A Dynamic Screening System for Early Detection of Multiple Interconnected Events

Zibo Tian and Peihua Qiu
Department of Biostatistics, University of Florida
2004 Mowry Road, Gainesville, FL 32610

Abstract

Sequential monitoring of temporal processes is essential in various fields for the early detection of critical events, such as airplane failures or disease occurrences. This task typically involves sequential decision-making across many related processes. Conventional statistical process control charts are generally designed for detecting an event of interest in a single process and thus inadequate for this task. Recently, several versions of the dynamic screening system (DySS) have been developed to monitor a population of processes. However, these DySS methods focus solely on the early detection of a single event (e.g., the occurrence of a particular disease) and cannot handle scenarios where multiple distinct events are involved. In practice, detecting multiple events is important and common. For example, we may be concerned with several medical conditions in patients or different types of failures in airplanes. This problem is significantly more challenging than single-event detection, as different events are often interconnected and can occur at different times. To address this complexity, we propose the concept of conditional risks for multiple events, utilizing single-index multinomial logistic regression modeling. Based on this, we develop a new DySS method for the early detection of multiple events by sequential monitoring of the related conditional risks. Numerical studies demonstrate that this method provides an effective analytical tool for the early detection of multiple interconnected events.

Key Words: Conditional risks; Dynamic screening system; Multiple diseases; Multivariate longitudinal data; Online process monitoring; Single-index model.

1 Introduction

In numerous fields, sequential monitoring of temporal processes is critical for the early detection of adverse events, such as disease outbreaks and mechanical failures, that can lead to health, economic, and other damages. For example, in manufacturing, real-time monitoring of machinery performance helps identify early signs of mechanical failure, minimizing unexpected downtime and costly repairs. Across various applications, multiple adverse events, such as different diseases or types of mechanical failures, often arise

from shared root causes or risk factors, making it more efficient to detect them collectively (Salisbury et al., 2011). For instance, medical conditions like heart disease, hypertension, and stroke are frequently linked through common risk factors like high cholesterol levels and smoking habits. Similarly, in cybersecurity, data breaches and other network vulnerabilities often stem from sources such as software exploits and phishing attacks (Tsochev et al., 2020). This paper aims to develop an analytical tool that effectively detects multiple interconnected events within a population of sequentially observed processes.

In the literature, early detection of a single adverse event in a population of sequentially observed processes can be achieved using the dynamic screening system (DySS), developed by Qiu and Xiang (2014, 2015). The DySS method identifies the adverse event by sequentially monitoring the observed risk factors for each individual process in three main steps. First, the *regular longitudinal pattern* of event risk factors is estimated using an in-control (IC) dataset collected in advance, containing observed risk factors from a group of well-functioning subjects. This step employs nonparametric longitudinal modeling approaches. Second, for a new subject under monitoring, the subject's observed risk factors at the current observation time are standardized based on the estimated regular longitudinal pattern. Finally, the standardized data are evaluated using a built-in statistical process control (SPC) chart to determine if the related process is IC or out-of-control (OC) at the current time point. Several modifications and generalizations of this method address issues like serial correlation, high dimensionality, and other data complexities. For a comprehensive description, see Qiu (2024).

In directly monitoring observed risk factors, the DySS approach traditionally treats all risk factors equally, which limits its ability to account for the heterogeneous impacts of different factors on the event of interest. Additionally, joint monitoring of multiple risk factors may become inefficient as dimensionality increases. To address these limitations, You and Qiu (2020) proposed an alternative DySS approach consisting of two main steps: i) A survival model is applied to observed times of the event in a training dataset, defining *event risk* as a linear combination of relevant risk factors, and ii) For a new subject under monitoring, event risk at each observation time is first calculated based on its observed risk factors and then monitored using a control chart. This approach enables univariate monitoring of event risk, with the heterogeneous impact of different risk factors captured through the estimated coefficients in the linear combination defining event risk. An improved version of this method was suggested in Qiu and You (2022).

In the literature, there are some other recent methodologies for online monitoring of dynamic processes. For instance, Liu et al. (2023) developed an in-profile monitoring (INPOM) control chart for monitoring a sequence of dynamic processes. Zhang et al. (2025) developed another INPOM chart based on generalized

likelihood ratio testing. Ashraf et al. (2021) proposed an EWMA chart for disease screening based on sequential monitoring of disease risk factors for individual patients. Motivated by the application of passenger flow surveillance, Li et al. (2024, 2025) developed dynamic modeling and online monitoring methods for analyzing tensor data streams.

All existing methods discussed above focus on detecting a single adverse event (e.g., a specific disease like stroke). To our knowledge, no existing methods are designed to detect multiple events in a population of sequentially observed processes. This paper aims to address this methodological gap. Detecting multiple adverse events poses a significantly greater challenge than detecting a single event, as multiple events may not necessarily occur at the same time point. Beyond detecting individual events, it is also valuable to detect events conditional on the presence of others, which introduces the concept of *conditional risks*, the risk of some events given the presence of others. For example, in multimorbidity research, a patient who has developed diabetes faces a significantly increased conditional risk of cardiovascular complications (Matheus et al., 2013). In finance, the default of one company within a supply chain can notably increase the risk of defaults in related businesses, underscoring the importance of monitoring conditional risks to understand cascading effects (Scheibe and Blackhurst, 2018).

In this paper, we propose a novel method for monitoring conditional risks of related events, offering a comprehensive tool for risk management and monitoring in applications requiring detection of multiple adverse events. To quantify the risk of a single event, single-index modeling is often employed to create a composite index that summarizes the impact of various risk factors on the likelihood of the event (e.g., Carroll et al., 1997; Cui et al., 2011). Related approaches have been extended to cases involving longitudinal observations (e.g., Chowdhury and Sinha, 2015; Yi et al., 2009) and scenarios with multiple events and longitudinal observations (Tian and Qiu, 2024a). While these methods allow for calculating marginal risks, they do not support the computation of conditional risks.

To address this limitation, we propose a single-index multinomial logistic regression model to analyze a pre-collected training dataset, deriving formulas to compute all possible conditional risks for the events of interest. For a given conditional risk that quantifies the risk of one set of events occurring given the presence of another set, its regular longitudinal pattern is first estimated using data from subjects in the training set who experienced the latter set of events but not the former. For a new subject under monitoring who meets this condition, its conditional risk at each observation time is then computed based on its observed risk factors. A control chart is used to monitor these values over time by comparing the subject's observed longitudinal pattern of conditional risk with the estimated regular pattern. The chart makes its decision

based on the cumulative difference between the two patterns up to the current time point. In other words, the control chart sequentially tests the null hypothesis that the subject's conditional risks up to the current time point follow the regular longitudinal pattern, continuing until a signal is triggered. Numerical studies show that this method provides an effective analytical tool for detecting multiple events across a variety of scenarios.

In contemporary machine learning literature, several approaches, such as the REverse Time AttentIoN model (RETAIN) and Random Forests for Survival, Longitudinal, and Multivariate data analysis (RF-SLAM), achieve dynamic risk modeling by incorporating time-varying covariates, historical event occurrences, and nonlinear interactions (Choi et al., 2016; Wongvibulsin et al., 2020). In the context of disease risk prediction, RETAIN employs a reverse-time attention mechanism to highlight the most influential past visits for forecasting future disease onset, whereas RF-SLAM extends random forests to handle time-to-event outcomes using time-varying covariates. While these methods provide accurate estimates of time-dependent marginal and conditional risks, they typically identify high-risk cases using *ad hoc* thresholding or observed upward trends. As such, they lack a formal sequential monitoring framework for controlling false-alarm rates over time. In contrast, our proposed method integrates risk estimation with SPC charts, enabling real-time surveillance of conditional risk trajectories in new subjects. This integration allows for proper control of the average time to false signals, thereby facilitating the prompt and reliable detection of emerging adverse events.

The remainder of the paper is organized as follows: Section 2 details each main step of the proposed method, Section 3 presents simulation results to assess its numerical performance, and Section 4 provides a real-data example to illustrate its application. Several concluding remarks are given in Section 5, with additional technical details and numerical results presented in Appendix.

2 Proposed Method

The proposed method involves the following key steps. First, a single-index multinomial logistic regression model is fitted to a training dataset, defining all possible conditional risks for the events of interest (e.g., diseases, machinery failures). Second, The regular longitudinal pattern of a given conditional risk is estimated using the observed data from appropriately defined "well-functioning" subjects within the training dataset. Third, for each new subject under monitoring, the conditional risk of interest is computed at the

current observation time, standardized, and decorrelated from its values at previous time points using the estimated regular longitudinal pattern. Finally, an SPC chart is constructed to monitor the standardized and decorrelated conditional risk, enabling the determination of the event status at the current time. A detailed explanation of each step is provided in the subsections below.

2.1 Single-index multinomial logistic regression modeling

Suppose we have q adverse events in concern, there are a total of p risk factors of these events, and a training dataset containing their longitudinal observations of M subjects is collected in advance. Let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijq})^T$ and $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ be the observations of the binary responses (i.e., status of q events) and covariates (i.e., risk factors) of the ith subject at time $t_{ij} \in [T_1, T_2]$, where $j = 1, \ldots, m_i$, $i = 1, \ldots, M$, and $[T_1, T_2]$ is a given time interval. For each i and j, there are a total of $q^* = 2^q$ different values for $\mathbf{Y}_{ij}$, denoting q^* possible event statuses. Let $Z_{ijk} = 1$ denote the presence of the kth event and 0 otherwise, for $k = 1, \ldots, q^*$. Without loss of generality, let $Z_{ijq^*} = 1$ denote the case when none of the q events are present at time t_{ij}.

The observed data in the training dataset are assumed to follow a single-index multinomial logistic regression model: for $j = 1, \ldots, m_i, i = 1, \ldots, M$, and $k = 1, \ldots, (q^* - 1)$,

$$\log\left\{\frac{\mathbb{P}(Z_{ijk} = 1|\mathbf{X}_{ij}, \mathbf{b}_i)}{\mathbb{P}(Z_{ijq^*} = 1|\mathbf{X}_{ij}, \mathbf{b}_i)}\right\} = \psi_k\left(\boldsymbol{\beta}_k^T \mathbf{X}_{ij}\right) + b_{ik}, \tag{1}$$

where $\{\psi_k(\cdot)\}$ are smooth but unknown link functions, and $\{\boldsymbol{\beta}_k\}$ represent single-index coefficient vectors, each satisfying the conditions $\boldsymbol{\beta}_k^T \boldsymbol{\beta}_k = 1$ and having nonnegative first element. For Model (1), the vector $\mathbf{b}_i = (b_{i1}, \ldots, b_{i(q^*-1)})^T$ captures random-effects for the ith subject to account for both between-subject variation and within-subject correlation. It is routinely assumed that $\{\mathbf{b}_i, i = 1, \ldots, M\}$ are independent and identically distributed (i.i.d.) following the common distribution $N_{(q^*-1)}(\mathbf{0}, \boldsymbol{\Sigma}_b)$, where $\boldsymbol{\Sigma}_b$ is a $(q^* - 1) \times (q^* - 1)$ covariance matrix. For simplicity, we assume that $\boldsymbol{\Sigma}_b = \mathbf{D}_b^{1/2} \mathbf{R}_b(\rho_b) \mathbf{D}_b^{1/2}$, where $\mathbf{D}_b = \text{diag}\{\sigma_{b1}^2, \ldots, \sigma_{b(q^*-1)}^2\}$ is a diagonal matrix of variances, and $\mathbf{R}_b(\rho_b)$ is a correlation matrix with 1s on the diagonal and ρ_b for all off-diagonal elements. While the single indices $\{\boldsymbol{\beta}_k^T \mathbf{X}_{ij}\}$ are linear combinations of the covariates, they allow for different coefficient vectors across events, enhancing flexibility. Together with the inclusion of the link functions $\{\psi_k(\cdot)\}$, this makes Model (1) quite adaptable to varying data structures.

