boldsymbol{\beta}}'\boldsymbol{x}(t)$ of a given person at different time points. In the simulation study, ATS_0 is fixed at 370, and the weighting parameter

Table 1: Estimated values of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ and their standard errors (in parentheses) under five different models of $\boldsymbol{x}_i(t)$.

Model	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
(Ia)	0.4878 (0.0028)	0.2843 (0.0021)	0.1957 (0.0023)
(Ib)	0.4887 (0.0027)	0.2900 (0.0021)	0.1925 (0.0022)
(Ic)	0.4887 (0.0027)	0.2882 (0.0020)	0.1967 (0.0022)
(II)	0.5020 (0.0023)	0.2911 (0.0023)	0.1958 (0.0023)
(III)	0.4969 (0.0025)	0.2969 (0.0022)	0.1961 (0.0023)
Truth	$\beta_1 = 0.5$	$\beta_2 = 0.3$	$\beta_3 = 0.2$

Table 2: Optimal values of h_θ, h_μ and h_σ, their selected values by the cross-validation procedures, and the corresponding MSE and/or MISE values.

Model		h_θ	MSE	h_μ	MISE	h_σ	MISE
(Ia)	Optimal	0.245	0.0182	0.085	0.2276	0.130	0.0891
	Selected	0.170	0.0183	0.045	0.3018	0.070	0.1203
(Ib)	Optimal	0.240	0.0169	0.085	0.2960	0.135	0.1085
	Selected	0.170	0.0170	0.045	0.3703	0.070	0.1456
(Ic)	Optimal	0.245	0.0160	0.085	0.3902	0.140	0.1456
	Selected	0.170	0.0161	0.045	0.4650	0.070	0.1917
(II)	Optimal	0.230	0.0149	0.210	0.3132	0.095	0.3971
	Selected	0.235	0.0149	0.105	0.3808	0.050	0.5451
(III)	Optimal	0.310	0.0150	0.110	0.2840	0.125	0.2079
	Selected	0.155	0.0152	0.055	0.3680	0.065	0.2915

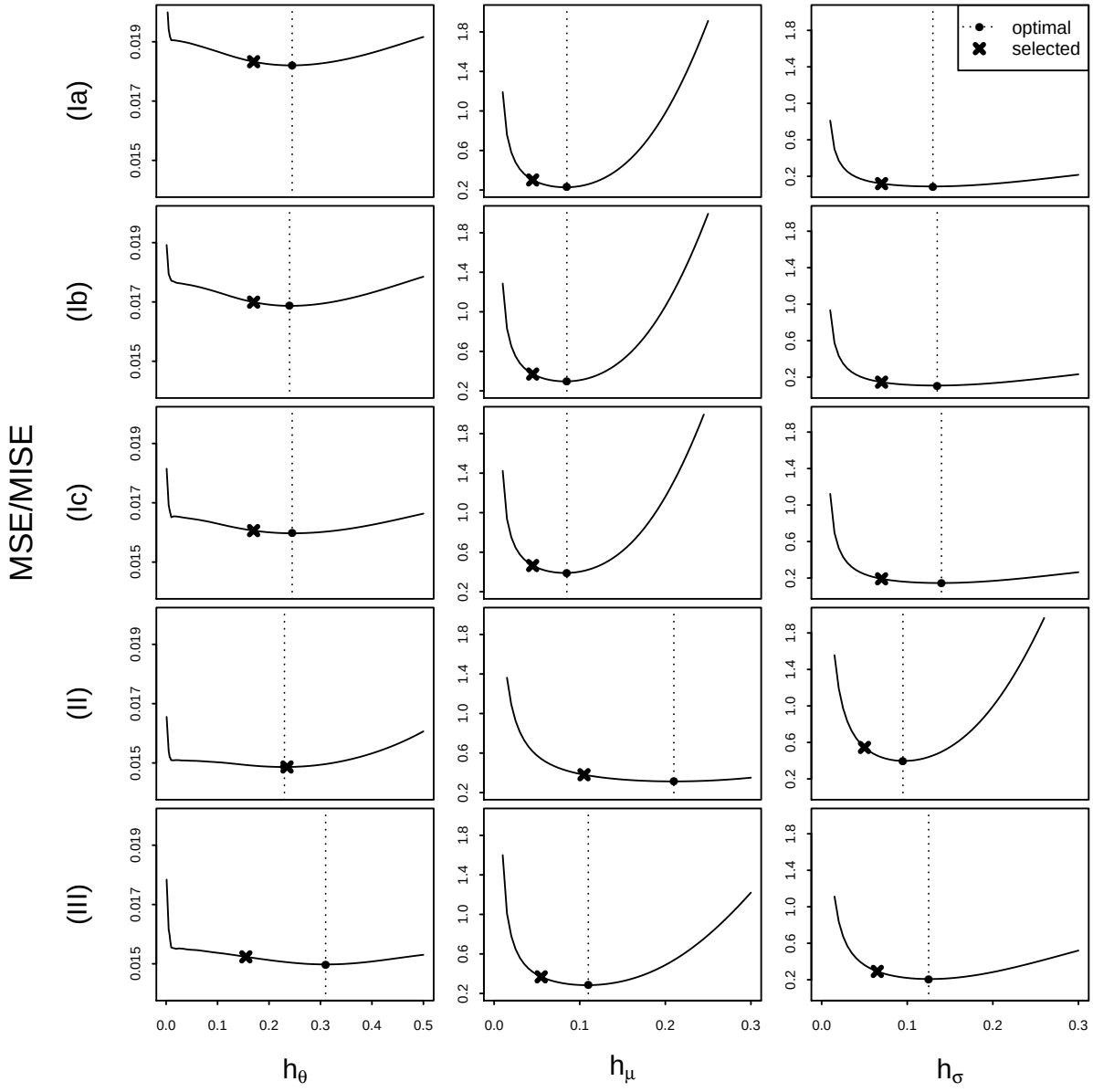

Figure 1: MSE or MISE values when their related bandwidths change around their optimal values (vertical lines and dark dots) in cases considered in Table 2. Dark crosses denote the selected bandwidths by CV.

λ is chosen to be 0.01, 0.02, 0.03 and 0.05. The shift size parameter δ changes between 0 and 5. The results of computed ATS_0 and ATS_1 values are presented in Figure 2, along with the 95% confidence intervals (CIs) of the ATS_1 values of New. From the plots in the figure, it can be seen that (i) the calculated ATS_0 values of the two methods are all close to the nominal ATS_0 value, and (ii) the calculated ATS_1 values of New are significantly smaller than those of DySS in all cases considered. Therefore, the proposed method New improves disease early detection by DySS in this example. If we compare the ATS_1 values of New in cases of (Ia)-(Ic) carefully, it can be seen that New performs slightly worse when the within-subject variability gets larger (i.e., when τ gets larger from Model (Ia) to Model (Ic)).

Next, we consider cases when there are time-independent covariates $\boldsymbol{z}_i$ in the model (1). First, in model Ia described above, we add a two-dimensional time-independent covariate vector $\boldsymbol{z}_i$ generated from the distribution $\mathrm{N}(\mathbf{0}, \sigma_{\boldsymbol{z}}^2 \boldsymbol{\Sigma}_{\boldsymbol{z}})$, where $\sigma_{\boldsymbol{z}} = 0.5$, $\boldsymbol{\Sigma}_{\boldsymbol{z}} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, and $\{\boldsymbol{z}_i, i = 1, 2, \ldots, n\}$ are independent of each other. The resulting model is labelled as Ia*. Second, we add a two-dimensional time-independent covariate vector $\boldsymbol{z}_i$ to model II, where observations of the first covariate z_{i1} are generated from the discrete distribution Binomial$(0.5, 1)$, observations of the second covariate z_{i2} are generated from $\mathrm{N}(0, 0.25)$, and $\{\boldsymbol{z}_i = (z_{i1}, z_{i2})', i = 1, 2, \ldots, n\}$ are independent of each other. The resulting model is labelled as II*. The true regression coefficients of $\boldsymbol{\gamma}$ in both models Ia* and II* are assumed to be $(0.2, 0.2)'$, and other settings remain the same as those in models Ia and II. The numerical results on estimates of $\boldsymbol{\theta}$ based on 1,000 replicated simulations are presented in Table 3, where the bandwidths h_θ, h_μ, and h_σ are chosen by CV, as discussed in Section 2.1. From the table, it can be seen that estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are all close to their true values in both scenarios considered here. Next, we evaluate the numerical performance of the proposed online monitoring chart New, in comparing with DySS. In the same setup as that of Figure 2, the results of computed ATS_0 and ATS_1 values are presented in Figure 3. It can be seen from the figure that similar conclusions can be made here to those from Figure 2.

