A Nonparametric Control Chart for Dynamic Disease Risk Monitoring

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

Abstract

Some deadly diseases can be treated or even prevented if they or some of their symptoms are detected early. Disease early detection and prevention is thus important for our health improvement. In this paper, we suggest a novel and effective new method for disease early detection. By this method, a patient's risk to the disease is first quantified at each time point by survival data analysis of a training dataset that contains patients' survival information and longitudinally observed disease predictors (e.g., disease risk factors and other covariates). To improve the effectiveness of the proposed method, variable selection is used in the survival analysis to keep only important disease predictors in disease risk quantification. Then, the longitudinal pattern of the quantified risk is monitored sequentially over time by a nonparametric control chart. A signal will be given by the chart once the cumulative difference between the risk pattern of the patient under monitoring and the risk pattern of a typical person without the disease in concern exceeds a control limit.

Key Words: Disease screening; Disease early detection; Dynamic process; Longitudinal data; Statistical process control; Survival data.

1 Introduction

One of the primary objectives of a disease screening program is to give early signals to patients who have the disease in concern or who are at high risk of having the disease, so that these patients can receive timely intervention and treatment (Qiu and Xiang 2014). This paper aims to develop a novel and effective method for disease screening.

Medical research has identified major predictors of many diseases. For instance, the major predictors for cardiovascular diseases include high blood pressure, high cholesterol level, obesity,

tobacco use, lack of physical activity, diabetes, unhealthy diet, age, gender, family history, and some others (e.g., Mendis et al. 2011). For disease screening, patients often take scheduled disease screening examinations over time to have their medical conditions evaluated. To identify high-risk patients through the data collected during the screening examinations, an effective statistical tool is needed. This type of research problem is called dynamic screening (DS) problem in Qiu and Xiang (2014), because medical data are collected sequentially over time from patients, data distribution would change over time, and decisions about the disease status need to be made sequentially as well during the process of data collection. Qiu and Xiang (2014) proposed a dynamic screening system (DySS) to monitor a single disease predictor over time for handling the DS problem. In their method, they first model the regular longitudinal pattern of the disease predictor by a nonparametric longitudinal model estimated from an in-control (IC) dataset that contains observed data of the disease predictor of patients without the disease in concern. Then, to monitor the disease predictor of a new individual, they constructed a statistical process control (SPC) chart to detect undesirable deviations and/or changes in the longitudinal pattern of the disease predictor of the individual under monitoring from the estimated regular longitudinal pattern. By employing a cumulative sum (CUSUM) control chart, this method makes use of the observed data at the current time point and all history data efficiently, and it has been demonstrated to good performance in many applications. In subsequent research, Qiu and Xiang (2015) further extended the DySS method to multivariate cases where multiple disease predictors are considered. A multivariate control chart was proposed to jointly monitor all disease predictors. Some other extensions of the DySS method include those discussed in Li and Qiu (2016, 2017) and You and Qiu (2018) where serially correlated data are considered, and the one discussed in Qiu et al. (2018) where unequally-spaced observation times were accommodated in the construction of the control chart. Qiu et al. (2019) proposed a new metric for evaluating the numerical performance of DS methods.

In practice, there could be many different disease predictors involved. Some of them might be more important than the others in predicting the occurrence of the disease in question. But, in the multivariate DySS methods mentioned above, all disease predictors are treated equally in constructing the related multivariate control charts, which would make the charts less effective in predicting the disease. To overcome this limitation, You and Qiu (2019) recently proposed a new method consisting of the following two steps: i) estimation of a survival model from a training dataset and the estimated survival model is then used for quantifying the disease risk of a person,

where the quantified disease risk is a linear combination of all disease predictors, and (ii) sequential monitoring of the quantified risks over time using a control chart. In the estimated survival model, more important covariates will receive more weights in the linear combination of the disease predictors. Thus, the effectiveness of the control chart is improved. Nonetheless, the aforementioned method still uses all disease predictors when defining disease risk. Intuitively, if certain disease predictors actually contain little useful information about the disease in concern, then they should be removed from disease screening. Based on this intuition, we propose a new method in this paper, in which variable selection by LASSO is incorporated in survival data modelling, so that the redundant disease predictors are deleted during survival model estimation. It will be shown that this new method is more effective than the original one by You and Qiu (2019) in various different cases.

The remaining parts of the article is organized as follows. In Section 2, the proposed model and its estimation for disease risk quantification will be introduced. In Section 3, some simulation studies will be presented to evaluate the performance of the proposed method. The proposed method will be demonstrated in a real-data example in Section 4. Finally, Section 5 will conclude the article with some discussions about certain future research topics.

2 Proposed Method

In this section, we describe the proposed disease screening method in detail. Our proposed method consists of two main steps. In the first step, a survival model is fitted from a training dataset that contains observations of the survival times and disease predictors of certain individuals. The fitted model can then be used to quantify people's disease risk at a given time. In this step, we will also discuss how to select important disease predictors by a LASSO variable selection method. In the second step, the quantified disease risk of a specific individual is monitored sequentially over time by a control chart. These two steps are discussed in detail in the following two parts.