Model (1) can be estimated by using Monte Carlo expectation-maximization (EM) algorithm and kernel local smoothing methods, as described below. First, the log-likelihood function of Model (1) can be written

as:

$$\begin{aligned} & l(\boldsymbol{\theta}; \mathbf{Z}, \mathbf{b}) \\ = & \log\{f(\mathbf{Z}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\}) f(\mathbf{b}|\boldsymbol{\Sigma}_b)\} \\ = & \sum_{i=1}^{M}\sum_{j=1}^{m_i}\sum_{k=1}^{q^*} Z_{ijk}\log(\mu_{ijk}) + \sum_{i=1}^{M}\left\{-\frac{q^*-1}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_b| - \frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\right\}, \end{aligned}$$

where for $k = 1, \ldots, (q^*-1)$,

$$\begin{aligned} \mu_{ijk} = \mathbb{P}(Z_{ijk} = 1|\mathbf{X}_{ij}, \mathbf{b}_i) &= \frac{\exp\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) + b_{ik}\}}{1 + \sum_{s=1}^{q^*-1}\exp\{\psi_s(\boldsymbol{\beta}_s^T\mathbf{X}_{ij}) + b_{is}\}}, \\ \mu_{ijq^*} = \mathbb{P}(Z_{ijq^*} = 1|\mathbf{X}_{ij}, \mathbf{b}_i) &= \frac{1}{1 + \sum_{s=1}^{q^*-1}\exp\{\psi_s(\boldsymbol{\beta}_s^T\mathbf{X}_{ij}) + b_{is}\}}. \end{aligned}$$

Let $\mathbf{Z}_{ij\cdot} = (Z_{ij1}, \ldots, Z_{ijq^*})^T$ and $\mathbf{Z}_{i\cdot\cdot} = (\mathbf{Z}_{i1\cdot}^T, \ldots, \mathbf{Z}_{im_i\cdot}^T)^T$ for each i and j, $\mathbf{Z} = (\mathbf{Z}_{1\cdot\cdot}^T, \ldots, \mathbf{Z}_{M\cdot\cdot}^T)^T$, $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_M^T)^T$, and $\boldsymbol{\theta}$ denotes a combination of $\{\boldsymbol{\beta}_k\}$, $\{\sigma_{bk}^2\}$, ρ, and $\{\psi_k(\cdot)\}$ used in Model (1). Then, for a given k, $\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij})$ could be approximated by a local linear expression

$$\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}) + \psi_k'(\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'})\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})$$

when $\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}$ is close to $\boldsymbol{\beta}_k^T\mathbf{X}_{ij}$. If the values of ψ_k and ψ_k' evaluated at $\boldsymbol{\beta}_k^T\mathbf{X}_{ij}$ are denoted as as a_{ijk} and c_{ijk}, respectively, then the conditional log-likelihood $\log f(\mathbf{Z}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\})$ can be approximated by

$$\begin{aligned} & \log f(\mathbf{Z}|\mathbf{X}, \mathbf{b}, \{\boldsymbol{\beta}_k\}, \mathbf{a}, \mathbf{c}) \\ = & \sum_{i,i'=1}^{M}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_{i'}}\left[\sum_{k=1}^{q^*-1} Z_{ijk}\eta_{iji'j'k} - \log\left\{1 + \sum_{k=1}^{q^*-1}\exp(\eta_{iji'j'k})\right\}\right] w_{iji'j'}, \end{aligned}$$

where

$$\begin{aligned} w_{iji'j'} &= \frac{\prod_{k=1}^{q^*-1} K_{h_{0k}}[\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})]}{\sum_{i=1}^{M}\sum_{j=1}^{m_i}[\prod_{k=1}^{q^*-1} K_{h_{0k}}\{\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})\}]}, \\ \eta_{iji'j'k} &= a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + b_{ik}, \end{aligned}$$

$\mathbf{a} = (\mathbf{a}_{11\cdot}^T, \mathbf{a}_{12\cdot}^T, \ldots, \mathbf{a}_{1m_1\cdot}^T, \mathbf{a}_{21\cdot}^T, \ldots, \mathbf{a}_{Mm_M\cdot}^T)^T$, $\mathbf{c} = (\mathbf{c}_{11\cdot}^T, \mathbf{c}_{12\cdot}^T, \ldots, \mathbf{c}_{1m_1\cdot}^T, \mathbf{c}_{21\cdot}^T, \ldots, \mathbf{c}_{Mm_M\cdot}^T)^T$, $\mathbf{a}_{ij\cdot} = (a_{ij1}, \ldots, a_{ij(q^*-1)})^T$, $\mathbf{c}_{ij\cdot} = (c_{ij1}, \ldots, c_{ij(q^*-1)})^T$, for $j = 1, \ldots, m_i$ and $i = 1, \ldots, M$, $K_{h_{0k}}(\cdot) = K(\cdot/h_{0k})/h_{0k}$, for $k = 1, \ldots, (q^*-1)$, $K(\cdot)$ is a density kernel function, and $\{h_{0k}\}$ are q^*-1 bandwidths for estimating the q^*-1 link functions. The kernel function in this paper is chosen to be the Epanechnikov kernel function $K(u) = 0.75(1-u^2)I(|u| \leq 1)$ for its good statistical properties (Epanechnikov, 1969).

To estimate the unknown parameters, we minimize the loss function $l_1(\boldsymbol{\theta})$, defined below as the conditional expectation of the approximate log-likelihood over the posterior distribution of the random-effects coefficients $\mathbf{b}$ with respect to $\{\boldsymbol{\beta}_k\}$, $\mathbf{a}$, $\mathbf{c}$, $\{\sigma_{bk}^2\}$, and ρ_b:

$$l_1(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{b}|\mathbf{Z},\boldsymbol{\Sigma}_b,\{\boldsymbol{\beta}_k\},\mathbf{a},\mathbf{c}} \left\{ \log f(\mathbf{Z}|\mathbf{X},\mathbf{b},\{\boldsymbol{\beta}_k\},\mathbf{a},\mathbf{c}) \log f(\mathbf{b}|\boldsymbol{\Sigma}_b) \,|\mathbf{Z},\boldsymbol{\Sigma}_b,\{\boldsymbol{\beta}_k\},\ \mathbf{a},\mathbf{c}\right\}. \tag{2}$$

Appendix A provides a detailed description of an iterative algorithm for solving this minimization, along with practical guidelines for selecting bandwidths in estimating the link functions.

In practice, a large number of potential risk factors for adverse events are often considered to avoid omitting important ones. However, this approach can inadvertently include factors with minimal impact on the events, increasing the variability of parameter estimates and weakening the model's predictive power. To mitigate this effect, we suggest using the LASSO method (Tibshirani, 1996) to shrink the estimated coefficients of unimportant risk factors toward zero. Specifically, we apply the following adaptive LASSO procedure to estimate the parameters:

$$l_2(\boldsymbol{\theta}) = -\frac{1}{\sum_{i=1}^{M} m_i} l_1(\boldsymbol{\theta}) + \gamma \sum_{k=1}^{q^*-1} \sum_{d=1}^{p} w_{kd}|\beta_{kd}|, \tag{3}$$

where γ is a non-negative regularization parameter, and $\{w_{kd} = 1/|\widetilde{\beta}_{kd}|, k = 1,\ldots,(q^*-1), d = 1,\ldots,p\}$ are the adaptive LASSO weights (Zou 2006). The vectors $\widetilde{\boldsymbol{\beta}}_k = (\widetilde{\beta}_{k1},\ldots,\widetilde{\beta}_{kp})^T$ for $k = 1,\ldots,(q^*-1)$ represent the estimates of $\{\boldsymbol{\beta}_k\}$, obtained by minimizing the objective function in (2). Appendix B provides practical guidelines for minimizing the objective function in (3) and for choosing γ. The final estimates of $\{\psi_k\}$ and $\{\boldsymbol{\beta}_k\}$ obtained from this procedure are denoted as $\{\widehat{\psi}_k\}$ and $\{\widehat{\boldsymbol{\beta}}_k\}$, respectively.

2.2 Conditional risks and their regular longitudinal patterns

After fitting the single-index multinomial logistic regression model (1) based on the training dataset, we can compute different conditional risks. The conditional risks are defined to quantify the likelihood of one or more adverse events given the presence of some others. As a demonstration, consider the case when there are $q = 3$ adverse events. In this setting, there could be $q^* = 2^3 = 8$ possible event statuses, and the single-index multinomial logistic regression model (1) with $q^* - 1 = 7$ non-redundant levels becomes: for $j = 1,\ldots,m_i, i = 1,\ldots,M$, and $k = 1,\ldots,7$,

$$\log\left\{\frac{\mathbb{P}(Y_{ij1} = y_1, Y_{ij2} = y_2, Y_{ij3} = y_3|\mathbf{X}_{ij},\mathbf{b}_i)}{\mathbb{P}(Y_{ij1} = 0, Y_{ij2} = 0, Y_{ij3} = 0|\mathbf{X}_{ij},\mathbf{b}_i)}\right\} = \psi_k\left(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}\right) + b_{ik}, \tag{4}$$

where $y_1, y_2, y_3 = 0$ or 1 but $(y_1, y_2, y_3) \neq (0, 0, 0)$.

Various *conditional risks* can be computed from Model (4), as demonstrated in the following three examples. In Example 1, assume that we would like to detect event 1 given the absence of events 2 and 3. In such a case, the conditional risk can be defined to be: for each i and j,

$$\begin{aligned}
&\text{logit}\{\mathbb{P}(Y_{ij1}=1|Y_{ij2}=0,Y_{ij3}=0,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&=\log\left\{\frac{\mathbb{P}(Y_{ij1}=1|Y_{ij2}=0,Y_{ij3}=0,\mathbf{X}_{ij},\mathbf{b}_i=0)}{1-\mathbb{P}(Y_{ij1}=1|Y_{ij2}=0,Y_{ij3}=0,\mathbf{X}_{ij},\mathbf{b}_i=0)}\right\}\\
&=\log\left\{\frac{\mathbb{P}(Y_{ij1}=1,Y_{ij2}=0,Y_{ij3}=0|\mathbf{X}_{ij},\mathbf{b}_i=0)}{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=0,Y_{ij3}=0|\mathbf{X}_{ij},\mathbf{b}_i=0)}\right\}\\
&=\psi_{(1,0,0)}(\boldsymbol{\beta}_{(1,0,0)}^T\mathbf{X}_{ij}),
\end{aligned}$$

where the logit function is defined to be $\text{logit}(u)=\log[u/(1-u)]$ for $u\in(0,1)$.

In Example 2, the conditional risk for detecting event 2 given the absence of event 1 and presence of event 3 can be defined to be: for each i and j,

$$\begin{aligned}
&\text{logit}\{\mathbb{P}(Y_{ij2}=1|Y_{ij1}=0,Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&=\log\left\{\frac{\mathbb{P}(Y_{ij2}=1|Y_{ij1}=0,Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)}{1-\mathbb{P}(Y_{ij2}=1|Y_{ij1}=0,Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)}\right\}\\
&=\log\left\{\frac{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=1,Y_{ij3}=1|\mathbf{X}_{ij},\mathbf{b}_i=0)}{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=0,Y_{ij3}=0|\mathbf{X}_{ij},\mathbf{b}_i=0)}\right\}\\
&-\log\left\{\frac{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=0,Y_{ij3}=1|\mathbf{X}_{ij},\mathbf{b}_i=0)}{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=0,Y_{ij3}=0|\mathbf{X}_{ij},\mathbf{b}_i=0)}\right\}\\
&=\psi_{(0,1,1)}(\boldsymbol{\beta}_{(0,1,1)}^T\mathbf{X}_{ij})-\psi_{(0,0,1)}(\boldsymbol{\beta}_{(0,0,1)}^T\mathbf{X}_{ij}).
\end{aligned}$$

In Examples 1 and 2 above, the conditional risks are for detecting a single event given pre-existing conditions. In Example 3, we consider detecting multiple events. Specifically, the conditional risk for detecting events 1 and 2 given the presence of event 3 can be defined as follows: for each i and j,

$$\begin{aligned}
&\text{logit}\{\mathbb{P}(Y_{ij1}=1,Y_{ij2}=1|Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&=\log\{\mathbb{P}(Y_{ij1}=1,Y_{ij2}=1|Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&-\log\{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=1|Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&-\log\{\mathbb{P}(Y_{ij1}=1,Y_{ij2}=0|Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&-\log\{\mathbb{P}(Y_{ij1}=0,Y_{ij2}=0|Y_{ij3}=1,\mathbf{X}_{ij},\mathbf{b}_i=0)\}\\
&=\log\left\{\frac{\exp(\psi_{(1,1,1)}(\boldsymbol{\beta}_{(1,1,1)}^T\mathbf{X}_{ij}))}{\exp(\psi_{(0,1,1)}(\boldsymbol{\beta}_{(0,1,1)}^T\mathbf{X}_{ij}))+\exp(\psi_{(1,0,1)}(\boldsymbol{\beta}_{(1,0,1)}^T\mathbf{X}_{ij}))+\exp(\psi_{(0,0,1)}(\boldsymbol{\beta}_{(0,0,1)}^T\mathbf{X}_{ij}))}\right\}.
\end{aligned}$$