At the end of this section, some cases with the time-varying model (7) are considered. Again, the models Ia and II are modified, and the resulting models are labelled as Ia** and II**, respectively. The true regression coefficients are assumed to be $\boldsymbol{\theta}(t) = \boldsymbol{\beta}(t) = (0.5 + 0.5t, 0.2 + 2(t - 0.5)^2, 0.6 - (t - 0.3)^2)'$ in model Ia**, and $\boldsymbol{\theta}(t) = \boldsymbol{\beta}(t) = (0.5 + 0.5\sin(t), 0.7 - 0.5t^2, 0.6 - 0.3\log(1 + 2t))'$ in model II**. In both models, a constant baseline hazard $h_0(t) = 0.2$ is assumed, and all censoring

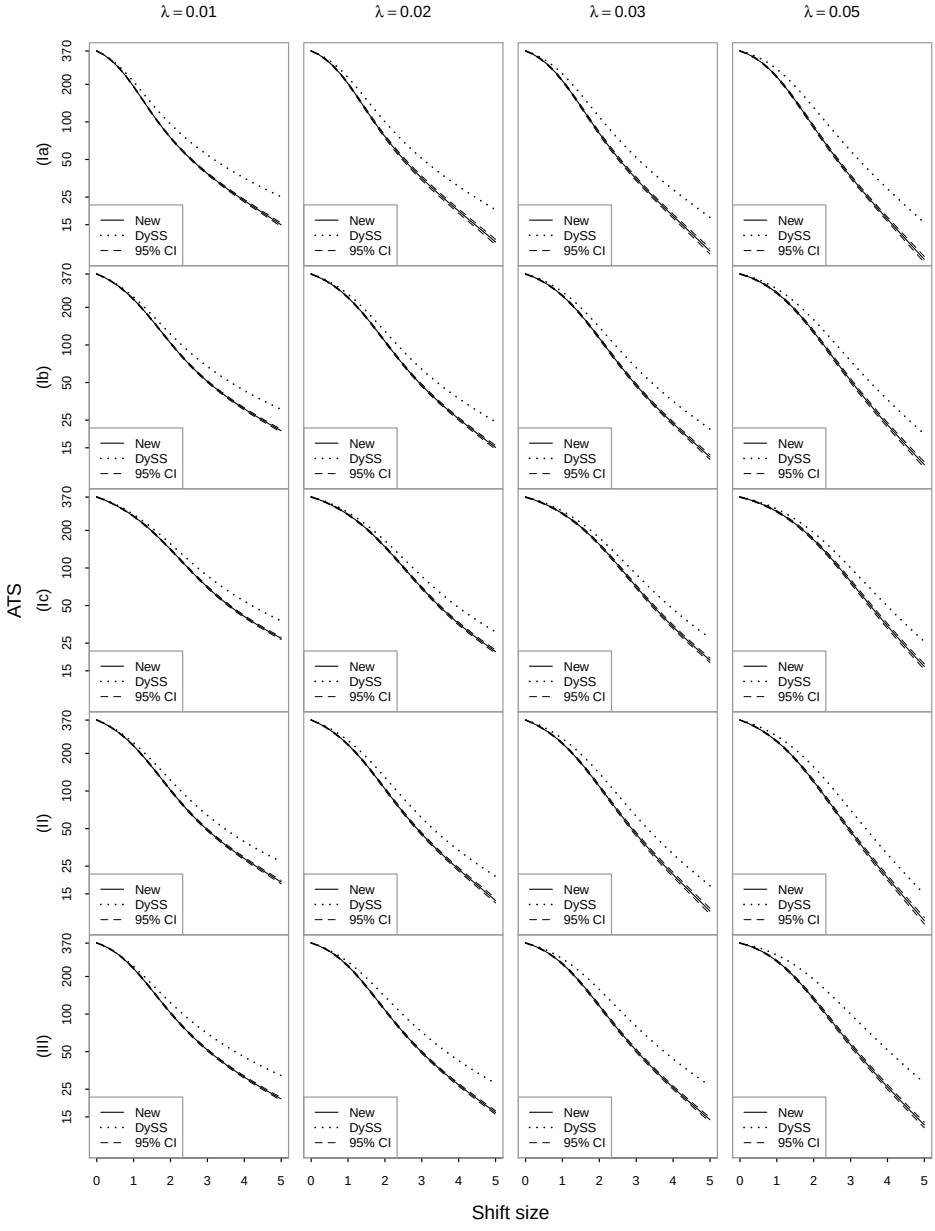

Figure 2: Calculated ATS_0 and ATS_1 values of the proposed method New and the DySS method in Qiu and Xiang (2015) when the weighting parameter λ and the shift size parameter δ change among several values. The dashed lines denote the 95% confidence intervals (CIs) of the ATS_1 values of New. In this example, the nominal ATS_0 is chosen to be 370.

Table 3: Estimated values of $\boldsymbol{\theta} = (\beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2)'$ and their standard errors (in parentheses) under two different models Ia* and II*.

Model	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$
(Ia*)	0.4941 (0.0028)	0.2898 (0.0022)	0.1964 (0.0023)	0.2082 (0.0056)	0.2010 (0.0055)
(II*)	0.5037 (0.0022)	0.2914 (0.0021)	0.1986 (0.0022)	0.2028 (0.0055)	0.1935 (0.0053)
Truth	$\beta_1 = 0.5$	$\beta_2 = 0.3$	$\beta_3 = 0.2$	$\gamma_1 = 0.2$	$\gamma_2 = 0.2$

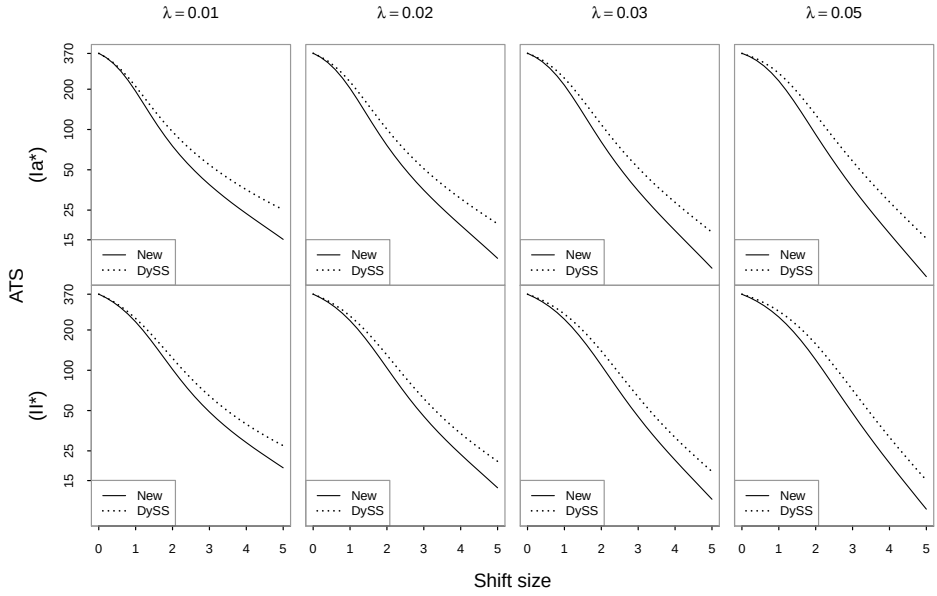

Figure 3: Calculated ATS_0 and ATS_1 values of the proposed method New and the DySS method in Qiu and Xiang (2015) under the models Ia* and II* and other setups are the same as those in Figure 2.

times C_i are set to be 1 to create more events. Other setups are the same as those in the example of Table 1. In cases when the bandwidths are all fixed at 0.1, the point estimates and the 90% pointwise confidence intervals for $\boldsymbol{\beta}(t)$ based on 1,000 replicated simulations are shown in Figure 4. It can be seen from the figure that the point estimates are reasonably good, although the pointwise confidence intervals are quite wide at certain places, due mainly to limited sample sizes. We also compared our proposed method New with DySS in this example under the same setup as that of Figure 2. The results are shown in Figure 5, from which it can be seen that New still outperforms DySS in terms of ATS_1.

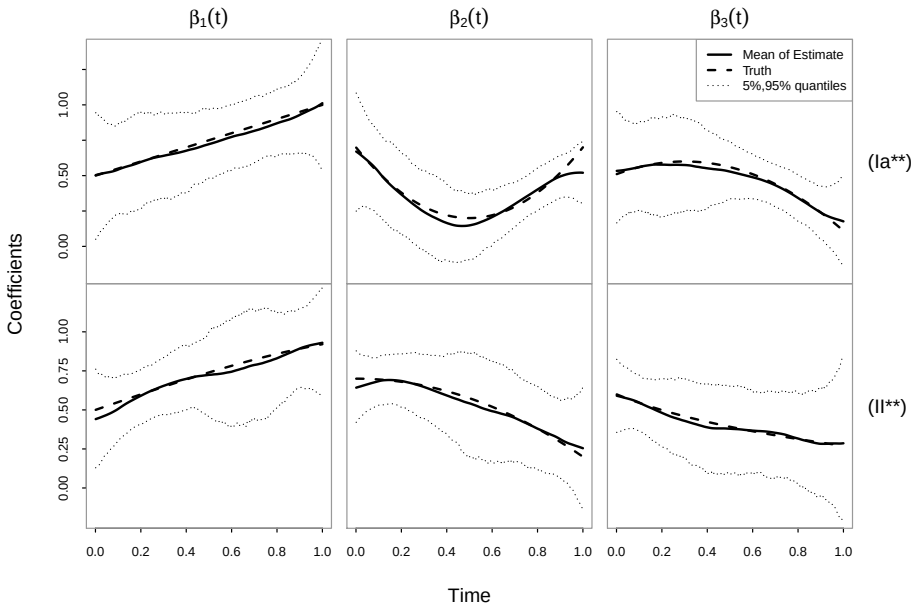

Figure 4: Point estimates (solid curves) and 90% pointwise confidence intervals (dotted curves) of $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t), \beta_3(t))'$. The dark-dashed curves in the plots are the true functions of $\boldsymbol{\beta}(t)$.

