Detection of Multiple Diseases by Sequential Risk Monitoring Using a Dynamic Screening System

Abstract

Early detection and prevention of diseases, particularly chronic conditions such as cardiovascular diseases (CVD), are essential for improving patient outcomes and reducing the burden on healthcare systems. Since disease risk factors are typically observed sequentially over time, early detection inherently involves *sequential decision-making*. This process is complicated by challenges such as time-varying data distributions and serial correlations in the observed data. While existing statistical process control (SPC) methods can effectively detect a single disease, there is an increasing demand for approaches capable of addressing multiple diseases simultaneously. In practice, patients often present with multimorbidities, such as myocardial infarction (MI) and stroke, which share common risk factors like high blood pressure, elevated cholesterol, and abnormal blood glucose levels. Developing a unified framework to jointly address these conditions can enhance the efficiency and cost-effectiveness of disease detection and prevention strategies. This paper introduces a novel dynamic screening system for the early detection of multiple diseases. The proposed method quantifies a patient's risk for various diseases at each observation point using a flexible longitudinal data modeling approach. It then compares the patient's risk profile with the baseline risk pattern of non-diseased individuals. By sequentially monitoring cumulative risk deviations, the system triggers a signal when a patient exhibits abnormally high risk for any disease of interest. Numerical studies demonstrate the robustness and effectiveness of the proposed approach under various scenarios. The method is applied to a dataset from the Framingham Heart Study. The results showcase its ability to identify early risks for MI and stroke, two major manifestations of CVD, highlighting its potential as a valuable tool in clinical practice.

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

1 Introduction

Chronic diseases, particularly cardiovascular diseases (CVD), represent a major global health burden, leading to significant morbidity, disability, and mortality (Christensen et al., 2009; Kennedy et al., 2014; Roth et al., 2020). Early detection of such conditions is critical for improving patient outcomes and reducing the overall burden on healthcare systems (Greenberg and Pi-Sunyer, 2019). As CVD often develop over many years without noticeable symptoms, identifying individuals at risk before the onset of acute clinical events

such as myocardial infarction (MI) or stroke allows for timely intervention. It also helps prevent or delay the progression of these diseases, reducing the risk of severe complications and improving patients' quality of life (Bauer et al., 2014; Hacker, 2024; Perel et al., 2015).

The Framingham Heart Study (FHS) is one of the longest-running and most comprehensive epidemiological studies of CVD and related chronic conditions (Andersson et al., 2019). Launched with an initial cohort of over 5,000 participants from Framingham, Massachusetts, the study has expanded over time to include multiple generations of participants. This expansion has provided researchers with invaluable insights into the genetic, environmental, and lifestyle factors contributing to CVD (Fox, 2010; Hajar, 2017; Staerk et al., 2020). Data collected through this longitudinal study has underpinned many modern guidelines and risk assessment models used globally by healthcare professionals. For example, the Framingham Risk Score, developed from the study, estimates an individual's 10-year risk of developing coronary heart disease based on their unique risk profile (Lloyd-Jones, 2004). While identifying risk factors is crucial, it is insufficient for effectively preventing or managing CVD. Recognizing risk factors highlights who might be at greater risk, but it does not inherently lead to timely or targeted interventions. The critical next step is implementing robust screening and early detection strategies for multiple CVD manifestations, such as MI and stroke. This paper focuses on addressing this essential need.

In the literature, most early disease detection strategies focus on a single target disease and rely on screening a predefined set of disease risk factors. As discussed by Zhao and Wu (2008) and Ma et al. (2012), one common approach involves constructing pointwise confidence intervals for the mean functions of these risk factors using data from non-diseased individuals. A new patient is then flagged as having the disease of interest if their observed risk factors fall outside the constructed confidence intervals at a given observation time. To implement this strategy, nonparametric longitudinal modeling approaches, such as those proposed by Li (2011) and Xiang et al. (2013), can be employed to construct the necessary confidence intervals. However, these methods determine a patient's disease status solely based on observed data at the current time point, without efficiently utilizing historical data. This limitation poses a significant challenge for early disease detection, which inherently involves sequential decision-making. Since disease risk factors are typically observed sequentially over time, methods that fail to incorporate historical data may lack the sensitivity and effectiveness needed for timely disease detection.