In the above three examples, a single conditional risk is considered. However, in some scenarios, it becomes necessary to evaluate multiple conditional risks simultaneously, as illustrated in Example 4 below. Suppose we want to detect the occurrence of exactly one event given the absence of other events. In such cases, the following three conditional risks can be jointly monitored: for each i and j,

$$\begin{bmatrix} \text{logit}\{\mathbb{P}(Y_{ij1}=1|Y_{ij2}=0,Y_{ij3}=0,\mathbf{X}_{ij},\mathbf{b}_i=0)\} \\ \text{logit}\{\mathbb{P}(Y_{ij2}=1|Y_{ij1}=0,Y_{ij3}=0,\mathbf{X}_{ij},\mathbf{b}_i=0)\} \\ \text{logit}\{\mathbb{P}(Y_{ij3}=1|Y_{ij1}=0,Y_{ij2}=0,\mathbf{X}_{ij},\mathbf{b}_i=0)\} \end{bmatrix} = \begin{bmatrix} \psi_{(1,0,0)}(\boldsymbol{\beta}_{(1,0,0)}^T\mathbf{X}_{ij}) \\ \psi_{(0,1,0)}(\boldsymbol{\beta}_{(0,1,0)}^T\mathbf{X}_{ij}) \\ \psi_{(0,0,1)}(\boldsymbol{\beta}_{(0,0,1)}^T\mathbf{X}_{ij}) \end{bmatrix}.$$

From the above examples, it is evident that various conditional risks can be defined using Model (4) for detecting one or more events given the pre-existing status. For the ith subject at the jth observation time t_{ij}, let $\mathbf{r}_i(t_{ij}) = (r_{i1}(t_{ij}),\ldots,r_{i\ell}(t_{ij}))^T$ represent the ℓ-dimensional vector of conditional risks of interest, where $\ell \geq 1$. Once Model (4) is estimated, the estimate of $\mathbf{r}_i(t_{ij})$ can be computed, denoted as $\widehat{\mathbf{r}}_i(t_{ij}) = (\widehat{r}_{i1}(t_{ij}),\ldots,\widehat{r}_{i\ell}(t_{ij}))^T$. Next, we aim to estimate the *regular longitudinal pattern* of $\widehat{\mathbf{r}}_i(t_{ij})$ to facilitate early identification of the event status of interest for new subjects. Without loss of generality, let us assume that the first M_1 subjects in the training data are well-functioning (or IC) during the whole time interval $[T_1,T_2]$. The remaining $M-M_1$ subjects are thus considered OC during $[T_1,T_2]$. Here, "well-functioning" or "IC" subjects corresponding to a conditional risk of interest are those who meet the condition of the conditional risk and do not experience the events to detect in the study period $[T_1,T_2]$. For example, if the conditional risk of event 3 given the presence of events 1 and 2 is our concern, then "well-functioning" subjects related to this conditional risk are those who experience events 1 and 2 but do not experience event 3 in $[T_1,T_2]$.

For the M_1 IC subjects, the quantities $\widehat{\mathbf{r}}_i(t_{ij})$ are assumed to follow the multivariate nonparametric longitudinal model:

$$\widehat{\mathbf{r}}_i(t_{ij}) = \boldsymbol{\mu}_r(t_{ij}) + \boldsymbol{\varepsilon}_{r,i}(t_{ij}),\ \ j=1,\ldots,m_i,\ \ i=1,\ldots,M_1, \tag{5}$$

where $\boldsymbol{\mu}_r(t_{ij}) = (\mu_{r,1}(t_{ij}),\ldots,\mu_{r,\ell}(t_{ij}))^T$ is the mean vector, and $\boldsymbol{\varepsilon}_{r,i}(t_{ij}) = (\varepsilon_{r,i1}(t_{ij}),\ldots,\varepsilon_{r,i\ell}(t_{ij}))^T$ represents the zero-mean error term. For any $s,t \in [0,T]$, the covariance structure of the error term is described by the covariance matrix function $\mathbf{V}_r(s,t) = \text{cov}(\boldsymbol{\varepsilon}_{r,i}(s),\boldsymbol{\varepsilon}_{r,i}(t))$. To estimate Model (5), we employ a modified version of the estimation procedures described in Li (2011) and Xiang et al. (2013). Detailed information on the estimation procedure for Model (5) is provided in Appendix C.

2.3 Online monitoring of conditional risks

To detect target events early for a new subject, given the subject's current event status, the estimated conditional risks, denoted as $\widehat{\mathbf{r}}(t_j^*)$, can be computed using the observed risk factors and the estimated model (4). Here, $\{t_j^*, j \geq 1\}$ represent the observation times of the new subject. When calculating $\widehat{\mathbf{r}}(t_j^*)$, a boundary issue may arise. From the definition of conditional risks in Section 2.2, $\widehat{\mathbf{r}}(t_j^*)$ is derived using the estimated link functions $\{\widehat{\psi}_k\}$ and single-index coefficients $\{\widehat{\boldsymbol{\beta}}_k\}$ through the expression $\widehat{\psi}_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}(t_j^*))$, both of which are obtained from the training dataset. However, $\{\widehat{\psi}_k\}$ are defined only within the range of $\{\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij}, i = 1, \ldots, M, j = 1, \ldots, m_i, k = 1, \ldots, q^* - 1\}$, denoted as $[u_{min}, u_{max}]$, which is determined by the training data. For a given $k = 1, \ldots, q^* - 1$, the values of $\{\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}(t_j^*), j \geq 1\}$ for the new subject could fall outside the range $[u_{min}, u_{max}]$. This would render some values of $\widehat{\mathbf{r}}(t_j^*)$ undefined.

To address this boundary issue, we propose fitting a linear model for the following data points respectively: $\{(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij}, \widehat{\psi}_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})), \widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij} \in [u_{min}, u_{min} + h_{0k}]\}$ and $\{(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij}, \widehat{\psi}_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij})), \widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_{ij} \in [u_{max} - h_{0k}, u_{max}]\}$. Using these linear models, we can extrapolate to estimate the values of $\widehat{\psi}_k(\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}(t_j^*))$ for cases where $\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}(t_j^*)$ lies outside the range $[u_{min}, u_{max}]$.

Next, we employ an SPC chart to sequentially monitor the new subject's estimated conditional risks, $\{\widehat{\mathbf{r}}(t_j^*), j \geq 1\}$. Conventional control charts, typically designed for scenarios where IC process observations are independent and identically distributed (i.i.d.) (Qiu, 2014), are not suitable for the current problem as the i.i.d. assumptions may be violated. To address this limitation, Li and Qiu (2016, 2017) introduced sequential data decorrelation and standardization procedures for both univariate and multivariate process monitoring. Subsequently, You and Qiu (2019) and Tian and Qiu (2024b) proposed computationally more efficient versions of these procedures. The multivariate approach described in Tian and Qiu (2024b), which focuses on data decorrelation and standardization for multivariate processes, is briefly outlined below. The univariate procedure can be implemented in a similar manner.

- At the first time point t_1^*, the standardized value of $\widehat{\mathbf{r}}(t_1^*)$ is defined to be $\widehat{\mathbf{e}}(t_1^*) = \{\widehat{\mathbf{V}}_r(t_1^*, t_1^*)\}^{-1/2}\{\widehat{\mathbf{r}}(t_1^*) - \widehat{\boldsymbol{\mu}}_r(t_1^*)\}$. Define $\widehat{\mathbf{U}}_1 = \{\widehat{\mathbf{V}}_r(t_1^*, t_1^*)\}^{-1/2}$ for subsequent use.

- At t_j^* when $j \geq 2$, $\widehat{\mathbf{r}}(t_j^*)$ needs to be standardized and decorrelated with conditional risks at all previous time points. Let $\widehat{\mathbf{R}}_j = (\widehat{\mathbf{r}}(t_1^*)^T, \ldots, \widehat{\mathbf{r}}(t_j^*)^T)^T$ and denote the estimate of $\text{cov}(\widehat{\mathbf{R}}_j, \widehat{\mathbf{R}}_j)$ as

$$\widehat{\boldsymbol{\Sigma}}_{j,j} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_{j-1,j-1} & \widehat{\mathbf{A}}_{j-1,j} \\ \widehat{\mathbf{A}}_{j-1,j}^T & \widehat{\mathbf{V}}_r(t_j^*, t_j^*) \end{pmatrix},$$

where $\widehat{\mathbf{A}}_{j-1,j} = (\widehat{\mathbf{V}}_r(t_1^*, t_j^*)^T, \ldots, \widehat{\mathbf{V}}_r(t_{j-1}^*, t_j^*)^T)^T$. Then, the standardized and decorrelated version of $\widehat{\mathbf{r}}(t_j^*)$ is defined to be

$$\widehat{\mathbf{e}}(t_j^*) = \widehat{\mathbf{D}}_j^{-1/2} \left\{ \widehat{\mathbf{r}}(t_j^*) - \widehat{\boldsymbol{\mu}}_r(t_j^*) - \widehat{\mathbf{L}}_j^T \widehat{\mathbf{Z}}_{j-1}^* \right\},$$

where $\widehat{\mathbf{L}}_j = \widehat{\mathbf{U}}_{j-1}\widehat{\mathbf{A}}_{j-1,j}$, $\widehat{\mathbf{Z}}_{j-1}^* = (\widehat{\mathbf{e}}(t_1^*), \ldots, \widehat{\mathbf{e}}(t_{j-1}^*))^T$, $\widehat{\mathbf{D}}_j = \widehat{\mathbf{V}}_r(t_j^*, t_j^*) - \widehat{\mathbf{L}}_j^T \widehat{\mathbf{L}}_j$, and $\widehat{\mathbf{U}}_j$ can be recursively updated by

$$\widehat{\mathbf{U}}_j = \begin{pmatrix} \widehat{\mathbf{U}}_{j-1} & \mathbf{0} \\ -\widehat{\mathbf{D}}_j^{-1/2} \widehat{\mathbf{L}}_j^T \widehat{\mathbf{U}}_{j-1} & \widehat{\mathbf{D}}_j^{-1/2} \end{pmatrix}.$$

As a side note, $\widehat{\mathbf{D}}_j^{-1/2}$ may not always exist, as $\widehat{\mathbf{D}}_j$ might not be a positive semi-definite matrix. If $\widehat{\mathbf{D}}_j$ is not positive semi-definite, we recommend modifying it to ensure positive semi-definiteness using the algorithm proposed by Higham (1988).