4 Application to the Stroke Data

In this section, we apply the proposed method to a dataset obtained from the SHARe Framingham Heart Study. In the data, there were a total of 1,055 people involved, among which 27 people experienced stroke during the study period and the remaining 1,028 people did not. Each person was followed 7 times, and three medical indices, including the systolic blood pressure (mmHg), diastolic blood pressure (mmHg), and total cholesterol level (mg/100ml) were recorded at each

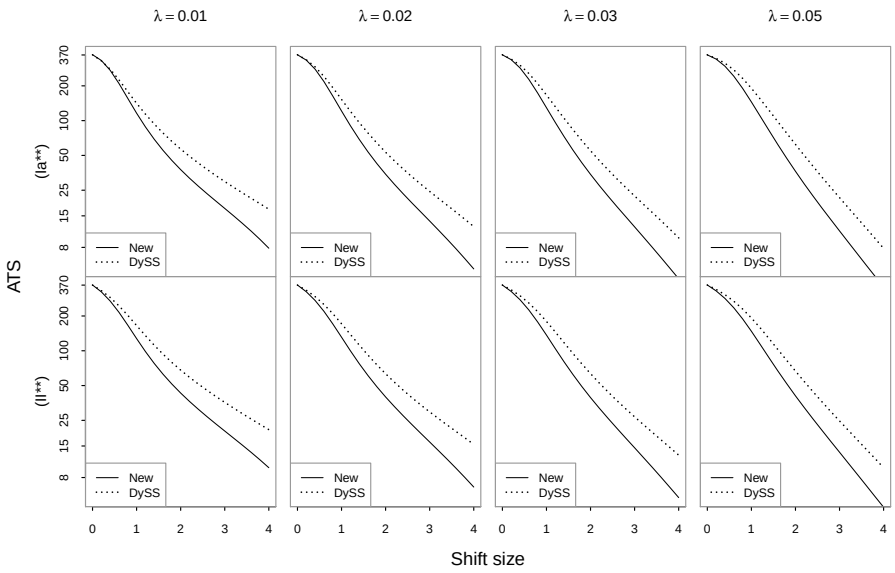

Figure 5: Calculated ATS_0 and ATS_1 values of the proposed method New and the DySS method in Qiu and Xiang (2015) under the models Ia** and II** and other setups are the same as those in Figure 2.

time. The data are shown in Figure 6. To apply the proposed method, we randomly select two-thirds of the stroke people and two-thirds of the non-stroke people as the training data and the remaining are used as the test data for online monitoring. In the entire dataset, all observation times are between 16 and 83 years old, the mean interval length between two consecutive observation times is $\bar{\Delta} = 4.37$ (in years).

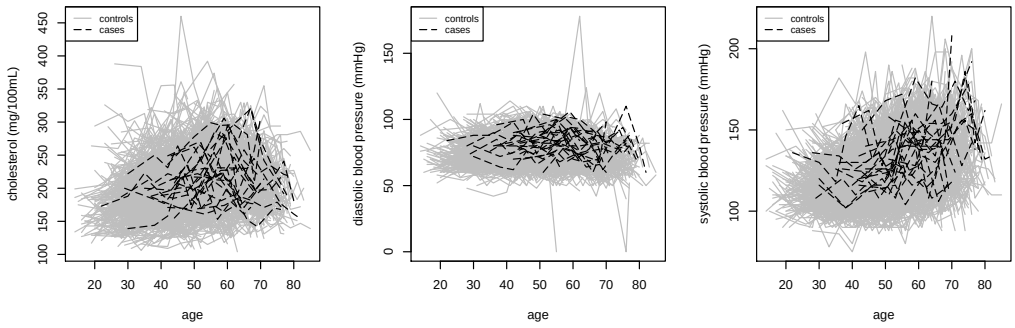

Figure 6: Three risk factors of the stroke data: systolic blood pressure (mmHg), diastolic blood pressure (mmHg), and total cholesterol level (mg/100ml). The solid gray lines are the longitudinal observations of the non-stroke people, and the black dashed lines are the longitudinal observations of the stroke people.

We first check the proportional hazards assumption in model (1) by applying the goodness-of-fit test discussed in Park and Qiu (2014) to the training data. To this end, all subjects in the training data set are partitioned into 30 groups based on the estimated baseline risks, i.e., the estimated risk at the first observation time point $\widehat{r}_i(t_{i1})$, for $i = 1, 2, \ldots, n$. The number of groups, 30, is chosen to be a simple integer that is close to $2n^{2/5} = 27.546$, as suggested in Park and Qiu (2014). The groups are determined by the $(k/30)$-th quantiles of the estimated baseline risks, for $k = 1, \ldots, 30$. Then, the test statistic is defined by

$$T = (H_1, H_2, \ldots, H_{29})\widehat{\Sigma}_H^{-1}(H_1, H_2, \ldots, H_{29})',$$

where H_k denotes the sum of all approximate martingale residuals within the kth group, for $k = 1, 2, \ldots, 29$, and $\widehat{\Sigma}_H$ is the estimated covariance matrix of $(H_1, H_2, \ldots, H_{29})'$. The 30th group is not included above because the matrix Σ_H would be singular otherwise. Details about the calculation of $\widehat{\Sigma}_H^{-1}$ can be found in Grønnesby and Borgan (1996). Here, the approximate martingale residuals and $\widehat{\Sigma}_H^{-1}$ are all adjusted using the proposed kernel smoothing technique, as discussed in Section 2.1 about the equation (2). When the proportional hazards assumption holds, the test statistic T should be asymptotically χ^2-distributed with 29 degrees of freedom. The resulting p-value for the goodness-of-fit test is 0.621, implying that no significance evidence is found for the violation of the proportional hazards assumption.

The training data are then used for estimating the survival model (1), and the estimated regression coefficients are $\widehat{\boldsymbol{\beta}} = (0.0001, 0.0047, 0.0269)'$. Consequently, the estimates of $\mu(t)$ and $\sigma^2(t)$ can be obtained by (4) and (5). In the control chart (8)-(9), the weighting parameter λ is chosen to be 0.1. The control charts for monitoring the 9 people in the test data when ATS_0 is chosen to be 10 or 15 are shown in Figure 7. From the plots in the figure, it can be seen that when $ATS_0 = 10$, the proposed method gives signals to 9 out of 9 stroke people and to 187 out of 342 non-stroke people in the test data. When $ATS_0 = 15$, it gives signals to 7 out of 9 stroke people and to 125 out of 342 people in the test data. As a comparison, the DySS method suggested by Qiu and Xiang (2015) in the same setup gives signals to all 9 stroke people and to 202 non-stroke people in the test data when $ATS_0 = 10$. When $ATS_0 = 15$, it only gives signals to 2 stroke people and to 52 non-stroke people in the test data. The above results are summarized in Table 4. From the table, it can be seen that New performs better than DySS when $ATS_0 = 10$, because their numbers of signals to stroke patients are the same, but New gives less signals to non-stroke patients. In the case when $ATS_0 = 15$, it is hard to compare their performance by just looking at the numbers

in the table, because DySS gives less signals to both stroke and non-stroke patients. To overcome this difficulty, we first compute the odds ratios of both New and DySS. The odds ratio of New is $7 * (342 - 125)/[(9 - 7) * 125] = 6.076$ and its 95% confidence interval is $(1.243, 29.700)$, which indicates that the odds ratio of New is significantly larger than 1. On the other hand, the odds ratio of DySS is $2 * (342 - 52)/[(9 - 2) * 52] = 1.593$ and its 95% confidence interval is $(0.322, 7.884)$, which indicates that the odds ratio of DySS is not significantly different from 1. By comparing the odds ratios of the two methods, it seems that New has a larger odds ratio than DySS. To further compare their performance in the case when $ATS_0 = 15$, we adjust the weighting parameter in the EWMA chart of the DySS method, so that the modified DySS method gives the same number of signals to stroke patients as that of New. As a result, when the weighting parameter λ is chosen to be 0.027, it gives signals to 7 stroke patients and 142 non-stroke patients in the test data. Thus, when $ATS_0 = 15$, New is still better because it give signals to less non-stroke patients, compared to DySS, when their numbers of signals to stroke patients are kept the same. Then, we use the results in rows 1, 2, 3 and 5 of Table 4 to perform a Cochran-Mantel-Haenszel chi-squared test for examining whether the proportion of signals to non-stroke patients given by New is significantly smaller than that of DySS in both cases when $ATS_0 = 10$ and 15. This test gives a p-value of 0.0393, implying a significant improvement of the proposed method New over the DySS method in terms of the proportion of signals given to non-stroke patients. In Table 4, the calculated ATS_1 values of the methods New and DySS are also presented. It can be seen that ATS_1 values of New are considerably smaller than those of DySS. As a side note, in cases when $ATS_0 = 15$, there are still 125 out of 342 non-stroke patients receive signals from New. Figure 8 shows the estimated disease risks of these 125 non-stroke patients and the remaining $342 - 125 = 217$ non-stroke patients who do not receive any signals. The bold dashed and solid lines are their means. From the plot, it can be seen that the estimated disease risks of the 125 non-stroke patients who receive signals are generally higher than those of the non-stroke patients who do not receive signals, which confirms that New is indeed effective in detecting upward mean shifts in the estimated disease risks.