A growing body of recent literature has focused on modeling disease probability or risk dynamically using modern machine learning techniques. The mixed-effects Bayesian additive regression trees (mixed-BART) provide flexible nonlinear modeling with uncertainty quantification for longitudinal and clustered

data (Spanbauer and Sparapani, 2021). Tree-based ensemble methods, including GPBoost, combine gradient boosting with random effects or Gaussian process components to capture nonlinear covariate effects and within-subject correlation in longitudinal risk prediction (Sigrist, 2022). In parallel, deep learning models for electronic health records, such as the REverse Time AttentIoN (RETAIN) model and more recent transformer-based architectures, have been developed to learn complex temporal dependencies and time-varying risk patterns from high-dimensional patient histories (Choi et al., 2016; Yang et al., 2023). While these methods can yield accurate estimates of time-dependent probabilities or risks of the target disease, they are primarily designed for offline prediction rather than online monitoring and detection. In practice, high-risk individuals are typically identified using ad hoc thresholding rules, ranking criteria, or visually observed upward trends in predicted risk trajectories. As a result, these approaches do not provide a formal sequential monitoring framework for controlling false-alarm rates or quantifying the long-run operating characteristics of detection procedures.

Statistical process control (SPC) charts offer a powerful analytical framework for addressing sequential decision-making problems (Qiu, 2014). Conventional SPC charts are primarily designed for monitoring a single sequential process under the assumption that IC process observations are independent and identically distributed (i.i.d.) at different observation times. However, this assumption is often violated in modern applications, where data streams are multivariate, dynamic, and serially correlated, even under IC operating conditions. To address these challenges, A sizable recent SPC literature has developed monitoring procedures for multivariate and serially correlated data by incorporating dependence-aware charting statistics and calibration schemes. Common strategies include model-based residual monitoring, where a time-series model such as ARIMA/VAR or related dynamic models is fitted to IC data and control charts are applied to the resulting decorrelated residuals (Yao et al., 2023); control charts coupled with recursive standardization/decorrelation and resampling calibration (i.e., block bootstrap) to achieve stable false-alarm control under dependence (Xue and Qiu, 2021); and copula-parameterized CUSUM schemes that incorporate the correlated data into the likelihood ratio formulation (Perry, 2024). Nevertheless, these methods are still developed within the traditional SPC paradigm, where the goal is to monitor a single process or system over time. In contrast, early disease detection problems involve monitoring a group of patient-specific longitudinal data streams, where each subject exhibits heterogeneous dynamics and disease onset times. This fundamental difference motivates the development of frameworks tailored to patient-level monitoring.

To address these challenges, Qiu and Xiang (2014, 2015) proposed a dynamic screening system (DySS) for the early detection of a single disease. DySS accommodates time-varying data distributions and certain

serial correlation patterns by first estimating the *regular longitudinal pattern* of disease risk factors from an IC dataset comprising data from non-diseased individuals. The system then detects the target disease for a new patient by comparing the observed risk factors of the patient with the estimated regular pattern. Several modifications and generalizations of this approach have been proposed in subsequent work, including Li and Qiu (2016, 2017), Qiu et al. (2018), and You and Qiu (2019). However, these methods share common limitations. In practice, different disease risk factors often vary in their relevance to the occurrence of the disease; some risk factors may have minimal influence. These methods fail to account for such heterogeneity, treating all risk factors equally. Additionally, the joint monitoring of all risk factors may become inefficient when the number of factors is large. To address these issues, You and Qiu (2020) introduced an alternative approach that quantifies *disease risk* at a given time point using a survival model and then monitors the quantified risk. Qiu and You (2022) improved upon this by integrating survival and longitudinal data modeling. Despite these advances, survival-model-based methods have their own limitations, as they typically exclude data observed after the first occurrence of the target disease, potentially reducing their effectiveness. For a comprehensive overview of existing DySS methods, readers are referred to the recent monograph by Qiu (2024).