After the above data standardization and decorrelation, the transformed conditional risks become $\{\widehat{\mathbf{e}}(t_j^*), j \geq 1\}$, which are asymptotically uncorrelated, with each having an asymptotic mean of $\mathbf{0}$ and an asymptotic identity covariance matrix. To sequentially monitor these transformed conditional risks, we propose using the following multivariate control chart. This chart accommodates unequally spaced observation times and is designed to effectively detect upward mean shifts in the conditional risks. Let $\omega > 0$ represent a *basic time unit* such that all observation times are integer multiples of ω. Thus, we have $\{t_j^* = n_j^* \omega, j \geq 1\}$, where n_j^* is a positive integer denoting the jth observation time in units of ω. At time t_j^*, the charting statistic is defined to be $\mathbf{E}_j^{+T}\mathbf{E}_j^+$, where $\mathbf{E}_j^+ = (E_{j1}^+, \ldots, E_{j\ell}^+)^T$,

$$E_{j\ell^*}^+ = \max\left[0, \{1 - \Lambda(t_j^*)\}E_{j-1,\ell^*}^+ + \Lambda(t_j^*)\widehat{e}_{\ell^*}(t_j^*)\right], \text{ for } j \geq 1, \ \ \ell^* = 1, \ldots, \ell, \tag{6}$$

$\mathbf{E}_0^+ = 0$, $\widehat{e}_{\ell^*}(t_j^*)$ is the ℓ^*th element of $\widehat{\mathbf{e}}(t_j^*)$, $\Lambda(t_1^*) = 1 - (1-\lambda)^{\overline{\Delta}}$, $\overline{\Delta} = \mathbb{E}(\Delta_j)$, $\Delta_j = n_j^* - n_{j-1}^*$, λ is a smoothing parameter, and

$$\Lambda(t_j^*) = \frac{\Lambda(t_{j-1}^*)}{(1-\lambda)^{\Delta_j} + \Lambda(t_{j-1}^*)}, \text{ for } j \geq 2.$$

The chart signals an upward mean shift in the multivariate conditional risks if

$$\mathbf{E}_j^{+T}\mathbf{E}_j^+ > \rho, \tag{7}$$

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

From (6), it is evident that λ and $\{\Delta_j, j \geq 1\}$ determine the weights assigned to historical data when calculating the charting statistic values. These weights decay exponentially for older data, while the un-

equally spaced observation times are accounted for through $\{\Delta_j, j \geq 1\}$. Since the adverse events of interest would increase the associated conditional risks, a one-sided structure is used in (6) to prevent signals from being triggered by downward shifts in any dimensions of the conditional risks.

In the SPC literature, the performance of a control chart is typically measured by the average run length (ARL) (Qiu, 2014). However, when observation times are unequally spaced, the ARL metric is not a meaningful performance measure. To address this limitation, Qiu and Xiang (2014) proposed using the average time to signal (ATS) in the basic time unit ω for such scenarios. With this performance metric, the in-control ATS value (ATS_0) is typically pre-specified at a desired level, and a chart is considered more effective in detecting a given shift if its OC ATS value (ATS_1) is smaller.

In control chart (6)-(7), the smoothing parameter λ is usually pre-specified, while the control limit ρ must be chosen to achieve the desired ATS_0 value. Since the transformed conditional risks $\{\widehat{\mathbf{e}}(t_1^*), \ldots, \widehat{\mathbf{e}}(t_j^*)\}$ may not follow a normal distribution, the bisection search procedure that relays on Monte Carlo simulations with standard normal data, as discussed in Qiu and Xiang (2014) and Li and Qiu (2017), is not applicable for selecting ρ. To address this challenge, we recommend using the block bootstrap procedure described in Qiu and Xiang (2014). In this approach, the training data is divided into two halves. The first half is used to estimate Model (1) and the regular longitudinal pattern of the event risks of interest, and the second half is utilized to determine the value of ρ.

To conclude the section, Figure 1 summarizes key components of the proposed DySS method and illustrates the sequential decision-making procedure.

3 Simulation Studies

In this section, we evaluate the numerical performance of the proposed method through simulation results. The simulation settings are described as follows. First, observation times for all subjects are confined to the study period $[0, 1]$, with a basic time unit of $\omega = 0.01$. For simplicity, the number of observation times for each subject is assumed to be equal and denoted by m. The jth observation time for the ith subject, t_{ij}, is defined as $n_{ij}\omega$, where n_{ij} is the ceiling of a random value generated from the uniform distribution Unif$(100(j-1)/m, 100j/m)$, for each i and j.

Second, we assume there are p risk factors associated with $q = 3$ adverse events of interest. We begin by focusing on the case where $p = 10$ in Sections 3.1 and 3.2, and then examine the impact of

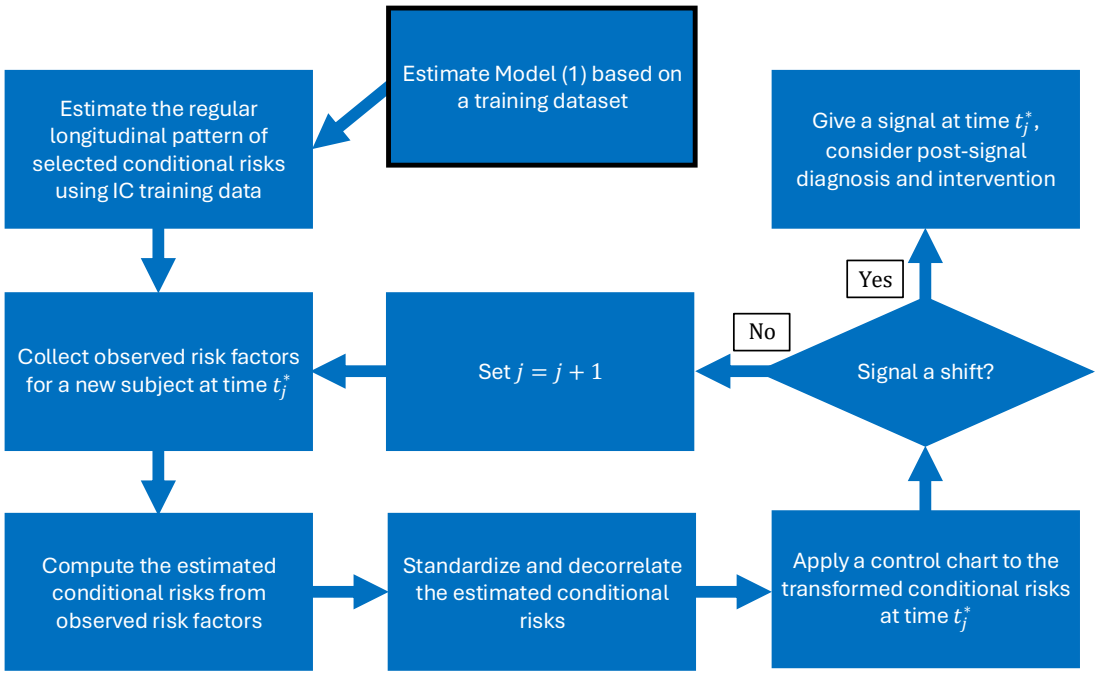


Figure 1: Diagram of the proposed DySS method for online monitoring of selected conditional risks.

increasing p on the performance of the proposed methods in Section 3.3. When $p = 10$, let $\mathbf{X}_i(t) = (X_{i1}(t), X_{i2}(t), \ldots, X_{i10}(t))^T$ denote the vector of these risk factors for the ith subject at time t. Their IC observations are generated using the following mixed-effects models:

$$X_{id}(t) = \mu_{x,d}(t) + \zeta_{id} + \tau_2 \sigma_{x,d}(t)\epsilon_{id}(t), \text{ for } d = 1, \ldots, 10, \tag{8}$$

where the mean functions $\{\mu_{x,d}(t), d = 1, \ldots, 10\}$ are defined as follows: $\mu_{x,1}(t) = 0$, $\mu_{x,d}(t) = t/2$ for $d = 2, 3, 4$, $\mu_{x,d}(t) = 0.5t^2 + 0.5t$ for $d = 5, 6, 7$, and $\mu_{x,d}(t) = \cos(2\pi t)/2$ for $d = 8, 9, 10$. The random-effects term $(\zeta_{i1}, \ldots, \zeta_{i10})^T$ is generated from $\mathrm{N}_{10}(\mathbf{0}, \mathbf{\Sigma}_x)$, where $\mathbf{\Sigma}_x$ is an exchangeable covariance matrix with diagonal elements 1 and off-diagonal elements τ_1. Thus, τ_1 and τ_2 are two parameters controlling the magnitude of between-risk-factor correlation, while τ_2 also controls the magnitude of the within-subject correlation. For the simulation studies, we set $\tau_1 = 0.2$ or 0.5 and $\tau_2 = 1$ or 1.5, allowing for scenarios with small or large within-subject and between-risk-factor variability. The variance functions $\sigma_{x,d}(t)$ are defined as follows: $\sigma_{x,1}(t) = 1$, $\sigma_{x,d}(t) = \log(1 + 5t)/2$ for $d = 2, 3, 4$, $\sigma_{x,d}(t) = 1/(1 + t)$ for $d = 5, 6, 7$, and $\sigma_{x,d}(t) = \sin(3\pi t)/3$ for $d = 8, 9, 10$. The pure measurement error $\epsilon_{il}(t)$ for $d = 1, \ldots, 10$ are i.i.d. and generated from either a standard normal distribution or a standardized χ^2_5 distribution with zero mean and unit variance.

Third, for the $q = 3$ adverse events, there are a total of $q^* = 8$ possible event statuses at any given time. The IC probability distribution of these statuses is assumed to follow a single-index multinomial logistic regression model: for $j = 1, \ldots, m$ and $i = 1, \ldots, M$,

$$\log\left\{\frac{\mathbb{P}(Y_{ij1} = y_1, Y_{ij2} = y_2, Y_{ij3} = y_3|\mathbf{X}_i(t_{ij}), \mathbf{b}_i)}{\mathbb{P}(Y_{ij1} = 0, Y_{ij2} = 0, Y_{ij3} = 0|\mathbf{X}_i(t_{ij}), \mathbf{b}_i)}\right\} = \psi_\kappa\left(\boldsymbol{\beta}_\kappa^T \mathbf{X}_i(t_{ij})\right) + b_{i\kappa}, \tag{9}$$

where $\kappa \in \mathbf{K} = \{(1,1,1), (0,1,1), (1,0,1), (1,1,0), (1,0,0), (0,1,0), (0,0,1)\}$. The random-effects term $\mathbf{b}_i = (b_{i(1,1,1))}, \ldots, b_{i(0,0,1)})^T$ is generated from a multivariate normal distribution $\mathrm{N}_7(\mathbf{0}, \boldsymbol{\Sigma}_b)$, where $\boldsymbol{\Sigma}_b$ is an exchangeable covariance matrix with diagonal elements 1 and off-diagonal elements 0.1. The true values of the single-index coefficients are set as follows:

$$\begin{aligned}
\boldsymbol{\beta}_{(1,1,1)} &= (3,2,1,1,2,1,1,2,1,1)/\sqrt{15},\\
\boldsymbol{\beta}_{(0,1,1)} &= (2,0,0,0,2,1,1,2,1,1)/\sqrt{10},\\
\boldsymbol{\beta}_{(1,0,1)} &= (2,2,1,1,0,0,0,2,1,1)/\sqrt{10},\\
\boldsymbol{\beta}_{(1,1,0)} &= (2,2,1,1,2,1,1,0,0,0)/\sqrt{10},\\
\boldsymbol{\beta}_{(1,0,0)} &= (1,2,1,1,0,0,0,0,0,0)/\sqrt{5},\\
\boldsymbol{\beta}_{(0,1,0)} &= (1,0,0,0,2,1,1,0,0,0)/\sqrt{5},\\
\boldsymbol{\beta}_{(0,0,1)} &= (1,0,0,0,0,0,0,2,1,1)/\sqrt{5}.
\end{aligned}$$

By using these values, it is assumed that the first risk factor X_{ij1} contributes to all three adverse events, the risk factors X_{ij2} to X_{ij4} contribute only to the first event, the risk factors X_{ij5} to X_{ij7} contribute only to the second event, and the risk factors X_{ij8} to X_{ij10} contribute only to the third event. The link functions are assumed to be

$$\begin{aligned}
\psi_1(u) = u/2,\ \psi_2(u) = \psi_3(u) = \psi_4(u) &= \{\sin(\pi u/8)/2 + 1\}u/2 - 1/2,\\
\psi_5(u) = \psi_6(u) = \psi_7(u) &= \{\exp(u/5)/5 + 1\}u/2 - 1/2.
\end{aligned}$$

In each simulation study, a training dataset with M subjects is first generated. Three-fourths of these subjects are generated from the IC model (9), while the remaining one-fourth are generated from an OC model. The OC model is identical to the IC model except that the risk factors have an OC mean given by $\boldsymbol{\mu}_x(t) + \mathbf{1}_{10}^T \delta^*/\sqrt{10}$, where $\boldsymbol{\mu}_x(t) = (\mu_{x,1}(t), \ldots, \mu_{x,10}(t))^T$ is the IC mean, $\mathbf{1}_{10}$ is a 10-dimensional vector of ones, and δ^* is a random number generated from Unif$[0, 3]$. Thus, $\delta^*/\sqrt{10}$ represents the shift size in each risk factor. From the discussions about conditional risks in Section 2.2, it can be easily checked that

such shifts in risk factors would lead to shifts in the corresponding conditional risks to be monitored. Since conditional risks are not directly observable, it is more practical to work with risk factors when generating observed data for a given number of IC or OC subjects in the training dataset.