5 Discussion and Concluding Remarks

In this paper, a new disease early detection method based on risk estimation and risk monitoring is proposed. Numerical results and theoretical arguments show that it performs well in practice

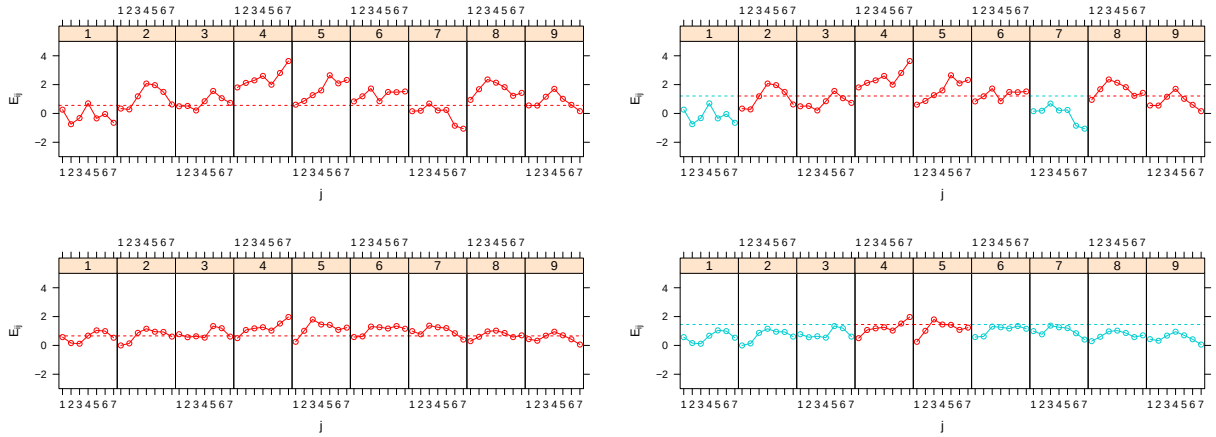

Figure 7: Control charts New (upper) and DySS (bottom) for monitoring the 9 stroke patients in the test data when $ATS_0 = 10$ (left) and $ATS_0 = 15$ (right). The dashed horizontal line in each plot denotes the control limit.

Table 4: Numbers of signals and calculated ATS_1 values of the two methods New and DySS in the stroke data example when they are applied to the 351 people contained in the test data.

ATS_0	λ	Method	ATS_1	# Signals to Stroke Patients (9 in total)	# Signals to Non-Stroke Patients (342 in total)
10	0.1	New	7.56	9	187
	0.1	DySS	8.56	9	202
15	0.1	New	11.57	7	125
	0.1	DySS	17.50	2	52
	0.027	DySS	15.71	7	142

and outperforms the existing method that monitors the original disease risk factors directly. There are still many issues about the proposed method for us to address in the future research. For the time-varying coefficient model (7), the consistency of proposed estimate for $\boldsymbol{\theta}(t)$ needs to be further validated theoretically. By using Model (7), the proposed cross-validation method for selecting h_θ discussed in section 2.1 will become highly computationally intensive, which should be addressed properly in the future research. Also, although Model (7) is already quite flexible, it can be further generalized to a nonparametric model (e.g., Chen and Zhou 2007). It is still unknown to us whether such a generalization would improve the performance of the proposed method in a substantial way. The proposed chart (8)-(9) depends on the weighting parameter λ. In the SPC literature, this

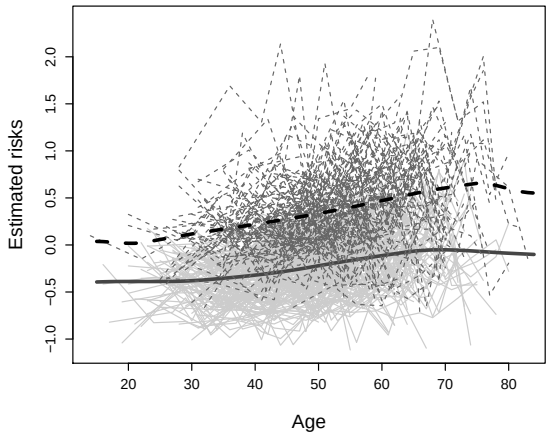

Figure 8: Dashed thin lines denote the estimated disease risks of the 125 non-stroke patients who received signals from New, and solid thin lines denote the estimated disease risks of the non-stroke patients who did not receive any signals from New. The bold dashed and solid lines are their means.

parameter is usually pre-specified, a large λ value is good for detecting a large shift, and a small λ value is good for detecting a small shift. In some applications, it is often unknown whether a future shift is large or small. So, pre-specification of λ becomes difficult. In the literature, there are two approaches to overcome this difficulty. One is to use the so-called adaptive EWMA chart, in which the shift size is sequentially estimated during online process monitoring and then the λ value is adjusted accordingly (cf., Qiu 2014, Section 5.4). Another strategy is to use a set of λ values and the charting statistic at each time point is defined to be the maximum of the charting statistics with the individual λ values (e.g., Qiu et al. 2018). Both strategies will be considered in our future research.

Supplementary Materials

ComputerCodesAndData.zip: This zip file contains computer codes of our proposed method and the stroke data used in the paper.

Acknowledgments: The authors thank the editor, the associate editor and two referees for their constructive suggestions and comments, which improved the quality of the paper substantially.

Appendix

For ease of presentation, we will simply write h_β, h_μ, h_σ as h (or h_n if we want to emphasize that they are a sequence indexed by n). In all appendices, for matrices $\mathbf{A}$ and $\mathbf{B}$, the notation $\mathbf{A} \geq 0$ means that $\mathbf{A}$ is a positive-semidefinite matrix, while $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is a positive-semidefinite matrix. Let $\tilde{\mathbf{x}}_i(t) = (\boldsymbol{x}_i'(t), \boldsymbol{z}_i')'$ denote the stacked vector of $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i$. The gradient of a function $f(\boldsymbol{\theta})$ with respect to the vector $\boldsymbol{\theta}$ will be written as ∇f, and $\nabla^2 f = (\nabla \cdot \nabla) f$ denotes the second derivative of f. For clarity, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0')'$ is used to denote the true value of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$.

A Proof of Theorem 1

Lemma A1 (Cauchy-Schwarz inequality). For random variables $\{w_i, i = 1, 2, \ldots, n\}$ and random vectors $\{\boldsymbol{v}_i, i = 1, 2, \ldots, n\}$, we have

$$\left[\sum_{i=1}^n w_i\right]\left[\sum_{i=1}^n w_i \boldsymbol{v}_i \boldsymbol{v}_i'\right] \geq \left[\sum_{i=1}^n w_i \boldsymbol{v}_i\right]\left[\sum_{i=1}^n w_i \boldsymbol{v}_i\right]',$$

$$\mathrm{E}[w_i]\,\mathrm{E}[w_i \boldsymbol{v}_i \boldsymbol{v}_i'] \geq \mathrm{E}[w_i \boldsymbol{v}_i]\,\mathrm{E}[w_i \boldsymbol{v}_i]'.$$

Lemma A2. Under the assumptions (a)-(d) in Theorem 1, we have $\mathrm{P}\left(y_i(t) = 1, \text{ for all } t \in [0,1]\right) > 0$.

Proof. Since $y_i(t) = 1$ for all $t \in [0, \mathcal{T}]$ if and only if $T_i \geq \mathcal{T}$, we only need to verify $\mathrm{P}(T_i \geq \mathcal{T}) > 0$. It can be checked that

$$\begin{aligned}\mathrm{P}(D_i \geq \mathcal{T}) &= \mathrm{E}\{\exp[-\textstyle\int_0^{\mathcal{T}} \exp(r_i(t))\, dt]\} \\ &\geq \mathrm{E}\{\exp[-\exp(\sup{}_{t\in[0,\mathcal{T}]}|r_i(t)| + \log \mathcal{T})]\}.\end{aligned}$$

Since $\exp(-\exp(x))$ is a concave function of x, by the Jensen's inequality, we have

$$\mathrm{P}(D_i \geq \mathcal{T}) \geq \exp\Big[-\exp\Big(\mathrm{E}[\sup{}_{t\in[0,\mathcal{T}]}|r_i(t)|]\Big)\Big] > 0.$$

□

Lemma A3. The logarithm of the partial likelihood function

$$l(\boldsymbol{\theta}) = \sum_{i:\delta_i=1} \left\{ \boldsymbol{\theta}' \tilde{\mathbf{x}}_i(T_i) - \log\left[\sum_{l\in R(T_i)} \sum_{j=1}^{m_l} K_h(T_i - t_{lj}) \exp(\boldsymbol{\theta}' \tilde{\mathbf{x}}_i(T_i))\right]\right\}$$

is a concave function of $\boldsymbol{\theta}$.