2.1 Risk estimation and variable selection

Suppose that a training dataset containing observations of the longitudinal disease predictors and survival times of n individuals. The survival outcomes of the ith individual are described by

(δ_i, T_i), where T_i is the last-follow-up time and δ_i is the survival indicator with $\delta_i = 1$ indicating the occurrence of an disease at the last follow-up time T_i, and $\delta_i = 0$ otherwise. Following the notations of survival models in the literature, we use D_i to denote the true disease time and C_i to denote the censoring time. Then, the survival outcomes can be expressed as $T_i = \min\{D_i, C_i\}$ and $\delta_i = I(D_i \leq C_i)$. For simplicity of presentation, we will also use $R(t) = \{i : T_i \geq t\}$ to denote the set of all individuals who are at risk of disease at a given time t (i.e., they are still under monitoring in the study at time t). The q-dimensional longitudinal disease predictor of the ith individual is denoted as $\mathbf{x}_i(t)$, and it is repeatedly and sequentially observed at times $t_{i1}, \ldots, t_{im_i}$, where these observation times can be unequally spaced and $t_{im_i} = T_i$. Let $\lambda_i(t) = \lim_{dt \to 0} \mathrm{P}\{D_i \in [t, t+dt] | D_i \geq t\}/dt$ be the hazard function of the disease in question for the ith individual. Then, the following Cox proportional hazards model is assumed (cf., Klein and Moeschberger 1997):

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i(t)), \tag{1}$$

where $\boldsymbol{\beta}$ is a q-dimensional vector of coefficients and $\lambda_0(t)$ is the baseline hazard function. By using model (1), the linear combination $\boldsymbol{\beta}' \mathbf{x}_i(t)$ can measure the disease risk of the ith individual at time t, and it is denoted as $r_i(t)$. Namely, we define

$$r_i(t) = \boldsymbol{\beta}' \mathbf{x}_i(t).$$

To estimate model (1), You and Qiu (2019) suggested using the following kernel-smoothed likelihood:

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i(T_i))}{\sum_{l \in R(T_i)} \sum_{j=1}^{m_l} K_h(T_i - t_{lj}) \exp(\boldsymbol{\beta}' \mathbf{x}_l(t_{lj}))},$$

where $K_h(s) = K(s/h)/h$, $K(s)$ is a density kernel function, and $h > 0$ is a bandwidth. The use of kernel smoothing in Cox proportional hazards model is motivated by some existing works on estimating time-varying coefficients model in the literature (e.g., Cai and Sun 2003, Tian et al. 2005). The corresponding log-likelihood function is given by

$$l(\boldsymbol{\beta}) = \sum_{i:\delta_i=1} \left[\boldsymbol{\beta}' \mathbf{x}_i(T_i) - \log \left\{ \sum_{l \in R(T_i)} \sum_{j=1}^{m_l} K_h(T_i - t_{lj}) \exp(\boldsymbol{\beta}' \mathbf{x}_l(t_{lj})) \right\} \right]. \tag{2}$$

Then, $\boldsymbol{\beta}$ can be estimated by the maximizer of (2), denoted as $\tilde{\boldsymbol{\beta}}$, which can be obtained by using the Newton-Raphson algorithm.

So far, we assume that all disease predictors in $\mathbf{x}_i(t)$ have substantial impact on the disease risk. In reality, because we do not know which disease predictors are important and which are not,

we often include many potential disease predictors in $\mathbf{x}_i(t)$, to avoid important disease predictors being overlooked. Thus, some disease predictors in $\mathbf{x}_i(t)$ may not have much prediction power for the specific disease. This will be reflected in the regression coefficients in $\boldsymbol{\beta}$, which some of them could be 0 or small. However, the estimate $\tilde{\boldsymbol{\beta}}$ given by the partial log-likelihood function in (2) usually would not contain elements that are exactly 0. Thus, it cannot serve the purpose of variable selection. To properly select important disease predictors and exclude unimportant ones, we need to identify zero elements in the regression coefficient $\boldsymbol{\beta}$, which can be achieved by using the LASSO method (Tibshirani 1996). The main idea of LASSO is to add a penalty term on the regression coefficients to shrink the coefficients of unimportant disease predictors toward zero. In this paper, we choose to use the following L_1 adaptive LASSO penalty (cf., Zou 2006):

$$p_\gamma(\boldsymbol{\beta}) = \gamma \sum_{k=1}^{q} w_k |\beta_k|,$$

where γ is a non-negative regularization parameter, and $\mathbf{w} = (w_1, \ldots, w_q)'$ is a vector of adaptive weights. The adaptive weights $\{w_k\}$ can be simply chosen to be $1/|\tilde{\beta}_k|$, where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_k)$ is the estimate of $\boldsymbol{\beta}$ obtained from (2), as discussed above. The LASSO penalized estimate of $\boldsymbol{\beta}$ is then defined to be the minimizer of the following penalized log-likelihood function:

$$-l(\boldsymbol{\beta}) + p_\gamma(\boldsymbol{\beta}). \tag{3}$$

The above penalized log-likelihood function is not differentiable with respect to $\boldsymbol{\beta}$ at $\mathbf{0}$ and thus $\widehat{\boldsymbol{\beta}}$ cannot be obtained by the Newton-type optimization algorithms. Here, we can use the coordinate optimization algorithm discussed in Friedman et al. (2007) and Simon et al. (2011) to find the estimate, which is denoted as $\widehat{\boldsymbol{\beta}}$.