Next, half of the training data, consisting of $M_0 = M/2$ subjects, is used to estimate Model (1), and the regular longitudinal pattern of the estimated conditional risks computed based on the estimated model (1) is determined using the observed data from the well-functioning subjects in that part of the training data, as discussed in Sections 2.1 and 2.2. The observed data from the well-functioning subjects in the second half of the training dataset are used to determine the control limit of the chart (6)-(7), as discussed in Section 2.3. When splitting the training data, we maintain a $3:1$ ratio of IC to OC subjects within each subset.

Once the training data are generated, the conditional ATS_0 (or ATS_1) value, given the training data, is computed based on 1,000 simulations of online process monitoring. This entire process, from generating the training data to calculating the ATS_0 (or ATS_1) value, is repeated 100 times. The average of these 100 conditional ATS_0 (or ATS_1) values serves as an estimate of the actual ATS_0 (or ATS_1) value, with its standard error also computed. In all simulation studies, the nominal ATS_0 value is fixed at 25.

We begin by presenting numerical results for estimating Model (9). Table 1 shows the sum of squares of estimation error (SSE) for the single-index coefficient vectors $\boldsymbol{\beta}_\kappa : \kappa \in \mathbf{K}$ under scenarios with and without the adaptive LASSO penalty in model estimation. These results are based on cases where the pure measurement error follows a normal distribution. The SSE is defined as

$$SSE = \sum_{\kappa \in \mathbf{K}} \sum_{d=1}^{p} \left(\widehat{\beta}_{\kappa d} - \beta_{\kappa d} \right)^2 .$$

The corresponding standard errors are shown in parentheses in the table. From the results in Table 1, we observe that the SSE and its standard error decrease as the number of subjects (M_0) and/or the number of observation times (m) increase, indicating improved estimation accuracy with larger sample sizes. Furthermore, for the same sample size, the use of the adaptive Lasso penalty effectively reduces both the SSE and its standard error, demonstrating its utility in enhancing estimation performance. Results for cases where the pure measurement error follows a chi-square distribution are presented in Table A.1 in Appendix D. These results lead to similar conclusions.

Next, we examine the numerical performance of the control chart (6)-(7) across four scenarios corresponding to Examples 1–4 discussed in Section 2.2. In Scenario 1, the focus is on monitoring the conditional risk of Event 1, given the absence of Events 2 and 3. For a new subject with observation times $\{t_j^*, j^* \geq 1\}$,

Table 1: Sum of squares of estimation error (SSE) of the estimated single-index coefficients in Model (9) across various cases when the pure measurement error in Model (8) has a normal distribution. Standard errors are provided in parentheses.

		$m = 5$		$m = 10$	
(τ_1, τ_2)	M_0	No penalty	LASSO penalty	No penalty	LASSO penalty
(0.2,1)	200	2.906 (0.170)	2.435 (0.104)	1.541 (0.052)	1.419 (0.045)
	300	1.032 (0.068)	0.838 (0.068)	0.489 (0.032)	0.458 (0.030)
	500	0.390 (0.024)	0.262 (0.020)	0.290 (0.012)	0.218 (0.011)
(0.2,1.5)	200	3.179 (0.226)	2.626 (0.136)	1.541 (0.049)	1.422 (0.045)
	300	1.818 (0.156)	1.457 (0.092)	0.676 (0.034)	0.617 (0.031)
	500	1.338 (0.114)	0.895 (0.056)	0.580 (0.017)	0.437 (0.012)
(0.5,1)	200	1.564 (0.076)	1.289 (0.068)	0.531 (0.014)	0.481 (0.011)
	300	0.540 (0.036)	0.444 (0.031)	0.274 (0.020)	0.239 (0.018)
	500	0.314 (0.021)	0.204 (0.014)	0.119 (0.002)	0.090 (0.002)
(0.5,1.5)	200	2.565 (0.175)	2.076 (0.105)	1.826 (0.053)	1.611 (0.048)
	300	1.140 (0.057)	0.902 (0.053)	0.603 (0.018)	0.516 (0.014)
	500	0.403 (0.042)	0.269 (0.020)	0.201 (0.008)	0.149 (0.007)

the goal is to sequentially monitor the conditional risk:

$$\widehat{r}(t_j^*) = \widehat{\psi}_{(1,0,0)}(\widehat{\boldsymbol{\beta}}_{(1,0,0)}^T \mathbf{X}(t_j^*)).$$

In Scenario 2, the focus is on monitoring the conditional risk of Event 2, given the absence of Event 1 and the presence of Event 3. In this case, the conditional risk to be monitored is:

$$\widehat{r}(t_j^*) = \widehat{\psi}_{(0,1,1)}(\widehat{\boldsymbol{\beta}}_{(0,1,1)}^T \mathbf{X}(t_j^*)) - \widehat{\psi}_{(0,0,1)}(\widehat{\boldsymbol{\beta}}_{(0,0,1)}^T \mathbf{X}(t_j^*)).$$

In Scenario 3, the focus is on monitoring the conditional risk of Events 1 and 2, given the presence of Event 3. The conditional risk to be monitored has the expression:

$$\widehat{r}(t_j^*) = \log \left\{ \frac{\exp(\psi_{(1,1,1)}(\widehat{\boldsymbol{\beta}}_{(1,1,1)}^T \mathbf{X}(t_j^*)))}{\exp(\psi_{(0,1,1)}(\widehat{\boldsymbol{\beta}}_{(0,1,1)}^T \mathbf{X}(t_j^*))) + \exp(\psi_{(1,0,1)}(\widehat{\boldsymbol{\beta}}_{(1,0,1)}^T \mathbf{X}(t_j^*))) + \exp(\psi_{(0,0,1)}(\widehat{\boldsymbol{\beta}}_{(0,0,1)}^T \mathbf{X}(t_j^*)))} \right\}.$$

In Scenario 4, the goal is to detect each of the three events, given the absence of the others. In this case, we

need to sequentially monitor the following multivariate conditional risk:

$$\mathbf{r}(t_j^*) = \left[\widehat{\psi}_{(1,0,0)}(\widehat{\boldsymbol{\beta}}_{(1,0,0)}^T \mathbf{X}(t_j^*)), \widehat{\psi}_{(0,1,0)}(\widehat{\boldsymbol{\beta}}_{(0,1,0)}^T \mathbf{X}(t_j^*)), \widehat{\psi}_{(0,0,1)}(\widehat{\boldsymbol{\beta}}_{(0,0,1)}^T \mathbf{X}(t_j^*))\right]^T.$$

The proposed method described in Section 2.3, referred to as CRM (Conditional Risk Monitoring), is compared with four competing approaches. First, a simplified version of the proposed method that excludes the adaptive LASSO penalty in estimating Model (1) is denoted as CRM\L. Second, the multivariate DySS (MDySS) method introduced by Qiu and Xiang (2015), which monitors the observed event risk factors directly after standardization by the estimated regular longitudinal pattern, is denoted as MDySS-std. Third, the multivariate DySS method proposed by Li and Qiu (2017), which monitors standardized and decorrelated event risk factors, is referred to as MDySS-dec. Fourth, we include the RF-SLAM method (Wongvibulsin et al., 2020). In RF-SLAM, the longitudinal follow-up data of each subject are reformulated into a series of counting process information units (CPIUs), defined over a sequence of pre-specified time intervals. Each CPIU includes the subject identifier, interval index, baseline and time-varying risk factors, and event status indicators. For each simulation run, an ensemble of 1,000 trees is constructed using training data from M_0 subjects. For a new subject, the conditional risk at each observation time is defined as the probability of experiencing the target event, given the statuses of other events at that time. This risk is estimated based on the subject's previous risk factor values and event status indicators. To ensure comparability with the other methods, CPIUs are defined over consecutive intervals of equal length, with the length being the basic time unit $\omega = 0.01$. A signal is generated for a new subject if the estimated conditional risk exceeds a pre-specified threshold.

For the first four methods that use control charts, control limits are determined using the block bootstrap procedure described at the end of Section 2. For RF-SLAM, the threshold for triggering a signal is set as a specified percentile of the conditional risk distribution, computed from the observed data of IC subjects in the second half of the training dataset, as discussed about the block bootstrap procedure at the end of Section 2.

3.1 IC performance of the proposed method

For the case where $\tau_1 = 0.2$, $\tau_2 = 1$, the pure measurement error follows a standard normal distribution, and the conditional risk of interest is the one in Scenario 1, the actual ATS_0 values for the different methods are presented in Table 2 across various settings. From the table, the actual ATS_0 values for the first four

methods, CRM, CRM\L, MDySS-std, and MDySS-dec, fall within 5% of the nominal ATS_0 value of 25 in all cases. Additionally, as the number of subjects (M_0) increases, the actual ATS_0 values for all of them converge closer to the nominal ATS_0 value. These findings confirm the effectiveness of the block bootstrap procedure discussed in Section 2 for determining the control limits of the control charts. For RF-SLAM, the threshold is set at the 35th percentile of the estimated conditional risk of interest, calculated using the second half of the training data. While this threshold yields an actual ATS_0 value close to the nominal level of 25 when $M_0 = 200$ and $m = 5$, the actual ATS_0 values under other sample size settings deviate substantially from the nominal value. Between the two methods, CRM and CRM\L, it is observed that CRM consistently exhibits a smaller standard error for the computed actual ATS_0 values in all cases. This is likely due to the reduced variability in the estimated conditional risk achieved by CRM, which shrinks the estimated single-index coefficients for unrelated risk factors to zero. Additional results, including the actual ATS_0 values for Scenarios 2–4 and scenarios with chi-square distributed pure measurement error, are presented in Tables A.2–A.8 in Appendix D. Similar conclusions can be drawn from those results.

3.2 OC performance of the proposed method

Next, we evaluate the OC performance of various control charts under consideration when $M_0 = 500$ and $m = 10$. In Scenario 1, the observed risk factors are generated based on the following OC model:

$$\mathbf{X}^*(t) = \mathbf{X}(t) + (1, 2, 1, 1, 0, 0, 0, 0, 0, 0)^T \delta / \sqrt{10},$$

where $\mathbf{X}(t)$ represents the vector of IC risk factors discussed earlier, and δ controls the magnitude of the mean shift, varying from 0 to 3. Since Event 1 is assumed to be associated exclusively with the first four risk factors, it is reasonable to introduce shifts in these factors when detecting the first event in the absence of the other two events in Scenario 1. Similarly, in Scenario 2, the observed risk factors are generated using the following OC model:

$$\mathbf{X}^*(t) = \mathbf{X}(t) + (1, 0, 0, 0, 2, 1, 1, 0, 0, 0)^T \delta / \sqrt{10}.$$

In Scenario 3, the observed risk factors are generated using the OC model:

$$\mathbf{X}^*(t) = \mathbf{X}(t) + (2, 2, 1, 1, 2, 1, 1, 0, 0, 0)^T \delta / \sqrt{10}.$$

In Scenario 4, the observed risk factors are generated using the OC model:

$$\mathbf{X}^*(t) = \mathbf{X}(t) + (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)^T \delta / \sqrt{10}.$$

Table 2: Computed ATS_0 values (standard errors) of the charts CRM, CRM\L, MDySS-std, MDySS-dec and RF-SLAM in Scenario 1 when $\tau_1 = 0.2$, $\tau_2 = 1$, the pure measurement error in (8) is normally distributed, and the nominal ATS_0 value is fixed at 25 for all methods.