Proof. By definition, we have

$$S_0(\boldsymbol{\theta};t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t-t_{ij})y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij})),$$

$$\mathbf{S}_1(\boldsymbol{\theta};t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t-t_{ij})y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))\tilde{\mathbf{x}}_i(t_{ij}),$$

$$\mathbf{S}_2(\boldsymbol{\theta};t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t-t_{ij})y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))\tilde{\mathbf{x}}_i(t_{ij})\tilde{\mathbf{x}}_i'(t_{ij}).$$

Then,

$$\nabla l(\boldsymbol{\theta}) = \sum_{i:\delta_i=1}\left[\tilde{\mathbf{x}}_i(T_i) - \frac{\mathbf{S}_1(\boldsymbol{\theta};T_i)}{S_0(\boldsymbol{\theta};T_i)}\right],$$

$$\nabla^2 l(\boldsymbol{\theta}) = -\sum_{i:\delta_i=1}\left[\frac{S_0(\boldsymbol{\theta};T_i)\mathbf{S}_2(\boldsymbol{\theta};T_i) - \mathbf{S}_1(\boldsymbol{\theta};T_i)\mathbf{S}_1'(\boldsymbol{\theta};T_i)}{[S_0(\boldsymbol{\theta};T_i)]^2}\right].$$

By Lemma A1, we can show that $S_0(\boldsymbol{\theta};T_i)\mathbf{S}_2(\boldsymbol{\theta};T_i) \geq \mathbf{S}_1(\boldsymbol{\theta};T_i)\mathbf{S}_1'(\boldsymbol{\theta};T_i)$, which implies directly that $\nabla^2 l(\boldsymbol{\theta})$ is negative-semidefinite. Therefore, the log-likelihood function $l(\boldsymbol{\theta})$ is a concave function. □

Lemma A4. Under the assumptions in Theorem 1, we have

$$\sup_{t\in[0,\mathcal{T}]}\left|S_0(\boldsymbol{\theta},t) - \mathrm{E}[S_0(\boldsymbol{\theta},t)]\right| \xrightarrow{p} 0, \text{ as } n\to\infty.$$

Proof. Let $\zeta(t) = \int e^{-iut}K(u)\,du$ be the inverse Fourier transform of $K(t)$. So, we have $K(t) = \frac{1}{2\pi}\int e^{iut}\zeta(u)\,du$. Then,

$$\begin{aligned}
S_0(\boldsymbol{\theta},t) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{h_n}(t_{ij}-t)y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{1}{2\pi h_n}\int e^{iu(t_{ij}-t)/h_n}\zeta(u)\,du\right] y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij})) \\
&= \frac{1}{2\pi}\int \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} e^{iu(t_{ij}-t)}\zeta(uh_n)y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))\,du \\
&= \frac{1}{2\pi}\int \frac{1}{n}\sum_{i=1}^{n} I(T_i\geq t)\sum_{j=1}^{m_i} e^{iu(t_{ij}-t)}\zeta(uh_n)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))\,du \\
&= \frac{e^{-iut}}{2\pi}\int\left[\frac{1}{n}\sum_{i=1}^{n} I(T_i\geq t)\sum_{j=1}^{m_i} e^{iut_{ij}}\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))\right]\zeta(uh_n)\,du.
\end{aligned}$$

Let $\psi_i(u) = \sum_{j=1}^{m_i} e^{iut_{ij}}\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))$. Since $\tilde{\mathbf{x}}_i(t)$ is assumed to be bounded, $\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t_{ij}))$ is bounded by some constant M. So, $|\psi_i(u)|^2 \leq Mm_i^2$ uniformly in u, and $\sup_{t\in[0,\mathcal{T}]}\mathrm{E}[(I(T_i \geq$

$t)\psi_i(u))^2] \le \mathrm{E}[\psi_i^2(u)] \le M\,\mathrm{E}[m_i^2]$. By Lemma 1 of Cai and Sun (2003), it can be shown that

$$\sup_{t\in[0,\mathcal{T}]}\left|\frac{1}{n}\sum_{i=1}^{n} I(T_i \ge t)\psi_i(u) - \mathrm{E}[I(T_i \ge t)\psi_i(u)]\right| \le \Psi_n = O_p(1/\sqrt{n}).$$

Consequently,

$$\begin{aligned}\sup_{t\in[0,\mathcal{T}]}|S_0(t) - \mathrm{E}[S_0(t)]| &\le \frac{1}{2\pi}\left|\int\left\{\frac{1}{n}\sum_{i=1}^{n} I(T_i \ge t)\psi_i(u) - \mathrm{E}\left[I(T_i \ge t)\psi_i(u)\right]\right\}\zeta(uh_n)\,du\right| \\ &\le \frac{1}{2\pi}\int|\Psi_n||\zeta(uh_n)|\,du = O_p(1/\sqrt{n}h_n).\end{aligned}$$

□

Lemma A5. Under the assumptions in Theorem 1, we have

$$\begin{aligned}\sup_{t\in[0,\mathcal{T}]}\Bigg|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\left(\frac{t_{ij}-t}{h}\right)^l y_i(t)[r_i(t_{ij}) - \mu(t_{ij})] \\ - \mathrm{E}\left[\sum_{j=1}^{m_i} K_h(t_{ij}-t)\left(\frac{t_{ij}-t}{h}\right)^l y_i(t)[r(t_{ij}) - \mu(t_{ij})]\right]\Bigg| = O_p(1/\sqrt{n}h_n),\end{aligned}$$

for $l = 0, 1$, where $r_i(t) = \boldsymbol{\beta}'\boldsymbol{x}_i(t)$.

Proof. Let $\zeta(t) = \int e^{-iut}u^l K(u)\,du$ be the inverse Fourier transform of $t^l K(t)$. So, we have $t^l K(t) = \frac{1}{2\pi}\int e^{iut}\zeta(u)\,du$, for $l = 0, 1, 2$. Then, the remaining part of the proof is similar to the proof of the previous lemma and thus omitted. One can also refer to Lemma A.1 in Yao et al. (2005) for details. □

Proof of Result (i) in Theorem 1. Let $\|\cdot\|$ denote the maximum norm of a vector, $M_i^*(t) = M_i(t) - \int_0^t y_i(u)\phi(u)\,du$, $N_i^*(t) = N_i(t) - \int_0^t y_i(u)\lambda_i(u)\,du$,

$$\begin{aligned}S_0^\dagger(\boldsymbol{\theta}, t) &= \frac{1}{n}\sum_{i=1}^{n} y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t)), \\ \mathbf{S}_1^\dagger(\boldsymbol{\theta}, t) &= \frac{1}{n}\sum_{i=1}^{n} y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\tilde{\mathbf{x}}_i(t), \\ \mathbf{S}_2^\dagger(\boldsymbol{\theta}, t) &= \frac{1}{n}\sum_{i=1}^{n} y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\tilde{\mathbf{x}}_i(t)\tilde{\mathbf{x}}_i'(t),\end{aligned}$$

and

$$\begin{aligned}s_0(\boldsymbol{\theta}, t) &= \mathrm{E}\left[y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\right], \\ \mathbf{s}_1(\boldsymbol{\theta}, t) &= \mathrm{E}\left[y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\tilde{\mathbf{x}}_i(t)\right], \\ \mathbf{s}_2(\boldsymbol{\theta}, t) &= \mathrm{E}\left[y_i(t)\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\tilde{\mathbf{x}}_i(t)\tilde{\mathbf{x}}_i'(t)\right].\end{aligned}$$