Note that there are some components in $\widehat{\boldsymbol{\beta}}$ that are exactly 0. These components and the corresponding disease predictors can then be deleted from the subsequent analysis. Without loss of generality, after these components and the corresponding disease predictors are deleted, the estimated regression coefficient vector and the disease predictor vector are still denoted as $\widehat{\boldsymbol{\beta}}$ and $\mathbf{x}_i(t)$. Then, the estimated disease risk of the ith individual at time t is defined to be

$$\widehat{r}_i(t) = \widehat{\boldsymbol{\beta}}' \mathbf{x}_i(t).$$

Similar to the univariate DySS method discussed in Qiu and Xiang (2014), we can characterize the regular pattern of the disease risk by a nonparametric longitudinal model, with the mean to be $\mu(t) = \mathrm{E}[r_i(t)|T_i \geq t]$ and the variance to be $\sigma^2(t) = \mathrm{Var}[r_i(t)|T_i \geq t]$. Here, we only assume that

$\mu(t)$ and $\sigma^2(t)$ are two smooth functions of time, and they can be estimated by the local linear kernel smoothing procedure, as discussed in Qiu and Xiang (2014) and Xiang et al. (2013). The corresponding local linear kernel estimates of $\mu(t)$ and $\sigma^2(t)$ are given by

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

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

where h_μ and h_σ are two bandwidths that could be different from h used in (2), $\widehat{\epsilon}_i(t_{ij}) = \widehat{r}_i(t_{ij}) - \widehat{\mu}(t_{ij})$, and

$$W_{l,h}(t) = \frac{1}{n}\sum_{i\in R(t)}\sum_{j=1}^{m_i} K_h(t_{ij}-t)\Big(\frac{t_{ij}-t}{h}\Big)^l,$$

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

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

As a side note, in the above model, the regular disease risk pattern is characterized by the first and second moments $\mu(t)$ and $\sigma^2(t)$. Alternatively, we can characterize the regular disease risk pattern by the entire distribution of $r_i(t)$. To this end, let $F(y;t) = \mathrm{P}(r_i(t) \le y | T_i \ge t)$ be the conditional distribution function of the disease risk at time t. For given values of y and t, we can use the local linear kernel smoothing method to estimate this conditional distribution, as discussed in Fan et al. (1996) and Yu and Jones (1998). By following their ideas, we can consider minimizing the following objective function:

$$\sum_{i\in R(t)}\sum_{j=1}^{m_i}\Big[\Psi_{h_\psi}\big(\widehat{r}_i(t_{ij}) - y\big) - \alpha_0 - \alpha_1(t_{ij}-t)\Big]^2 K_{h_F}(t_{ij}-t),$$

where $\Psi(y)$ is a suitable kernel cumulative distribution function, h_ψ and h_F are two bandwidths, and $\Psi_h(y) = \Psi(y/h)$. Then, the estimate $\widehat{F}(y;t)$ of $F(y;t)$ can be defined by the minimizer with respect to α_0 in the above minimization problem, which has the expression

$$\widehat{F}(y;t) = \frac{S_0(y;t)W_{2,h_F}(t) - S_1(y;t)W_{1,h_F}(t)}{W_{0,h_F}(t)W_{2,h_F}(t) - W_{1,h_F}(t)^2}, \tag{6}$$

where for $l = 0, 1$,

$$S_l(y;t) = \frac{1}{n}\sum_{i\in R(t)}\sum_{j=1}^{m_i} K_{h_F}(t_{ij}-t)\Big(\frac{t_{ij}-t}{h_F}\Big)^l \Psi_{h_\psi}\big(\widehat{r}_i(t_{ij}) - y\big).$$

This idea will not be further explored in this paper, and will be studied in our future research.

In (3)-(6), there are several bandwidths to use. To determine the bandwidth h used in (3) for estimating $\boldsymbol{\beta}$, we can use the leave-one-out cross-validation (CV) criterion that is based on the martingale residuals (cf., Tian et al. 2005, You and Qiu 2019). The selected bandwidth can be calculated by minimizing the following function of h

$$CV_{\beta}(h) = \sum_{i=1}^{n} PE_i(h),$$

where

$$PE_i(h) = \left[\delta_i - \sum_{\substack{k \neq i,\, \delta_k = 1 \\ T_k \leq T_i}} \frac{\sum_{j=1}^{m_i} K_h(T_k - t_{ij}) \exp(\tilde{\boldsymbol{\beta}}'_{-i} \mathbf{x}_i(t_{ij}))}{\sum_{d \neq k, d \in R(T_k)} \sum_{j=1}^{m_d} K_h(T_k - t_{dj}) \exp(\tilde{\boldsymbol{\beta}}'_{-i} \mathbf{x}_d(t_{dj}))} \right]^2,$$