			Computed ATS_0 (SE)				
λ	m	M_0	CRM	CRM\L	MDySS-std	MDySS-dec	RF-SLAM
0.05	5	200	25.661 (0.243)	25.699 (0.246)	25.723 (0.066)	25.668 (0.082)	25.076 (0.103)
		300	25.252 (0.212)	25.188 (0.237)	25.173 (0.060)	25.450 (0.073)	23.960 (0.098)
		500	25.223 (0.203)	25.112 (0.222)	25.192 (0.060)	25.291 (0.070)	20.527 (0.081)
	10	200	25.250 (0.215)	25.216 (0.242)	25.299 (0.062)	25.362 (0.076)	23.036 (0.085)
		300	24.687 (0.208)	25.001 (0.209)	24.955 (0.059)	25.043 (0.070)	19.012 (0.082)
		500	24.848 (0.192)	24.814 (0.210)	24.856 (0.055)	24.896 (0.067)	16.776 (0.060)
0.1	5	200	25.524 (0.233)	25.454 (0.252)	25.678 (0.075)	25.673 (0.097)	25.076 (0.103)
		300	25.123 (0.215)	24.943 (0.237)	25.224 (0.064)	25.329 (0.083)	23.960 (0.098)
		500	25.108 (0.205)	25.262 (0.225)	25.216 (0.062)	25.349 (0.083)	20.527 (0.081)
	10	200	25.178 (0.218)	25.052 (0.239)	25.322 (0.071)	25.317 (0.084)	23.036 (0.085)
		300	24.793 (0.194)	24.713 (0.213)	24.812 (0.063)	25.057 (0.076)	19.012 (0.082)
		500	24.774 (0.191)	24.840 (0.207)	24.850 (0.059)	24.865 (0.076)	16.776 (0.060)
0.2	5	200	25.570 (0.235)	25.534 (0.241)	25.568 (0.079)	25.770 (0.107)	25.076 (0.103)
		300	25.139 (0.201)	25.166 (0.223)	25.155 (0.081)	25.398 (0.099)	23.960 (0.098)
		500	25.253 (0.199)	25.064 (0.219)	25.254 (0.073)	25.115 (0.097)	20.527 (0.081)
	10	200	25.251 (0.220)	25.388 (0.241)	25.226 (0.075)	25.316 (0.104)	23.036 (0.085)
		300	24.661 (0.197)	24.976 (0.217)	24.910 (0.073)	24.918 (0.098)	19.012 (0.082)
		500	24.804 (0.185)	24.791 (0.203)	24.818 (0.067)	24.839 (0.089)	16.776 (0.060)
0.5	5	200	25.789 (0.241)	25.601 (0.272)	25.625 (0.128)	25.571 (0.152)	25.076 (0.103)
		300	25.517 (0.226)	25.542 (0.242)	24.935 (0.113)	24.949 (0.127)	23.960 (0.098)
		500	25.208 (0.213)	25.293 (0.226)	25.218 (0.111)	25.233 (0.130)	20.527 (0.081)
	10	200	25.268 (0.223)	25.293 (0.241)	25.193 (0.121)	25.229 (0.141)	23.036 (0.085)
		300	24.883 (0.198)	24.764 (0.226)	24.711 (0.108)	24.666 (0.122)	19.012 (0.082)
		500	24.888 (0.197)	24.980 (0.212)	24.753 (0.105)	24.755 (0.119)	16.776 (0.060)

All other setups remain consistent with those described in Section 3.1.

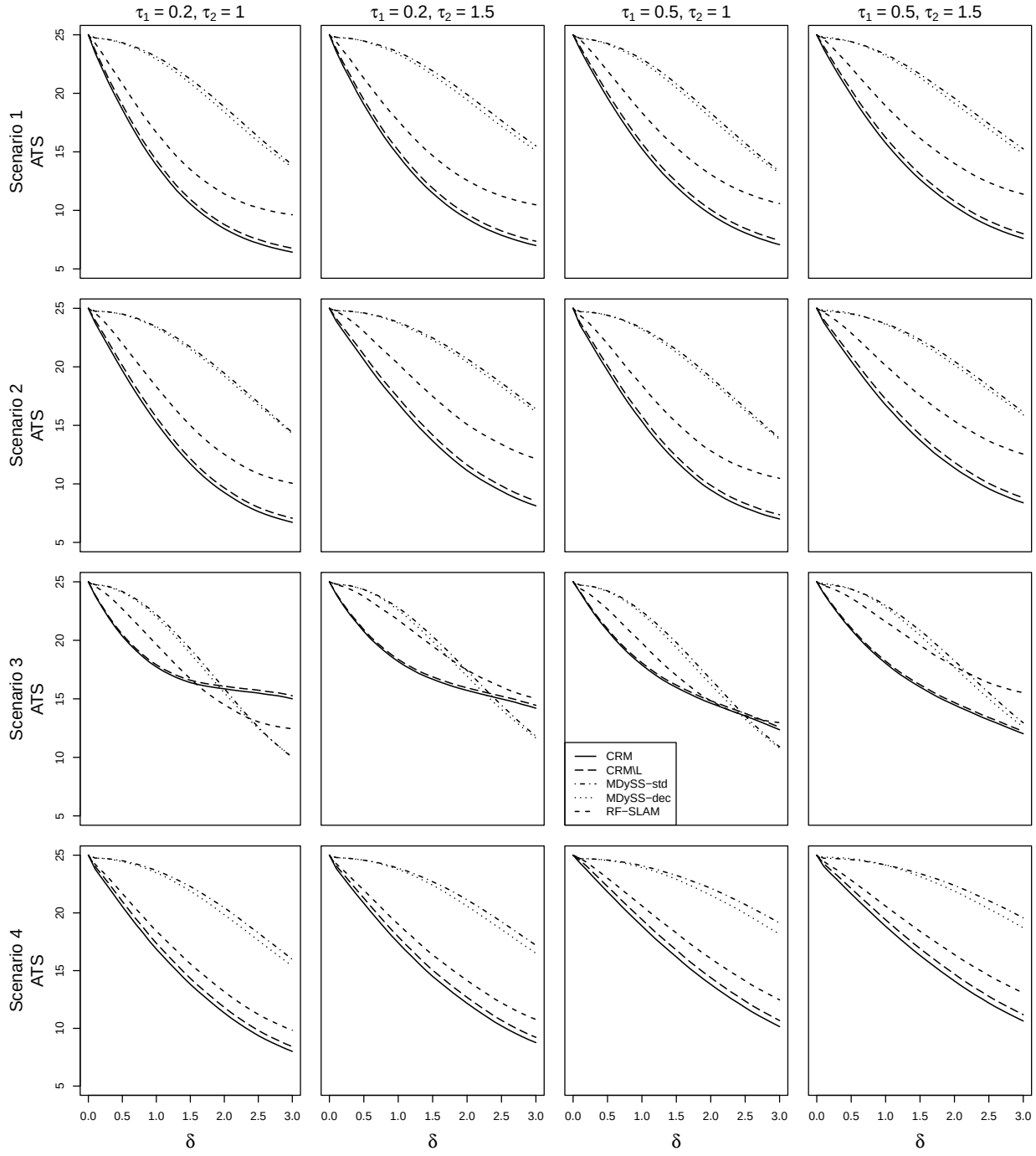
To ensure a fair comparison among different methods, their optimal ATS_1 values are evaluated in each case. The optimal ATS_1 for a given method in detecting a specific shift is defined as the minimal ATS_1 value achieved when adjusting the method's procedure parameter, while maintaining the nominal ATS_0 level (Qiu, 2014). Additionally, the threshold value for RF-SLAM is appropriately adjusted to ensure that the nominal ATS_0 is achieved in each case considered when no shift is present (i.e., $\delta = 0$).

Figure 2 presents the optimal ATS_1 values for the various methods across different scenarios, considering a range of values for τ_1, τ_2 and δ. In all these cases, the pure measurement error in Model (8) follows a standard normal distribution. The results in Figure 2 show that the proposed methods, CRM and its simplified version CRM\L, outperform the three competing methods, MDySS-std, MDySS-dec and RF-SLAM, in all cases within Scenarios 1, 2, and 4. In Scenario 3, the proposed methods maintain superior performance in most cases, except when the shift size is large (e.g., $\delta > 1.5$ when $\tau_1 = 0.2$ and $\tau_2 = 1$). Furthermore, CRM consistently outperforms CRM\L, and RF-SLAM outperforms the two MDySS methods in all cases.

These results can be intuitively explained as follows. The two MDySS methods monitor all risk factors directly and sequentially to detect relevant events. However, including risk factors that are unassociated with the events of interest would increase the variability of the charting statistics, reducing the effectiveness of these methods. In contrast, the proposed methods rely on an event risk framework that assigns different weights to risk factors based on their association with the events of interest. Additionally, the adaptive LASSO procedure in CRM further enhances detection by excluding irrelevant risk factors from process monitoring. Similar to CRM, RF-SLAM accounts for the influence of time-varying risk factors and the statuses of other events in estimating the conditional probability of the target event. This feature contributes to its superior performance compared to the two MDySS methods. However, because RF-SLAM generates signals based solely on the estimated conditional risk at the current time, without adequately incorporating the dynamic pattern of risk over time, its performance is generally inferior to that of the proposed CRM method in most cases considered.

The results in Figure 2 also reveal that MDySS-dec consistently outperforms MDySS-std, highlighting the advantage of decorrelating process observations in the MDySS framework. Figure A.1 in Appendix D displays the optimal ATS_1 values of the five methods in corresponding cases where the pure measurement error in Model (8) follows a standardized chi-square distribution. Similar conclusions can be drawn from these results.

Figure 2: Optimal ATS_1 values of the charts CRM, CRM\L, MDySS-std, MDySS-dec and RF-SLAM in various cases when the nominal ATS_0 value of each method is fixed at 25, the pure measurement error in Model (8) follows a normal distribution, and the observed risk factors have different mean shifts in Scenarios 1-4.



3.3 Impact of p on the performance of the proposed method

In this section, we investigate the impact of the dimensionality p on the estimation of Model (1) and the performance of the methods CRM, CRM\L, MDySS-std, MDySS-dec, and RF-SLAM discussed earlier. To this end, additional numerical studies were conducted for cases where $p = 20$ and $p = 30$. When generating risk factors for IC subjects, the first 10 dimensions of $\mathbf{X}_i(t)$ are generated according to Model (8). The remaining dimensions follow the same model structure, except that $\mu_{x,d}(t) = 0$ and $\sigma_{x,d} = 1$ for $d = 11, \ldots, p$. The p-dimensional random-effects term is generated similarly to that in Model (8), but with a larger exchangeable covariance matrix.

This study focuses on the case where $\tau_1 = 0.2$, $\tau_2 = 1$, and the pure measurement error follows a standard normal distribution. Given the observed risk factors, event statuses are generated as described in Model (9), with the exception that the index coefficients $\{\boldsymbol{\beta}_\kappa\}$ for dimensions 11 through p are set to zero. Furthermore, when generating data for OC subjects under the different scenarios described in Section 3.2, all components of the shift-multiplier vector beyond the tenth dimension are set to zero.

Table 3 presents the sum of squares of estimation error (SSE) for the estimated single-index coefficient vectors under scenarios with and without the adaptive LASSO penalty in model estimation. A similar relationship between SSE, its standard error, and the sample size is observed for both $p = 20$ and $p = 30$. However, for a fixed sample size, increasing the dimensionality p slightly increases the SSE. For a given sample size and value of p, incorporating the adaptive LASSO penalty reduces both the SSE and its standard error, demonstrating its effectiveness in improving model estimation.

For the case where $\tau_1 = 0.2$, $\tau_2 = 1$, the pure measurement error follows a standard normal distribution, and the conditional risk of interest is the one specified in Scenario 1, Tables A.9 and A.10 in Appendix D present the actual ATS_0 values for the five methods under $p = 20$ and $p = 30$, respectively. While the IC performance of CRM, CRM\L, MDySS-std, and MDySS-dec remains well-controlled in these higher-dimensional settings, the RF-SLAM method exhibits unreliable IC performance. Under the same settings with event occurrence specified earlier, Figure 3 displays the optimal ATS_1 values for the five methods across different OC models and values of p. The results suggest that similar conclusions regarding method performance hold as p increases from 10 to 30. Additionally, the optimal ATS_1 values for all methods show a slight increase with larger p. For a given shift magnitude δ, the performance gap in ATS_1 between CRM and CRM\L widens as p increases, highlighting the increasing importance of the adaptive LASSO procedure in high-dimensional settings.