Since $\tilde{\mathbf{x}}_i(t)$ is bounded, we can find M such that $|\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))|$, $\|\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\tilde{\mathbf{x}}_i(t)\|$, and $\|\exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(t))\tilde{\mathbf{x}}_i(t)\tilde{\mathbf{x}}_i'(t)\|$ are all bounded by M uniformly in $t \in [0, \mathcal{T}]$. Then by Lemma 1 in Cai and Sun (2003), $\sup_{t\in[0,\mathcal{T}]} \|S_0^\dagger(\boldsymbol{\theta}, t) - s_0(\boldsymbol{\theta}, t)\| \xrightarrow{p} 0$ and $\sup_{t\in[0,\mathcal{T}]} \|\mathbf{S}_l^\dagger(\boldsymbol{\theta}, t) - \mathbf{s}_l(\boldsymbol{\theta}, t)\| \xrightarrow{p} 0$ for $l = 1, 2$. By Lemma A4, $\sup_{t\in[0,\mathcal{T}],\boldsymbol{\theta}\in\mathscr{B}} \|S_0(\boldsymbol{\theta}, t) - \mathrm{E}[S_0(\boldsymbol{\theta}, t)]\| \xrightarrow{p} 0$. Note that $\mathrm{E}[S_0(\boldsymbol{\theta}, t)] = \mathrm{E}[\int_0^{\mathcal{T}} K_h(t-s) y_i(t) \exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(s)) dM_i(s)] = \int_0^{\mathcal{T}} K_h(t-s)\, \mathrm{E}[y_i(t) \exp(\boldsymbol{\theta}'\tilde{\mathbf{x}}_i(s))]\phi(s)\, ds$. So, by the Taylor expansion and all assumptions in the theorem, it is easy to show that $\sup_{t\in[h,\mathcal{T}-h]} \|\, \mathrm{E}[S_0(\boldsymbol{\theta}, t)] - s_0(\boldsymbol{\theta}, t)\phi(t)\| \to 0$. Consequently, we have $\sup_{t\in[h,\mathcal{T}-h]} \|S_0(\boldsymbol{\theta}, t) - s_0(\boldsymbol{\theta}, t)\phi(t)\| \xrightarrow{p} 0$. Then,

$$
\begin{aligned}
\frac{1}{n}[l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)] &= \frac{1}{n}\sum_{i=1}^{n} \int_0^{\mathcal{T}} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\tilde{\mathbf{x}}_i(t) - \log\left[\frac{S_0(\boldsymbol{\theta}, t)}{S_0(\boldsymbol{\theta}_0, t)}\right] \right\} dN_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_0^{\mathcal{T}} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\tilde{\mathbf{x}}_i(t) - \log\left[\frac{S_0(\boldsymbol{\theta}, t)}{S_0(\boldsymbol{\theta}_0, t)}\right] \right\} dN_i^*(t) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \int_h^{\mathcal{T}-h} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\tilde{\mathbf{x}}_i(t) - \log\left[\frac{S_0(\boldsymbol{\theta}, t)}{S_0(\boldsymbol{\theta}_0, t)}\right] \right\} y_i(t)\lambda_i(t)\, dt + O_p(h).
\end{aligned}
$$

Let

$$
\begin{aligned}
X_n(\boldsymbol{\theta}, t) &= \frac{1}{n}\sum_{i=1}^{n} \int_0^{t} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\tilde{\mathbf{x}}_i(s) - \log\left[\frac{S_0(\boldsymbol{\theta}, s)}{S_0(\boldsymbol{\theta}_0, s)}\right] \right\} dN_i^*(s), \\
A_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \int_h^{\mathcal{T}-h} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\tilde{\mathbf{x}}_i(s) - \log\left[\frac{S_0(\boldsymbol{\theta}, s)}{S_0(\boldsymbol{\theta}_0, s)}\right] \right\} y_i(s)\lambda_i(s)\, ds \\
&= \int_h^{\mathcal{T}-h} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{S}_1^\dagger(\boldsymbol{\theta}_0; s) - \log\left[\frac{S_0(\boldsymbol{\theta}, s)}{S_0(\boldsymbol{\theta}_0, s)}\right] S_0^\dagger(\boldsymbol{\theta}_0; s) \right\} \lambda_0(s)\, ds.
\end{aligned}
$$

Since $N_i^*(s)$ is a martingale, the quadratic variation process of $X_n(\boldsymbol{\theta}, \mathcal{T})$ can be written as

$$
\begin{aligned}
&\langle X_n(\boldsymbol{\theta}, \mathcal{T}), X_n(\boldsymbol{\theta}, \mathcal{T})\rangle \\
&\quad = \frac{1}{n^2}\sum_{i=1}^{n} \int_0^{\mathcal{T}} \left[(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\tilde{\mathbf{x}}_i(t) - \log\left[\frac{S_0(\boldsymbol{\theta}, s)}{S_0(\boldsymbol{\theta}_0, s)}\right] \right]^2 y_i(s)\lambda_i(s)\, ds \\
&\quad = \frac{1}{n}\int_0^{\mathcal{T}} \left[(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{S}_2^\dagger(\boldsymbol{\theta}_0, s)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - 2(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{S}_1^\dagger(\boldsymbol{\theta}_0, s) \log\left[\frac{S_0(\boldsymbol{\theta}, s)}{S_0(\boldsymbol{\theta}_0, s)}\right] \right. \\
&\qquad \left. + \left\{ \log\left[\frac{S_0(\boldsymbol{\theta}, s)}{S_0(\boldsymbol{\theta}_0, s)}\right] \right\}^2 S_0^\dagger(\boldsymbol{\theta}_0, s) \right] ds.
\end{aligned}
$$

By the convergence of $S_0(\boldsymbol{\theta}, t)$ and $\mathbf{S}_l^\dagger(\boldsymbol{\theta}, t)$, for $l = 0, 1, 2$, $n\langle X_n(\boldsymbol{\theta}, \mathcal{T}), X_n(\boldsymbol{\theta}, \mathcal{T})\rangle$ converges in probability to a finite limit. Then, by a corollary to the Lenglart inequality (cf., Theorem I.1(b) in Andersen and Gill 1982), the above result implies that $\lim_{n\to\infty} X_n(\boldsymbol{\theta}, \mathcal{T}) = 0$. Again, by the uniform convergence of $S_0(\boldsymbol{\theta}, t)$ and $\mathbf{S}_l^\dagger(\boldsymbol{\theta}, t)$, for $l = 0, 1, 2$, we have

$$
A_n(\boldsymbol{\theta}) \xrightarrow{p} \int_0^{\mathcal{T}} \left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{s}_1(\boldsymbol{\theta}_0; s) - \log\left[\frac{s_0(\boldsymbol{\theta}, s)}{s_0(\boldsymbol{\theta}_0, s)}\right] s_0(\boldsymbol{\theta}_0; s) \right\} \lambda_0(s)\, ds.
$$

Thus, $\frac{1}{n}[l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)]$ will converge to the same limit as $A_n(\boldsymbol{\theta})$. Let the above limit be $A(\boldsymbol{\theta})$. Then, we can conclude that $\frac{1}{n}[l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)] \xrightarrow{p} A(\boldsymbol{\theta})$. By Theorem 4.1 of Andersen and Gill (1982), we can evaluate the first and second derivatives of $A(\boldsymbol{\theta})$ in a compact and convex set $\mathscr{B}$ containing $\boldsymbol{\theta}_0$:

$$\nabla A(\boldsymbol{\theta}) = \int_0^{\mathcal{T}} \left[\frac{\mathbf{s}_1(\boldsymbol{\theta}_0; t)}{s_0(\boldsymbol{\theta}_0; t)} - \frac{\mathbf{s}_1(\boldsymbol{\theta}; t)}{s_0(\boldsymbol{\theta}; t)}\right] s_0(\boldsymbol{\theta}_0; t)\lambda_0(t)\, dt,$$
$$\nabla^2 A(\boldsymbol{\theta}) = -\int_0^{\mathcal{T}} \left[\frac{s_0(\boldsymbol{\theta}; t)\mathbf{s}_2(\boldsymbol{\theta}; t) - \mathbf{s}_1(\boldsymbol{\theta}; t)\mathbf{s}_1'(\boldsymbol{\theta}; t)}{[s_0(\boldsymbol{\theta}; t)]^2}\right] s_0(\boldsymbol{\theta}_0; t)\lambda_0(t)\, dt.$$

Furthermore, $\nabla A(\boldsymbol{\theta}_0) = 0$ and $\nabla^2 A(\boldsymbol{\theta}_0) \leq 0$ by Lemma A1. So, $A(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta}_0$ in $\mathscr{B}$. Since $\frac{1}{n}[l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)]$ is a convex function of $\boldsymbol{\theta}$, by Theorem II.1 and Corollary II.2 of Andersen and Gill (1982), $\frac{1}{n}[l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}_0)]$ converges uniformly to $A(\boldsymbol{\theta})$ in probability for all $\boldsymbol{\theta} \in \mathscr{B}$. Because $A(\boldsymbol{\theta})$ is also a convex function of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. □

Proof of Result (ii) in Theorem 1. We first proceed our arguments in the case when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and will plug in the consistent estimate $\widehat{\boldsymbol{\beta}}$ in the last step. Let $G(s,t) = \mathrm{E}[r_i(s)|T_i \geq t]$ and $R_l^*(t) = R_l(t) - \mu(t)W_l(t) - h\mu'(t)W_{l+1}(t)$. Then,

$$\widehat{\mu}(t) - \mu(t) = \frac{R_0^*(t)W_2(t) - R_1^*(t)W_1(t)}{W_0(t)W_2(t) - W_1(t)^2},$$

and

$$\begin{aligned}
R_l^*(t) =& R_l(t) - \mu(t)W_l(t) - h\mu'(t)W_{l+1}(t) \\
=& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij} - t)\Big(\frac{t_{ij} - t}{h}\Big)^l y_i(t)[r_i(t_{ij}) - \mu(t) - \mu'(t)(t_{ij} - t)] \\
=& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij} - t)\Big(\frac{t_{ij} - t}{h}\Big)^l y_i(t)[r_i(t_{ij}) - \mu(t_{ij})] \\
&+ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij} - t)\Big(\frac{t_{ij} - t}{h}\Big)^l y_i(t)[\mu(t_{ij}) - \mu(t) - \mu'(t)(t_{ij} - t)].
\end{aligned}$$