$\boldsymbol{\beta}_{-i}$ is the estimate of $\boldsymbol{\beta}$ when the ith individual is excluded from model estimation, and PE_i is the square of some estimate of the integrated martingale residual. To choose the regularization parameter γ in the LASSO penalty in (3), we propose to use the Akaike information criterion (AIC) (cf., Akaike 1992, Tibshirani 1997). Let $c(\boldsymbol{\beta})$ be the number of non-zero elements of the vector $\boldsymbol{\beta}$. Then, the AIC of the modified Cox partial likelihood is defined as

$$AIC(\gamma) = -2l(\widehat{\boldsymbol{\beta}}) + 2c(\widehat{\boldsymbol{\beta}})$$

where $\widehat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ with the regularization parameter being γ. The regularization parameter γ is then chosen to be the minimizer of $AIC(\gamma)$. The bandwidths h_μ and h_σ in (4) and (5) can be chosen using the leave-one-out CV procedure discussed in Qiu and Xiang (2014). In this chapter, all kernel functions are chosen to be the Epanechnikov kernel function (Epanechnikov 1969).

2.2 Online disease risk monitoring

To monitor the quantified disease risk of a new individual, assume that the disease predictors are sequentially observed at times $t_1^*, t_2^*, \ldots$, and the corresponding observations are $\mathbf{x}(t_1^*), \mathbf{x}(t_2^*), \ldots$. For simplicity of presentation, we further assume that $t_1^*, t_2^*, \ldots$ are multiplications of a basic time ω. Thus, we can write $t_j^* = n_j^* \omega$, for $j \geq 1$. When the disease risk pattern is characterized by the estimated mean and variance function $\widehat{\mu}(t)$ and $\widehat{\sigma}^2(t)$, we can define the standardized value of the estimated disease risk as

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

To monitor the quantified disease risks of the new individual and detect an undesirable upward shift in the longitudinal pattern of the disease risk when observations are sequentially obtained, SPC charts can be used. To this end, we consider using the following upward EWMA chart, based on the exponential smoothing idea that was discussed in Wright (1986) and Qiu et al. (2018) to account for irregularly spaced observation times:

$$E_1 = V_1 \widehat{e}(t_1^*), \tag{7}$$

$$E_j = (1 - V_j) E_{j-1} + V_j \widehat{e}(t_j^*), \tag{8}$$

where $V_1 = 1 - (1-\lambda)^{\bar{\Delta}}$, $V_j = V_{j-1}/[(1-\lambda)^{n_j^* - n_{j-1}^*} + V_{j-1}]$, λ is a weighting parameter in $[0, 1)$, and $\bar{\Delta}$ is an estimate of the mean of $n_j^* - n_{j-1}^*$ obtained from the IC dataset. The chart gives a signal at time t_j^* if

$$E_j > \rho$$

where $\rho > 0$ is a control limit. It should be pointed out that the upward chart is considered here because we are mainly concerned about upward shifts in disease risk in the current disease screening problem. In other problems, downward or two-sided charts might be more appropriate. Also, the observation times $t_1^*, t_2^*, \ldots$ are often unequally spaced in disease screening applications, and the above EWMA chart can accommodate the unequally spaced observation times well. With other types of control charts (e.g., CUSUM), we still do not know how to accommodate this properly in their chart construction.

The performance of control charts are traditionally evaluated by the average run length (ARL), which is the average number of collected observations before triggering a signal. When observation times are unequally spaced, a more sensible measure is the average time to signal (ATS) (cf., Qiu and Xiang 2014). In disease monitoring problems, there is also interest in the receiver operating characteristics (ROC) of monitoring schemes in terms of sensitivity and specificity, which are for evaluating whether the monitoring scheme can correctly identify patients who may or may not have the disease in the future. Recently, Qiu et al. (2019) has proposed a new measure, called process monitoring ROC curve, which attempts to combine the ATS measure and the ROC measures.

The control limit ρ is usually chosen such that the nominal IC ATS value is fixed at a given level when the monitoring schemes are applied to an IC dataset. To accommodate the within-subject data correlation, the block bootstrap procedure discussed in Qiu and Xiang (2014) can be used for searching for the control limit. When there are enough training data, we can split them into two

parts. The first part can be used for estimating model (1) to describe the regular pattern of the disease risk, and the IC individuals in the second part can be used for determining the control limit ρ. To expedite the searching algorithm, we can use the bisection method as discussed in Qiu (2014, Section 4.2).