Figure 3: Optimal ATS_1 values of the charts CRM, CRM\L, MDySS-std, MDySS-dec and RF-SLAM in various cases when the nominal ATS_0 value of each method is fixed at 25, $\tau_1 = 0.2$, $\tau_2 = 1$, the pure measurement error in Model (8) follows a normal distribution, and the dimension of the risk factors varies from $\{10, 20, 30\}$

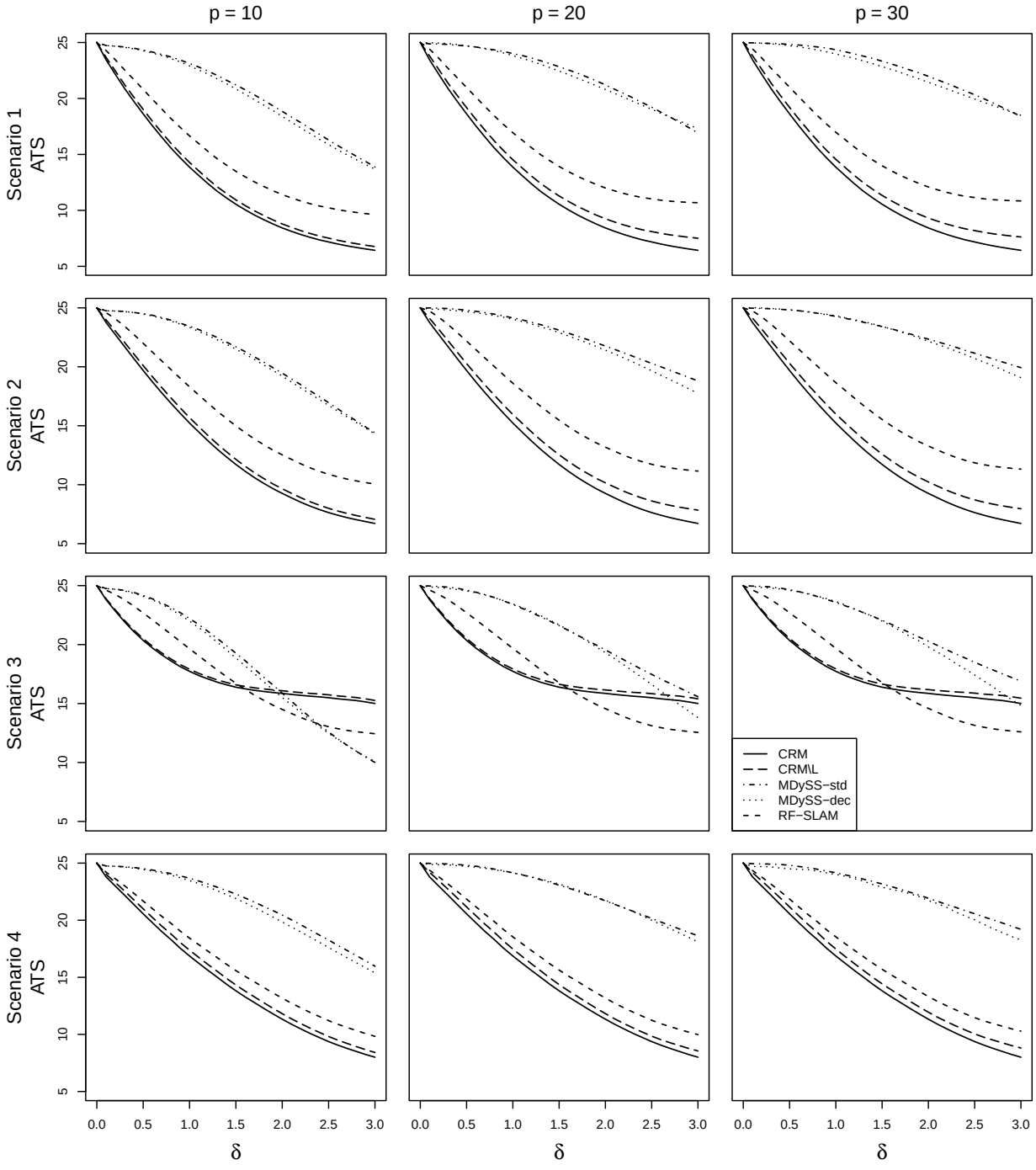


Table 3: Sum of squares of estimation error (SSE) of the estimated single-index coefficients in Model (9) when $\tau_1 = 0.2$, $\tau_2 = 1$, the pure measurement error in Model (8) has a normal distribution, and the dimension of the risk factors, p, equals 20 or 30. Standard errors of SSEs are provided in parentheses.

		$m = 5$		$m = 10$	
p	M_0	No penalty	LASSO penalty	No penalty	LASSO penalty
20	200	4.227 (0.251)	3.549 (0.155)	2.272 (0.076)	2.078 (0.067)
	300	1.260 (0.083)	1.006 (0.050)	0.601 (0.025)	0.571 (0.024)
	500	0.436 (0.028)	0.297 (0.013)	0.330 (0.013)	0.242 (0.009)
30	200	4.508 (0.267)	3.794 (0.161)	2.376 (0.081)	2.194 (0.071)
	300	1.439 (0.090)	1.154 (0.055)	0.662 (0.029)	0.599 (0.025)
	500	0.469 (0.031)	0.334 (0.015)	0.376 (0.015)	0.274 (0.010)

4 Case Study

In this section, we use data from the Framingham Heart Study (FHS) to demonstrate the application of the proposed method. The FHS is one of the most influential and longest-running cardiovascular studies in medical history. Initiated in 1948 in Framingham, Massachusetts, its primary goal was to identify risk factors for cardiovascular disease (CVD), which was then a leading cause of mortality. A comprehensive description of the FHS, including its study design, participant follow-up protocols, examination cycles, and phenotype and outcome measures, is available in Tsao and Vasan (2015).

In this example, we focus on detecting the onset of obesity, defined as a body mass index (BMI) exceeding 30, and CVD, which includes stroke, myocardial infarction, and congestive heart failure. The five considered risk factors for obesity and CVD are: systolic blood pressure (mmHg), diastolic blood pressure (mmHg), cholesterol level (mg/dL), blood glucose (mg/dL), and ventricular beats (beats per minute). To demonstrate the proposed method, we consider individuals without obesity or CVD at baseline and aim to detect the onset of either condition. In such cases, we jointly monitor i) the conditional risk of CVD given the absence of obesity and ii) the conditional risk of obesity given the absence of CVD.

The dataset comprises longitudinal observations of 1,078 participants aged 50 years and older, each with seven follow-up visits recording selected risk factors and diagnostic status for CVD and obesity. Of

these participants, 601 had no diagnosis of either disease (hereafter referred to as non-diseased participants), while 477 had at least one of the two diseases (diseased participants). The cohort was split into a training set and a test set: the training set included 301 non-diseased and 239 diseased participants, and the test set included 300 non-diseased and 238 diseased participants. The training set was used to fit the proposed single-index multinomial logistic regression model (1). Based on the fitted model, we computed the two conditional risks described above for all training participants. Using these estimates, we then derived the regular longitudinal patterns of the two conditional risks from the non-diseased participants in the training set.

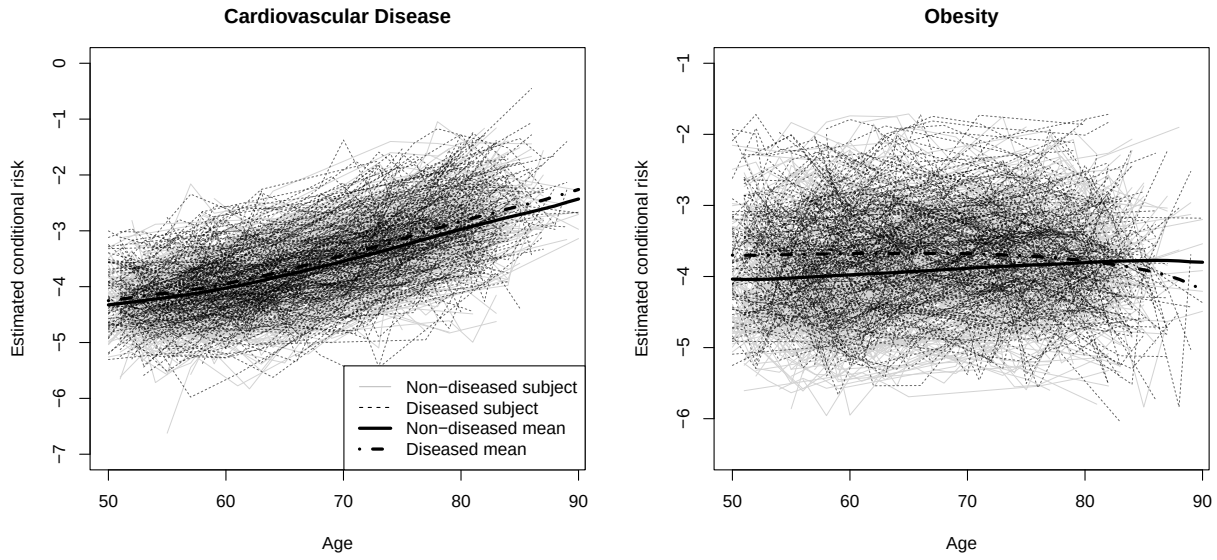
Figure 4 presents the estimated conditional risks for participants in the test set, computed using model (1) fitted with the adaptive LASSO penalty (see Expressions (2) and (3)). The figure reveals several patterns: i) The estimated conditional risks for both CVD and obesity, given the absence of the other disease, vary substantially across participants. ii) CVD conditional risks generally increase over time, whereas obesity conditional risks exhibit no clear temporal trend. In both cases, mean conditional risks are slightly higher for diseased than for non-diseased participants, with a modest exception for obesity among those aged 85 and older. These findings suggest that the conditional risks possess a meaningful degree of predictive and discriminatory ability for the diseases under study.

Five related methods – CRM, CRM\L, MDySS-std, MDySS-dec, and RF-SLAM – were applied to the test set to monitor each participant for early detection of CVD and obesity. For RF-SLAM, an ensemble of 1,000 trees was constructed. For a comprehensive comparison, the left panel of Figure 5 displays receiver operating characteristic (ROC) curves of the five methods, obtained with weighting parameter $\lambda = 0.2$ for the first four methods. Each ROC curve was obtained by computing the true positive rate (TPR) and false positive rate (FPR) across the test set at a given control limit (decision threshold) and then varying the control limit value. The plot shows that CRM and CRM\L outperform MDySS-dec, MDySS-std, and RF-SLAM, with CRM achieving a slightly larger Area Under the Curve (AUC) than CRM\L. Additionally, RF-SLAM outperforms MDySS-std in this setting. The right panel of Figure 5 shows the corresponding results when λ is optimized for each of the first four methods. Specifically, for a given FPR level, λ is selected to maximize the TPR for each method. In this case, CRM and CRM\L continue to outperform MDySS-dec, MDySS-std, and RF-SLAM, with CRM maintaining a slight advantage over CRM\L. Additionally, CRM achieves a 6.3% higher AUC than MDySS-dec and a 13.4% higher AUC than RF-SLAM, indicating substantial improvement in disease prediction performance.

Figure 6 presents the computed ATS_1 values of the five methods across various nominal ATS_0 levels.

The left panel shows results when the first four methods use the weighting parameter $\lambda = 0.2$, while the right panel displays results when λ is optimized. In both cases, CRM and CRM\L generally outperform MDySS-dec, MDySS-std, and RF-SLAM by achieving smaller ATS_1 values. Between CRM and CRM\L, the former yields slightly smaller ATS_1 values than the latter.

Figure 4: Estimated conditional risks for all participants in the test dataset, computed using the estimated model (1) by the adaptive LASSO procedure (3). Solid gray thin lines represent the estimated conditional risks for non-diseased participants, while black dashed thin lines represent those for diseased participants. In each plot, the solid and dot-dashed bold lines denote the sample means of the estimated conditional risks for non-diseased and diseased participants, respectively.



5 Concluding Remarks

In this paper, a new method for the early detection of one or multiple events conditional on the status of other events is developed. The method consists of several key steps. First, an innovative single-index multinomial logistic regression model is proposed to quantify all conditional event risks based on a training dataset collected in advance. The quantified risks are defined as linear combinations of the observed event risk factors, and their regular longitudinal patterns can be estimated nonparametrically from the observed data of subjects without the events of interest in the training dataset, as discussed in Section 2.2. Next, to detect events given the status of other events for a new subject, the relevant conditional event risks are computed. After proper data standardization and decorrelation, these computed conditional event risks are monitored

Figure 5: ROC curves for the five methods CRM, CRM\L, MDySS-std, MDySS-dec and RF-SLAM used to detect CVD and obesity in the 538 participants from the test dataset. The left panel shows the results when the weighting parameter λ is set to 0.2 for the first four methods, while the right panel displays the results when λ is optimized for each of these methods. The corresponding AUC values are also provided. The dashed gray diagonal line in each panel serves as a reference.