By Lemma A5, we have

$$\sup_{t\in[0,\mathcal{T}]} \left| \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij} - t)\Big(\frac{t_{ij} - t}{h}\Big)^l y_i(t)[r_i(t_{ij}) - \mu(t_{ij})] \right.$$
$$\left. - \mathrm{E}\left[\sum_{j=1}^{m_i} K_h(t_{ij} - t)\Big(\frac{t_{ij} - t}{h}\Big)^l y_i(t)[r_i(t_{ij}) - \mu(t_{ij})]\right] \right| \xrightarrow{p} 0.$$

Note that

$$\begin{aligned}
&\mathrm{E}\left[\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[r_i(t_{ij})-\mu(t_{ij})]\right] \\
&= \mathrm{E}\Big[\int_0^{\mathcal{T}} K_h(s-t)\Big(\frac{s-t}{h}\Big)^l y_i(t)[r_i(s)-\mu(s)]\phi(s)\,ds\Big] \\
&= \int_0^{\mathcal{T}} K_h(s-t)\Big(\frac{s-t}{h}\Big)^l \mathrm{E}\big[y_i(t)[r_i(s)-\mu(s)]\big]\phi(s)\,ds.
\end{aligned}$$

By the Taylor's expansion, it can be shown that the above quantity will tend to $\mathrm{E}[y_i(t)[r_i(t)-\mu(t)]] = \mathrm{E}[r_i(t)-\mu(t)|T_i \geq t]\,\mathrm{P}(T_i \geq t) = 0$ uniformly for t in $[h, \mathcal{T}-h]$. Therefore,

$$\sup_{t\in[0,\mathcal{T}]}\left|\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[r_i(t_{ij})-\mu(t_{ij})]\right| \xrightarrow{p} 0.$$

On the other hand, $\mu(s)-\mu(t)-\mu'(t)(s-t) = O(|s-t|^2)$, and $K_h(t_{ij}-t) = 0$ when $|t_{ij}-t| > h$. So,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\mu(t_{ij})-\mu(t)-\mu'(t)(t_{ij}-t)] = O_p(h)$$

uniformly for $t \in [0, \mathcal{T}]$. Therefore, the decomposed terms of R_l^* all tend to 0, and we have

$$\sup_{t\in[h,\mathcal{T}-h]} |R_l^*(t)| \xrightarrow{p} 0, \text{ for } l = 0, 1.$$

By similar arguments, we can show that

$$\begin{aligned}
&\sup_{t\in[h,\mathcal{T}-h]} |W_0(t) - \phi(t)\,\mathrm{P}(T_i \geq t)| \xrightarrow{p} 0, \\
&\sup_{t\in[h,\mathcal{T}-h]} |W_1(t)| \xrightarrow{p} 0, \\
&\sup_{t\in[h,\mathcal{T}-h]} |W_2(t) - K_2\phi(t)\,\mathrm{P}(T_i \geq t)| \xrightarrow{p} 0,
\end{aligned}$$

where $K_2 = \int s^2 K(s)\,ds$. Therefore, we have

$$\sup_{t\in[h,\mathcal{T}-h]} |\widehat{\mu}(t) - \mu(t)| = \sup_{t\in[h,\mathcal{T}-h]}\left|\frac{R_0^*(t)W_2(t) - R_1^*(t)W_1(t)}{W_0(t)W_2(t) - W_1(t)^2}\right| \xrightarrow{p} 0.$$

So far, we have showed the consistency of $\widehat{\mu}(t;\boldsymbol{\beta}_0)$. Because $\widehat{\mu}(t;\boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$, the same conclusion holds for $\widehat{\mu}(t;\widehat{\boldsymbol{\beta}})$. □

Proof of Result (iii) in Theorem 1. As in the proof for Result (ii) in Theorem 1, we only provide arguments in the case when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The results when $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ can be obtained at the final step

after replacing $\boldsymbol{\beta}_0$ with $\widehat{\boldsymbol{\beta}}$ and after using the consistency of $\widehat{\boldsymbol{\beta}}$. Let $\epsilon_i(t) = r_i(t) - \mu(t;\boldsymbol{\beta}_0)$ and $\widehat{\epsilon}_i(t) = r_i(t) - \widehat{\mu}(t;\boldsymbol{\beta}_0)$. Then,

$$\begin{aligned} Q_l^*(t) &= Q_l(t) - \sigma^2(t)W_l - h[\sigma^2(t)]'W_{l+1} \\ &= \frac{1}{n_t}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\widehat{\epsilon}^2(t_{ij}) - \sigma^2(t) - [\sigma^2(t)]'(t_{ij}-t)]. \end{aligned}$$

So, we have

$$\widehat{\sigma}^2(t) - \sigma^2(t) = \frac{Q_0^*(t)W_2(t) - Q_1^*(t)W_1(t)}{W_0(t)W_2(t) - W_1(t)^2}.$$

The convergence of $W_l(t)$, for $l = 0, 1, 2$, has been established in the proof of Result (ii), where we have

$$\begin{aligned} &\sup_{t\in[h,\mathcal{T}-h]} |W_0(t) - \phi(t)\,\mathrm{P}(T_i \geq t)| \xrightarrow{p} 0, \\ &\sup_{t\in[h,\mathcal{T}-h]} |W_1(t)| \xrightarrow{p} 0, \\ &\sup_{t\in[h,\mathcal{T}-h]} |W_2(t) - K_2\phi(t)\,\mathrm{P}(T_i \geq t)| \xrightarrow{p} 0. \end{aligned}$$

Again, we can consider the following decomposition of $Q_l^*(t)$:

$$\begin{aligned} Q_l^*(t) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\widehat{\epsilon}_i^2(t_{ij}) - \epsilon_i^2(t_{ij})] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\epsilon_i^2(t_{ij}) - \sigma^2(t_{ij})] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\sigma^2(t_{ij}) - \sigma^2(t) - [\sigma^2(t)]'(t_{ij}-t)]. \end{aligned}$$

For the first term on th right-hand-side, we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\widehat{\epsilon}_i^2(t_{ij}) - \epsilon_i^2(t_{ij})] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[(\widehat{\epsilon}_i(t_{ij}) + \epsilon_i(t_{ij}))(\widehat{\epsilon}_i(t_{ij}) - \epsilon_i(t_{ij}))] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[(\widehat{\epsilon}_i(t_{ij}) + \epsilon_i(t_{ij}))(\mu(t_{ij}) - \widehat{\mu}(t_{ij}))]. \end{aligned}$$

So, this term will uniformly converge to 0 in probability due to the uniform convergence of $\widehat{\mu}(t) -$

$\mu(t)$. By some arguments similar to those in the proof of Lemma A5, we have

$$\sup_{t\in[0,\mathcal{T}]}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\epsilon_i^2(t_{ij})-\sigma^2(t_{ij})]\right.$$
$$\left.-\,\mathrm{E}\left[\sum_{j=1}^{m_i}K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\epsilon_i^2(t_{ij})-\sigma^2(t_{ij})]\right]\right|\xrightarrow{p}0.$$

Thus,

$$\sup_{t\in[h,\mathcal{T}-h]}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\epsilon_i^2(t_{ij})-\sigma^2(t_{ij})]\right|\xrightarrow{p}0.$$

By combining these results with the following fact that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l y_i(t)[\sigma^2(t_{ij})-\sigma^2(t)-[\sigma^2(t)]'(t_{ij}-t)]=O_p(h),$$

which is uniformly true for $t\in[0,\mathcal{T}]$, we can conclude that

$$\sup_{t\in[h,\mathcal{T}-h]}|Q_l^*(t)|\xrightarrow{p}0,\ \text{for } l=0,1.$$

After plugging in all the established convergences of the related terms into the expression $\frac{Q_0^*(t)W_2(t)-Q_1^*(t)W_1(t)}{W_0(t)W_2(t)-W_1(t)^2}$, we have $\sup_{t\in[h,\mathcal{T}-h]}|\widehat{\sigma}^2(t)-\sigma^2(t)|\to 0$. □

B Derivation of Equation (6)