3 Simulation Study

Simulations were conducted to evaluate the numerical performance of the proposed method. We used a simulated training data set of $n = 500$ individuals to estimate the regular disease risk pattern. The whole design interval $[0, 1]$ is discretized in to 1000 basic time units. We considered three different cases with the dimension of covariates being $q = 10, 20$ and 30. The processes of $x_{ik}(t)$ are generated from the following random process model

$$x_{ik}(t) = -\sin(\pi t + \pi u_{ik1}) + 0.5\cos(10\pi t + 10\pi u_{ik2}) + \epsilon_{ik}(t), \quad \text{for } i = 1, \ldots, n, k = 1, \ldots, q, \quad (9)$$

where $\{u_{ik1}, u_{ik2}\}$ are independent realizations from the uniform $[0, 1]$ distribution, $\epsilon_{ik}(t)$ are generated from the Ornstein-Uhlenbeck processes with $d\epsilon_{ik}(t) = -\theta(m_{ik} - \epsilon_{ik}(t))dt + \sigma dW_{ik}(t)$, $W_{ik}(t)$ are independent realizations from the Wiener process, $\theta = 50$, $\sigma = 20$, $\epsilon_{ik}(0) \sim N(0, 0.2^2)$, and the random mean vectors $\mathbf{m}_i = (m_{i1}, \ldots, m_{iq})'$ are realizations from the multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix being

$$\begin{pmatrix} 0.5 & 0.1 & \cdots & 0.1 \\ 0.1 & 0.5 & \cdots & 0.1 \\ \vdots & \vdots & \ddots & \vdots \\ 0.1 & 0.1 & \cdots & 0.5 \end{pmatrix}.$$

The baseline hazard function in model (1) is chosen to be $\lambda_0(t) = 0.25$. The true regression coefficients $\boldsymbol{\beta}$ are sparse with the first three dimensions being 0.2 and all the remaining dimensions being 0, i.e., $\boldsymbol{\beta} = (0.2, 0.2, 0.2, 0, \ldots, 0)'$. In each simulation, we estimate the regular disease risk pattern using the simulated dataset of $n = 500$ individuals, and then determine the control limit ρ from another simulated dataset of 500 individuals. The control limit ρ here is chosen such that the nominal IC ATS is 370. Then, the proposed monitoring scheme is applied to simulated data of 10,000 new individuals to evaluate its performance. All the results presented in this section are based on 1000 replicated simulations.

We first present some results about the variable selection method using LASSO. In Table 1, we displayed the mean-squared errors of estimated regression coefficients for different dimensions $q = 10, 20, 30$. From the table, we can see that when there is a substantial percentage of zeros in the true regression coefficients, the LASSO penalized estimate $\widehat{\boldsymbol{\beta}}$ has a smaller mean-squared-error (MSE), compared to the MSE of the ordinary estimate $\tilde{\boldsymbol{\beta}}$ in this example. The relative efficiency of $\widehat{\boldsymbol{\beta}}$ with respect to $\tilde{\boldsymbol{\beta}}$, defined as the ratio of their MSE values, decreases as the dimension q increases. The implication is that when many covariates are unrelated to the disease risk, applying the LASSO penalty can indeed improve the efficiency of parameter estimates.

Table 1: MSE of LASSO penalized estimate $\widehat{\boldsymbol{\beta}}$ and unpenalized estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and their corresponding standard errors (in parentheses). The relative efficiency of $\widehat{\boldsymbol{\beta}}$ with respect to $\tilde{\boldsymbol{\beta}}$ is the ratio of the their MSE values, and the standard errors of relative efficiency is obtained by the delta method.

	$q = 10$	$q = 20$	$q = 30$
MSE of $\tilde{\boldsymbol{\beta}}$	0.0140 (0.0002)	0.0283 (0.0003)	0.0424 (0.0004)
MSE of $\widehat{\boldsymbol{\beta}}$	0.0084 (0.0002)	0.0115 (0.0003)	0.0139 (0.0003)
Relative efficiency of $\widehat{\boldsymbol{\beta}}$	0.6000 (0.0072)	0.4077 (0.0061)	0.3265 (0.0051)

Next, we present some results about the proposed online monitoring scheme. To examine whether the proposed method can effectively detect distributional shifts that lead to an increased disease risks, we considered a shift of $\mathbf{x}(t)$ in the direction of $\boldsymbol{\beta}$, namely,

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

where $\mathbf{x}(t)$ is simulated from model (9). The out-of-control ATS values of three different methods are presented in Figure 1, where DySS denotes the multivariate DySS by Qiu and Xiang (2015), NoSelection denotes the original risk monitoring method by You and Qiu (2019) without using the LASSO variable selection, and Selection is the proposed method in this paper. From the figure, we can see that (i) the proposed method Selection has the best performance among all three methods, (ii) DySS has the worst performance among all three methods, and (iii) the improvement from NoSelection to Selection is more pronounced as the dimensionality of $\mathbf{x}(t)$ increases.