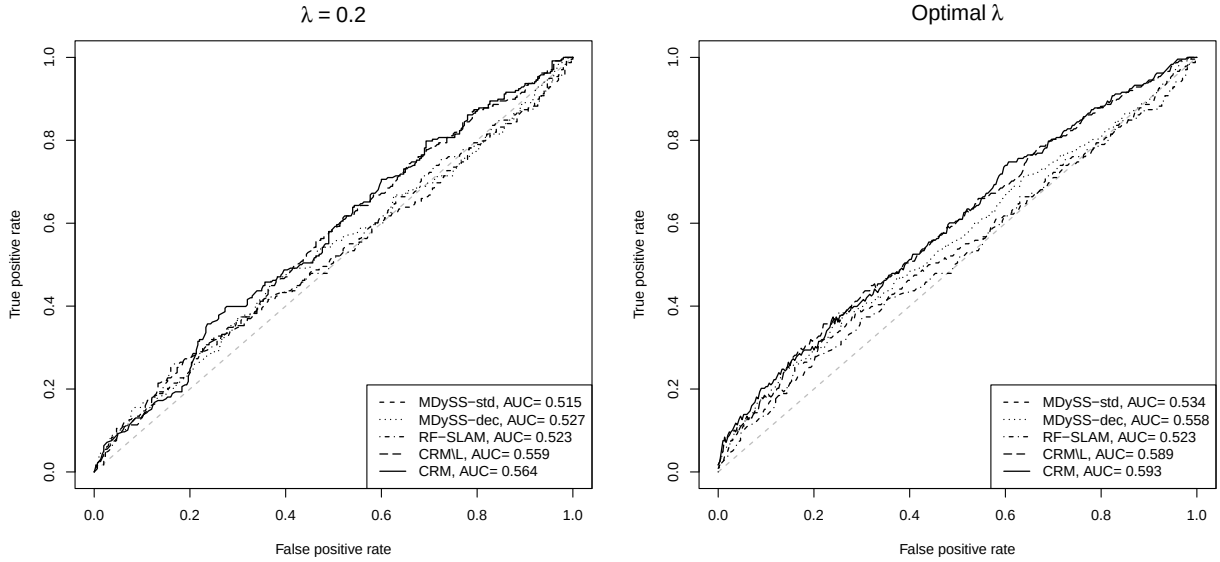
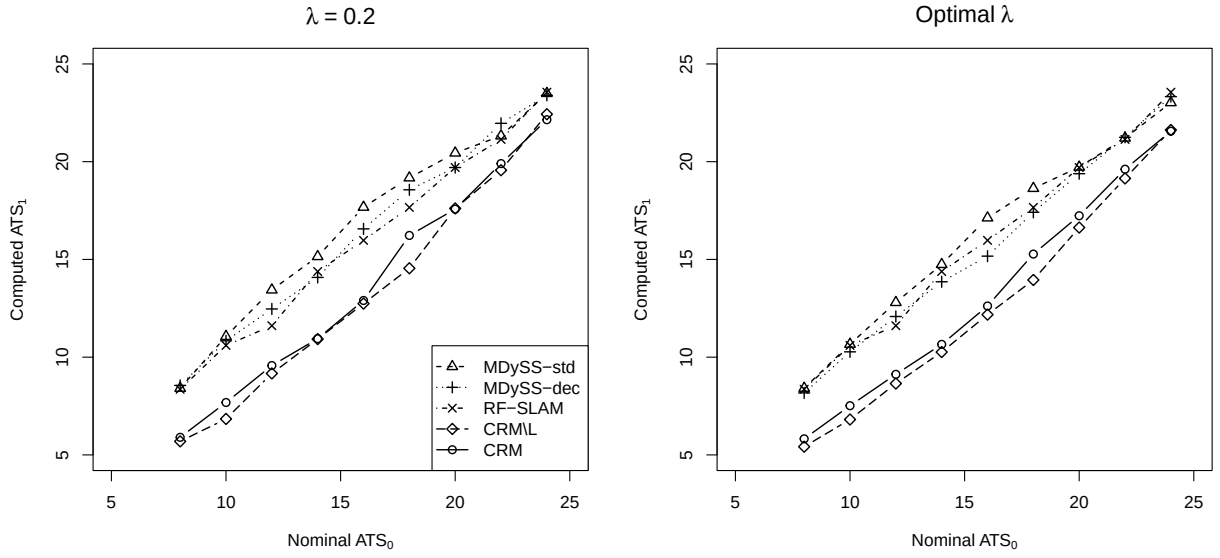


Figure 6: Computed ATS_1 values for the five methods CRM, CRM\L, MDySS-std, MDySS-dec and RF-SLAM used to detect CVD and obesity in the 538 participants from the test dataset under different nominal levels of ATS_0. The left panel shows the results when the weighting parameter λ is set to 0.2 for the first four methods, while the right panel displays the results when λ is optimized for each of these methods.



sequentially using the control chart (6)-(7), which can accommodate unequally spaced observation times. Numerical results show that the proposed method performs well in various scenarios.

In the single-index multinomial logistic regression modeling procedure described in Section 2.1, all possible conditional risks can theoretically be computed from the estimated Model (1). However, when one or more rare events are involved in the conditioning set, the resulting estimated conditional risks would exhibit relatively high variability because of small number of IC subjects in the training data for estimating the conditional risks. This increased variability may lead to less reliable detection of other events given these rare conditions. Therefore, the proposed method is expected to be more effective when commonly observed events are used as the conditioning variables. Additionally, in practice, it may occur that certain events are impossible given specific conditioning events (e.g., a car with a dead battery cannot display a warning light). In such cases, the corresponding conditional risks, computed from marginal risks based on the estimated Model (1) (cf. Section 2.2 for examples), may not be well defined because the probabilities of the associated event combinations are zero. As a result, those event combinations should be excluded from the formulation and estimation of Model (1).

There are several other aspects of the proposed method that warrant further investigation. For instance, when the number of events of interest q is large, estimating the single-index multinomial logistic regression model (1) can become computationally intensive. This is due to the model's flexibility, which allows for different link functions and single-index coefficients across a total of $q^* - 1 = 2^q - 1$ non-redundant event statuses. Therefore, developing more computationally efficient estimation procedures in such settings is essential. One possible approach to reduce the computational burden is to simplify the multinomial logistic regression model. For example, one could assume identity link functions or impose identical single-index coefficients across all event statuses. Exploring reasonable trade-offs between such simplified modeling strategies and the overall performance of the proposed method in early detection of multiple events is an important area for future research. Another issue concerns variable selection using the adaptive LASSO approach (cf. Expression (3)) when estimating Model (1). In its current form, the proposed method selects risk factors independently for each possible event status (i.e., each k in Model (1)). However, in practice, each event may have its own specific set of important risk factors. A promising direction is to group the $q^* - 1$ equations in Model (1) by individual events and apply a group-LASSO procedure (Yuan and Lin, 2006) for risk factor selection.

Supplementary Materials

Code_and_Data.zip: This zip file contains some computer codes to implement the proposed method and the real data used in the paper.

supp.pdf: This supplementary file contains i) some technical details about the estimation of the single-index multinomial logistic regression model (1), ii) guidelines on optimization of the objective function in (3), iii) estimation of the regular longitudinal pattern, and iv) some extra numerical results.

Acknowledgments

The authors thank the editor, the associate editor, and two referees for many constructive comments and suggestions, which greatly improved the quality of the paper.

References

Ashraf, A., Ali, S., and Shah, I. (2021). Online disease risk monitoring using DEWMA control chart. *Expert Systems with Applications*, **180**, 115059.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**(438), 477–489.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems, 29.

Chowdhury, S.K. and Sinha, S.K. (2015). Semiparametric marginal models for binary longitudinal data. *International Journal of Statistics and Probability*, **4**(3), 107.

Cui, X., Härdle, W. K., and Zhu, L. (2011). The EFM approach for single-index models. *The Annals of Statistics*, **39**(3), 1658—-1688.

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

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, **103**, 103–118.

Li, J. and Qiu, P. (2016). Nonparametric dynamic screening system for monitoring correlated longitudinal data. *IIE Transactions*, **48**(8), 772–786.

Li, J. and Qiu, P. (2017). Construction of an efficient multivariate dynamic screening system. *Quality and Reliability Engineering International*, **33**(8), 1969–1981.

Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, **98**(2), 355–370.

Li, W., Li, Y., Tsung, F., and Wu, C. (2025). Real-time monitoring of dynamic tensor data with longitudinal patterns: A tensor graphical LASSO approach. *Technometrics*, 1–23.

Li, Y., Wu, C., Li, W., Tsung, F., and Guo, J. (2024). Dynamic modeling and online monitoring of tensor data streams with application to passenger flow surveillance. *The Annals of Applied Statistics*, **18**, 1789–1814.

Liu, P., Du, J., Zang, Y., Zhang, C., and Wang, K. (2023). In-profile monitoring for cluster-correlated data in advanced manufacturing system. *Journal of Quality Technology*, **55**, 195–219.

Matheus, A. S. D. M., Tannus, L. R. M., Cobas, R. A., Palma, C. C. S., Negrato, C. A., and Gomes, M. D. B. (2013). Impact of diabetes on cardiovascular disease: an update. *International journal of hypertension*, **2013**(1), 653789.

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

Qiu, P. (2024). *Statistical Methods for Dynamic Disease Screening and Spatio-Temporal Disease Surveillance*. Boca Raton, FL, Chapman & Hall/CRC.

Qiu, P. and Xiang, D. (2014). Univariate dynamic screening system: an approach for identifying individuals with irregular longitudinal behavior. *Technometrics*, **56**(2), 248–260.

Qiu, P. and Xiang, D. (2015). Surveillance of cardiovascular diseases using a multivariate dynamic screening system. *Statistics in Medicine*, **34**(14), 2204–2221.

Qiu, P. and You, L. (2022). Dynamic disease screening by joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **71**(5), 1158–1180.

Salisbury, C., Johnson, L., Purdy, S., Valderas, J. M., and Montgomery, A. A. (2011). Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. *British Journal of General Practice*, **61**(582), e12–e21.

Scheibe, K. P., and Blackhurst, J. (2018). Supply chain disruption propagation: a systemic risk and normal accident theory perspective. *International Journal of Production Research*, **56**(1-2), 43-59.

Tian, Z. and Qiu, P. (2024a). Generalized single index modeling of longitudinal data with multiple binary responses. *Statistics in Medicine*, **43**(19), 3578–3594.

Tian, Z. and Qiu, P. (2024b). Online risk monitoring of multiple diseases by a dynamic screening system. *arXiv*.

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

Tsao, C. W., and Vasan, R. S. (2015). Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology*, **44**(6), 1800–1813.

Tsochev, G., Trifonov, R., Nakov, O., Manolov, S. and Pavlova, G. (2020). Cyber security: Threats and challenges. *2020 International Conference Automatics and Informatics*, 1–6.

Wongvibulsin, S., Wu, K. C., and Zeger, S. L. (2020). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Medical Research Methodology*, **20**, 1–14.

Xiang, D., Qiu, P., and Pu, X. (2013). Nonparametric regression analysis of multivariate longitudinal data. *Statistica Sinica*, **23**(2), 769–789.

Yi, G. Y., He, W., and Liang, H. (2009). Analysis of correlated binary data under partially linear single-index logistic models. *Journal of Multivariate Analysis*, **100**(2), 278–290.

You, L. and Qiu, P. (2019). Fast computing for dynamic screening systems when analyzing correlated data. *Journal of Statistical Computation and Simulation*, **89**(3), 379–394.

You, L. and Qiu, P. (2020). An effective method for online disease risk monitoring. *Technometrics*, **62**(2), 249–264.

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**(1), 49–67.

Zhang, W., He, Z., He, S., Zhanwen, N., Shang, Y., and Song, L. (2025). In-profile monitoring on univariate profile with application to industrial busbar. *IISE Transactions*, 1–16.

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