The leave-one-out cross-validation score in Equation (6) is modified from the prediction error score discussed in Tian et al. (2005). Let $\widehat{\boldsymbol{\theta}}_{-i}$ and $\widehat{\Lambda}_{-i}(t)$ be the estimates of $\boldsymbol{\theta}$ and baseline cumulative hazard function $\Lambda_0(t)$ when observations of the i subject are excluded, and $PE_i^{\theta}(h)$ be the squared

integrated value of the martingale residuals. Then, we have

$$
\begin{aligned}
PE_i^{\theta}(h) &= \int_0^{\mathcal{T}} \left[N_i(t) - \int_0^t y_i(s) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_i(s)\} \, d\widehat{\Lambda}_{-i}(s) \right]^2 dN_i(t) \\
&= \int_0^{\mathcal{T}} \left[N_i(t) - \int_0^t y_i(s) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_i(s)\} \, d\left(\sum_{k \neq i,\, \delta_k = 1} \frac{N_k(s)}{\sum_{d \neq i} y_d(T_k) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_d(s)\}} \right) \right]^2 dN_i(t) \\
&= \int_0^{\mathcal{T}} \left[N_i(t) - \sum_{k \neq i,\, \delta_k = 1} \frac{y_i(T_k) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_i(T_k)\}}{\sum_{d \neq i} y_d(T_k) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_d(T_k)\}} \right]^2 dN_i(t) \\
&\approx \int_0^{\mathcal{T}} \left[N_i(t) - \sum_{k \neq i \, \delta_k = 1} \frac{\sum_{j=1}^{m_i} y_i(T_k) K_h(T_k - t_{ij}) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_i(t_{ij})\}}{\sum_{d \neq i} \sum_{j=1}^{m_d} y_d(T_k) K_h(T_k - t_{dj}) \exp(\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_d(t_{dj}))} \right]^2 dN_i(t) \\
&= \left(N_i(T_i) - \sum_{\substack{k \neq i,\, \delta_k = 1 \\ T_k \leq T_i}} \frac{\sum_{j=1}^{m_i} K_h(T_k - t_{ij}) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_i(t_{ij})\}}{\sum_{d \neq i, d \in R(T_k)} \sum_{j=1}^{m_d} K_h(T_k - t_{dj}) \exp(\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_d(t_{dj}))} \right)^2 \\
&= \left(\delta_i - \sum_{\substack{k \neq i,\, \delta_k = 1 \\ T_k \leq T_i}} \frac{\sum_{j=1}^{m_i} K_h(T_k - t_{ij}) \exp\{\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_i(t_{ij})\}}{\sum_{d \neq i, d \in R(T_k)} \sum_{j=1}^{m_d} K_h(T_k - t_{dj}) \exp(\widehat{\boldsymbol{\theta}}'_{-i} \tilde{\mathbf{x}}_d(t_{dj}))} \right)^2 .
\end{aligned}
$$

The last equation is Equation (6).

References

Andersen, P. K., and Gill, R. D. (1982), "Cox's regression model for counting processes," *Annals of Statistics*, **10**, 1100–1120.

Cai, Z., and Sun, Y. (2003), "Local linear estimation for time-dependent coefficients in cox's regression models," *Scandinavian Journal of Statistics*, **30**, 93–111.

Chen, K., and Jin, Z. (2005), "Local polynomial regression analysis of clustered data," *Biometrika*, **92**, 59–74.

Chen, S., and Zhou, L. (2007). "Local partial likelihood estimation in proportional hazards regression," *The Annals of Statistics*, **35**, 888–916.

Cohn, J. N., Hoke, L., Whitwam, W., Sommers, P. A., Taylor, A. L., Duprez, D., Roessler, R., and Florea, N. (2003), "Screening for early detection of cardiovascular disease in asymptomatic individuals," *American Heart Journal*, **146**, 679–685.

Dupuy, J. F., Grama, I., and Mesbah, M. (2006), "Asymptotic theory for the cox model with missing time-dependent covariate," *Annals of Statistics*, **34**, 903–924.

Epanechnikov, V. A. (1969), "Non-parametric estimation of a multivariate probability density," *Theory of Probability and Its Applications*, **14**, 153–158.

Grønnesby, J. K., and Borgan, Ø. (1996), "A method for checking regression models in survival analysis based on the risk score." *Lifetime Data Analysis*, **2**, 315–328.

Hastie, T., and Tibshirani, R. (1993), "Varying-coefficient models." *Journal of the Royal Statistical Society, Series B*, **55,** 757–796.

Hawkins, D. M., and Olwell, D. H. (1998), *Cumulative Sum Charts and Charting for Quality Improvement*, New York: Springer-Verlag.

Hendry, D. J. (2014), "Data generation for the cox proportional hazards model with time-dependent covariates: a method for medical researchers," *Statistics in Medicine*, **33**, 436–454.

Hubbard, G., Macmillan, I., Canny, A., Forbat, L., and Neal, R. D. (2014), "Cancer symptom awareness and barriers to medical help seeking in Scottish adolescents: a cross-sectional study," *BMC Public Health*, **14**, 1117.

Kattlove, H., Liberati, A., Keeler, E., and Brook, R. H. (1995), "Benefits and costs of screening and treatment for early breast cancer. Development of a basic benefit package," *Journal of the American Medical Association*, **273**, 142–148.

Lee, S., Huang, H., and Zelen, M. (2004), "Early detection of disease and scheduling of screening examinations," *Statistical Methods in Medical Research*, **13**, 443–456.

Li, Y. (2011), "Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation," *Biometrika*, **98**, 355–370.

Li, W., Dou, W., Pu, X., and Xiang, D. (2018), "Monitoring individuals with irregular semiparametric longitudinal behaviour," *Quality Technology & Quantitative Management*, **15**, 37–52.

Li, J., and Qiu, P. (2016), "Nonparametric dynamic screening system for monitoring correlated longitudinal data," *IIE Transactions*, **48**, 772–786.

Li, J., and Qiu, P. (2017), "Construction of an efficient multivariate dynamic screening system," *Quality and Reliability Engineering International*, **33**, 1969–1981.

Liang, W., Liu, Z., Jin, L., and Wang, H. (2017), "Case study: statistical monitoring for the moisture content of the cut tobacco," *Journal of Industrial and Production Engineering*, **34**, 551–557.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, **73**, 13–22.

Lin, D. Y., and Ying, Z. (1993), "Cox regression with incomplete covariate measurements," *Journal of the American Statistical Association*, **88**, 1341–1349.

Lin, X., and Carroll, R. (2001), "Semiparametric regression for clustered data using generalized estimating equations," *Journal of the American Statistical Association*, **96**, 1045–1056.

Ma, S., Yang, L., and Carroll, R. (2012), "A simultaneous confidence band for sparse longitudinal regression," *Statistica Sinica*, **22**, 95–122.

Mendis, S., Puska, P., and Norrving, B. (2011), *Global Atlas on Cardiovascular Disease Prevention and Control,* World Health Organization, Geneva, Switzerland.

Paik, M. C., and Tsai, W. Y. (1997), "On using the cox proportional hazards model with missing covariates," *Biometrika*, **84**, 579–593.

Park, K. Y., and Qiu, P. (2014), "Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions," *Statistics in Medicine*, **33**, 4532–4546.

Qiu, P. (2014), *Introduction to Statistical Process Control*, Boca Raton, FL: Chapman Hall/CRC.

Qiu, P., Xia, Z., and You, L. (2019), "Process monitoring ROC curve for evaluating dynamic screening methods," *Technometrics*, DOI: 10.1080/00401706.2019.1604434.

Qiu, P., and Xiang, D. (2014), "Univariate dynamic screening system: an approach for identifying individuals with irregular longitudinal behavior," *Technometrics*, **56**, 248–260.

Qiu, P., and Xiang, D. (2015), "Surveillance of cardiovascular diseases using a multivariate dynamic screening system," *Statistics in Medicine*, **34**, 2204–2221.

Qiu, P., Zi, X., and Zou, C. (2018), "Nonparametric dynamic curve monitoring," *Technometrics* **60**, 386–397.

Radbill, B., Murphy, B., and LeRoith, D. (2008), "Rationale and strategies for early detection and management of diabetic kidney disease," *Mayo Clinic Proceedings*, **83**, 1373–1381.

Ridker, P. M. (2003), "Clinical application of C-reactive protein for cardiovascular disease detection and prevention," *Circulation*, **107**, 363–369.

Robb, K., Stubbings, S., Ramirez, A., Macleod, U., Austoker, J., Waller, and Hiom, J. S. (2009), "Public awareness of cancer in Britain: a population-based survey of adults," *British Journal of Cancer*, **101**, S18–S23.

Tian, L., Zucker, D., and Wei, L.J. (2005), "On the Cox model with time-varying regression coefficients," *Journal of the American Statistical Association*, **100**, 172–183.

Wang, N. (2003), "Marginal nonparametric kernel regression accounting for within-subject correlation," *Biometrika*, **90**, 43–52.

Wright, D. J. (1986). "Forecasting data published at irregular time intervals using an extension of Holt's method," *Management Science*, **32**, 499–510.

Xiang, D., Qiu, P., and Pu, X. (2013), "Local polynomial regression analysis of multivariate longitudinal data," *Statistica Sinica*, **23**, 769–789.

Yao, F., Müller, H. G., and Wang, J. L. (2005), "Functional data analysis for sparse longitudinal data." *Journal of the American Statistical Association*, **100**, 577–590.

You, L., and Qiu, P. (2019), "Fast computing for dynamic screening systems when analyzing correlated data," *Journal of Statistical Computation and Simulation*, **89**, 379–394.

Yu, Z., and Lin, X. (2010), "Semiparametric regression with time-dependent coefficients for failure time data analysis," *Statistica Sinica*, **20**, 853–869.

Zelen, M. (1993), "Optimal scheduling of examinations for the early detection of disease," *Biometrika*, **80**, 279–293.

Zhao, Z., and Wu, W. (2008), "Confidence bands in nonparametric time series regression," *Annals of Statistics*, **36**, 1854–1878.

Zucker, D.M., and Karr, A.F. (1990), "Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach," *The Annals of Statistics*, **18**, 329–353.