We then compare the three different methods when they are applied to individuals in a simulated training dataset, using the metrics of true positive rate (TPR) and false positive rates (FPR). Here, TPR is defined to be the percentage of individuals receiving signals among all diseased people,

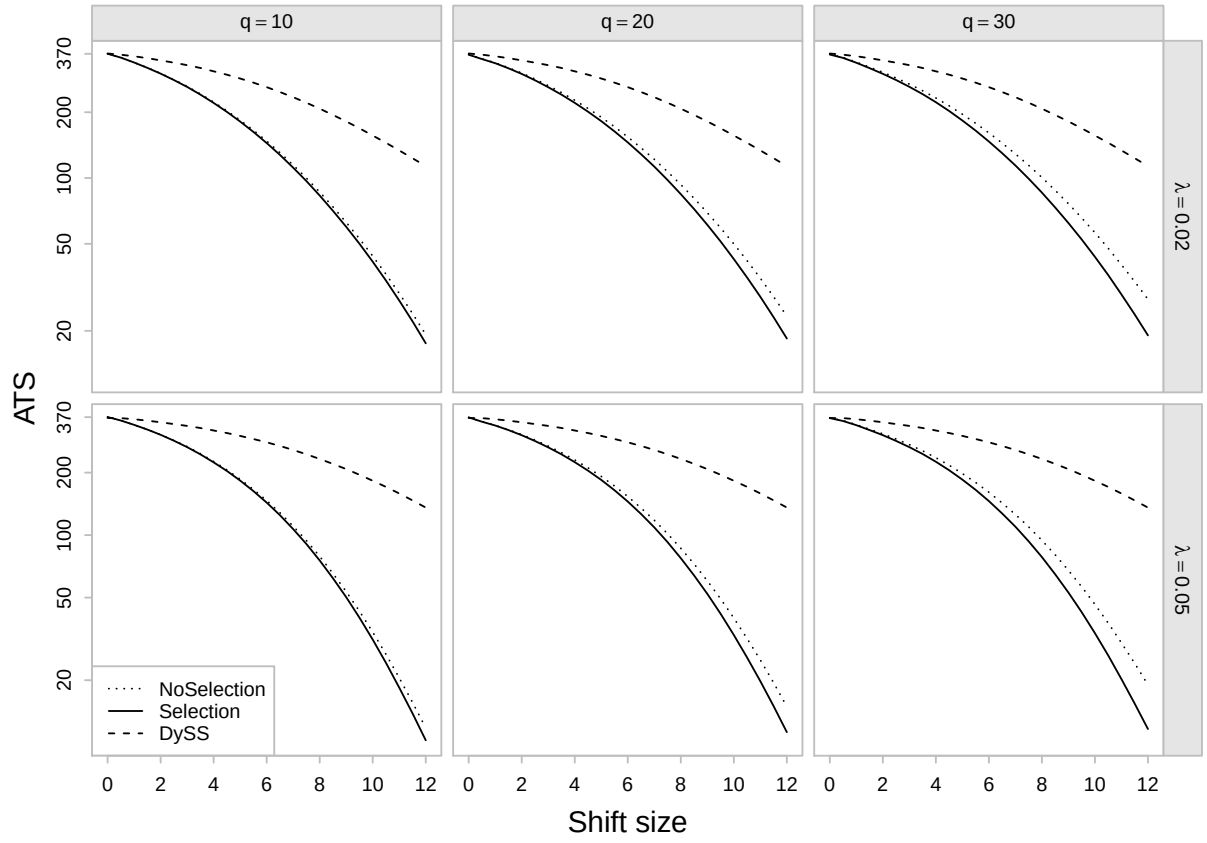

Figure 1: OC ATS values of different monitoring methods when $\lambda = 0.02, 0.05$ and $q = 10, 20, 30$. NoSelection is the method by You and Qiu (2019) where $\boldsymbol{\beta}$ is estimated from (2) without using the LASSO penalty. Selection is the proposed method in this paper. DySS is the multivariate DySS by Qiu and Xiang (2015).

and FPR is defined to be the percentage of individuals receiving signals among all non-diseased people. The results are presented in Table 2, from which we can see that i) DySS tends to have a high false positive rate in all cases considered, and ii) Selection and NoSelection have similar TPR values in most scenarios considered here, but the FPR values of NoSelection are lower than those of Selection in all cases considered here.

Table 2: TPR and FPR values of different monitoring methods when $\lambda = 0.02, 0.05$ and $q = 10, 20, 30$. Numbers in parentheses are the corresponding standard errors. NoSelection is the risk monitoring method by You and Qiu (2019) where $\boldsymbol{\beta}$ is estimated from (1) without using the LASSO penalty. Selection is the proposed risk monitoring method where $\boldsymbol{\beta}$ is estimated from (3). DySS is the multivariate DySS by Qiu and Xiang (2015).

	$q = 10$		$q = 20$		$q = 30$	
	TPR	FPR	TPR	FPR	TPR	FPR
$\lambda = 0.02$						
NoSelection	0.427 (0.002)	0.457 (0.003)	0.432 (0.002)	0.473 (0.003)	0.430 (0.002)	0.468 (0.002)
Selection	0.427 (0.002)	0.453 (0.003)	0.432 (0.002)	0.465 (0.003)	0.433 (0.002)	0.453 (0.003)
DySS	0.483 (0.002)	0.640 (0.002)	0.483 (0.002)	0.640 (0.002)	0.483 (0.002)	0.639 (0.002)
$\lambda = 0.05$						
NoSelection	0.426 (0.003)	0.467 (0.003)	0.431 (0.003)	0.482 (0.003)	0.428 (0.003)	0.478 (0.003)
Selection	0.426 (0.002)	0.463 (0.003)	0.430 (0.003)	0.471 (0.003)	0.430 (0.003)	0.462 (0.003)
DySS	0.392 (0.003)	0.509 (0.003)	0.392 (0.003)	0.508 (0.003)	0.392 (0.003)	0.509 (0.003)

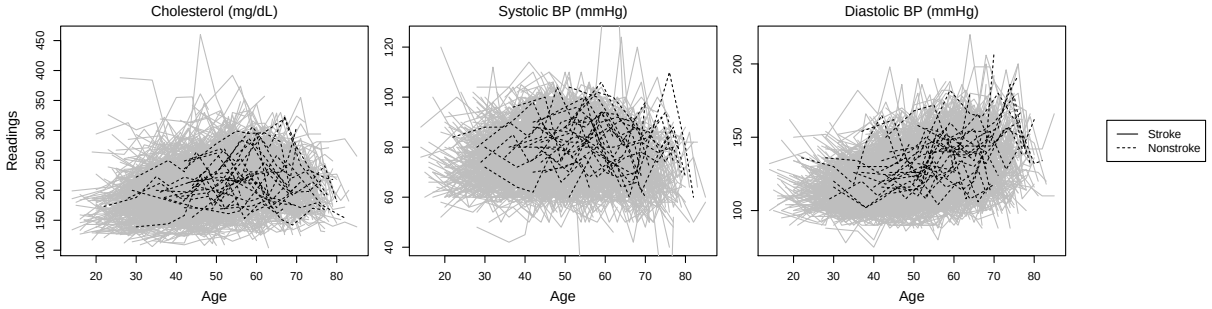

Figure 2: Cholesterol levels, systolic blood pressure readings and diastolic blood pressure readings of 1,055 participants of the Framingham heart study. Gray solid lines are longitudinal observations of 1,028 non-stroke participants, while black dashed lines are longitudinal observations of 27 stroke participants.

4 Real Data Example

In this section, we apply the proposed method to a real data example from the Framingham heart study. In this study, participants are regularly examined for risk factors of cardiovascular diseases. The data set contains observations of the cholesterol levels, systolic blood pressures and diastolic blood pressures of 1,055 participants. Each participant was followed for 7 times at their different ages. Among the 1,055 participants, 27 of them had strokes at least once during the study. This data

set is displayed in Figure 2. To implement and evaluate the proposed method, we randomly partition the original dataset into training and test datasets. The training dataset contains approximately two-thirds of all the participants, among which 18 of them had strokes during the study and 686 did not have any strokes. This training dataset is then used for estimating the regular disease risk pattern using (2)-(5). The test dataset contains approximately 1/3 of all the participants, among which 9 had strokes and 342 did not have any strokes. The test dataset is then used for evaluating the numerical performance of the proposed method. Its weighting parameter λ is chosen to be 0.2 and the nominal IC ATS is set to be 10 years.

The estimate of $\boldsymbol{\beta}$ by (2) without the LASSO penalty is $\tilde{\boldsymbol{\beta}} = (-0.0013, 0.0178, 0.0099)'$, and the LASSO estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (0.0000, 0.0169, 0.0092)'$ where the parameter γ is chosen to be 0.05. We can see that the first dimension of the LASSO estimate has been shrunk to 0. We then compare the performance of the three monitoring methods: NoSelection, Selection and DySS. A summary of the results is presented in Table 3. From the table, we can see that i) all the three methods considered here correctly gave signals to 8 out of 9 stroke patients in the test dataset, ii) NoSelection gives 132 signals to 342 non-stroke patients, and Selection gives only 9 less signals to the non-stroke patients, and iii) DySS has the worst performance because it gives more signals to the non-stroke patients than each of the other two methods. The three types of charts for monitoring the 9 stroke patients in the test dataset are shown in Figure 3.

Table 3: Number of signals when different methods are used to monitor patients in the test dataset.

	DySS		NoSelection		Selection	
	Signal	No signal	Signal	No signal	Signal	No signal
Stroke Patients	8	1	8	1	8	1
Non-Stroke Patients	167	175	132	210	123	219

5 Concluding remarks

In this chapter, we presented an improved version of the disease risk monitoring method suggested by You and Qiu (2019). The major contribution of the improved version is that variable selection is used when quantifying disease risks, in order to reduce variability of the quantified disease risks.

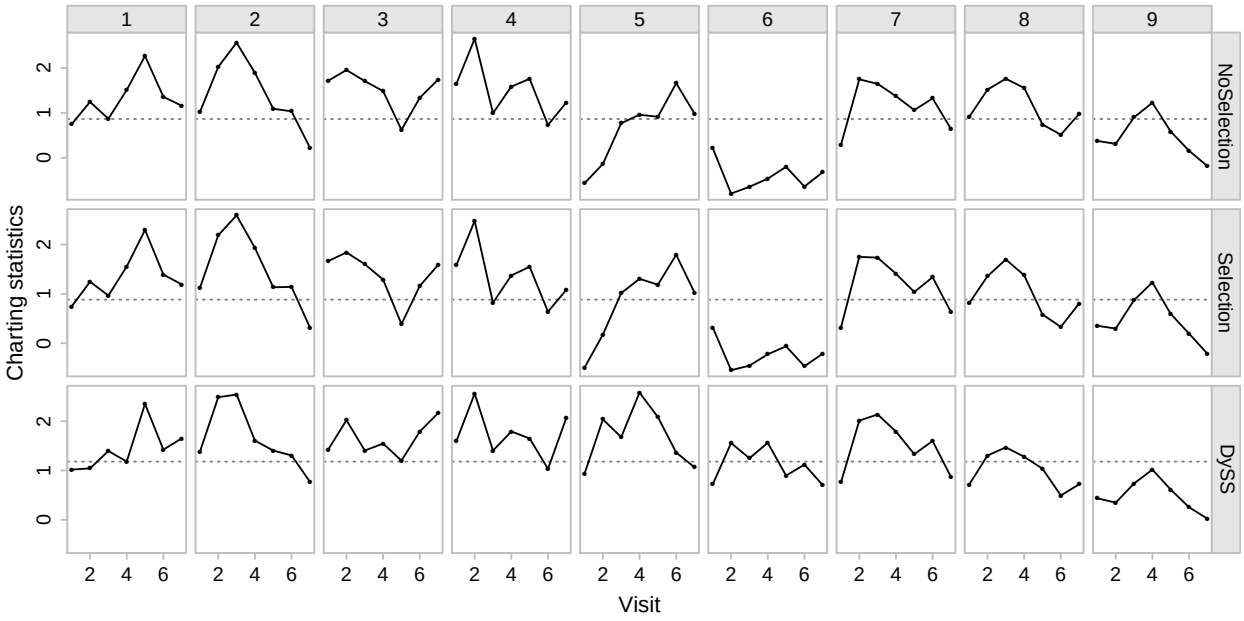

Figure 3: Charting statistics of the three types of charts for monitoring the 9 stroke patients in the test dataset. Horizontal dotted lines are the control limits.

Variable selection is achieved by using the LASSO penalty to reduce the dimensionality of the disease predictors in the related survival model. Through numerical simulations and a real data example, we have shown that when many disease predictors are included in the survival model, implementing variable selection before monitoring the quantified disease risks can often improve the performance of disease risk monitoring.

Our proposed method still has its own limitations and there are many issues to be addressed in the future research. For instance, in real-life disease screening practice, it is quite common that patients may miss some medical tests or have some incomplete medical examinations during a clinic visit. Future research is needed to extend existing methods to allow for missing data of different types. Also, the effect of disease predictors can be time-varying. Though You and Qiu (2019) has provided a method for estimating time-varying regression coefficients, the variable selection problems in a time-varying-effect model will be much more challenging than the problem considered here, which has not properly discussed yet. Finally, the proposed variable selection method is based on the L_1 adaptive LASSO penalty. In the literature, there are a series of alternative penalized regression methods for variable selection. For example, one may consider the alternative penalty functions like the elastic net (Zou and Hastie 2005) and the smoothly clipped absolute deviation

penalty (Fan and Li 2001) to reduce the bias of the LASSO estimates. When disease predictors come from many different groups, one may consider using the group LASSO (Yuan and Lin 2006) to select some groups of disease predictors for quantifying disease risks. It is of interest to study all these variable selection methods in the dynamic disease screening and monitoring problems in the future research.

Acknowledgments: The authors thank the editors for the invitation of this contribution to the editted book. One referee provided some comments about the paper for improvements, which is greatly appreciated. This research is supported in part by an NSF grant in USA.

References

Akaike, H. (1992), "Information theory and an extension of the maximum likelihood principle," In *Second International Symposium on Information Theory Proceeding*, Volume 1, pages 610–624.

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.

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

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J., Yao, Q., and Tong, H. (1996), "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," *Biometrika*, **83**, 189–206.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R. (2007), "Pathwise coordinate optimization," *The Annals of Applied Statistics*, **1**, 302–332.

Klein, J.P., and Moeschberger, M.L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer.

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.

Mendis, S., Puska, P., and Norrving, B. (2011), "Global atlas on cardiovascular disease prevention and control," World Health Organization. https://apps.who.int/iris/handle/10665/44701.

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.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, **39**, 1–13.

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

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Tibshirani, R. (1997), "The LASSO method for variable selection in the Cox model," *Statistics in Medicine*, **16**, 385–395.

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), "Nonparametric regression analysis of multivariate longitudinal data," *Statistica Sinica*, **23**, 769–789.

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.

You, L. and Qiu, P. (2019), "An effective method for online disease risk monitoring," *Technometrics*, DOI: 10.1080/00401706.2019.1625813.

Yu, K. and Jones, M. (1998), "Local linear quantile regression," *Journal of the American Statistical Association*, **93**, 228–237.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of The Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.