Multimorbidity, particularly the co-occurrence of multiple CVD conditions such as coronary heart disease, intermittent claudication, congestive heart failure, and stroke, is common in practice (Tsao and Vasan, 2015). Since these conditions often share risk factors, pathophysiological mechanisms, and treatment strategies, jointly detecting and managing them is more efficient and cost-effective than addressing each separately (Salisbury et al., 2011). However, existing methods discussed earlier are designed for detecting a single disease and cannot be easily generalized to manage multiple diseases. Thus, there is a pressing need for new approaches that can effectively accommodate the mutual correlations among multiple diseases, account for their heterogeneous associations with risk factors, and overcome the key limitations of existing methods. To bridge this gap, we propose a novel method in this paper for the effective early detection of multiple diseases.

The remainder of this paper is structured as follows. Section 2 provides a detailed description of the proposed method. Section 3 presents simulation results to evaluate the numerical performance of the method. Section 4 demonstrates the application of the proposed method to a dataset from FHS for the early detection of MI and stroke, two critical manifestations of CVD. Section 5 offers several concluding remarks. Additional technical details and supplementary numerical results are provided in the supplementary materials.

2 Proposed Method

The proposed method consists of the following four major steps: i) a multivariate single-index logistic regression model is fitted using training data that includes observed longitudinal data from both diseased and non-diseased individuals, ii) the fitted model is used to quantify the risks of the targeted diseases for a given patient, iii) the quantified risks are standardized and decorrelated with all historical data from the patient, and iv) a multivariate control chart is constructed to sequentially monitor the decorrelated risks over time. Each of these steps is described in detail in the following subsections.

2.1 Multivariate single-index logistic regression modeling

Assume that a training dataset is available and contains longitudinal observations of p disease risk factors and q binary response variables that indicate the status of q diseases in concern of M subjects. For the ith subject, Y_{ijk} denotes its jth observation of the kth response, for $j = 1, \ldots, m_i$, $k = 1, \ldots, q$, and $i = 1, \ldots, M$, where m_i denotes the number of longitudinal observations at times $\{t_{ij} \in [T_0, T_1], j = 1, \ldots, m_i\}$ in a given time interval $[T_0, T_1]$. Similarly, $\{\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^T, j = 1, \ldots, m_i, i = 1, \ldots, M\}$ denote the corresponding observations of the p disease risk factors. These observed data are assumed to follow the multivariate single-index logistic regression model below:

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \psi_k\left(\boldsymbol{\beta}_k^T \mathbf{X}_{ij}\right) + \mathbf{G}_k^T \mathbf{b}_i, \text{ for } j = 1, \ldots, m_i, i = 1, \ldots, M, k = 1, \ldots, q, \tag{1}$$

where $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \ldots, \mathbf{b}_{iq}^T)^T$ is a vector of random-effects for the ith subject, $\mathbf{b}_{ik} = (b_{ik1}, \ldots, b_{iks})^T$ denotes the subject-specific random-effects corresponding to the kth response variable, $\mathbf{G}_k$ is a design vector of the random-effects, $\mu_{ijk} = \mathbb{E}(Y_{ijk}|\mathbf{X}_{ij}, \mathbf{b}_i)$, $\psi_k(\cdot)$ is an unknown link function, and $\boldsymbol{\beta}_k$ is a p-dimensional vector of index coefficients. The random-effects $\{\mathbf{b}_i, i = 1, \ldots, M\}$ are used to account for between-subject variation and within-subject data correlation. They are routinely assumed to be i.i.d. with the common distribution $N_{qs}(\mathbf{0}, \boldsymbol{\Sigma}_b)$. For example, when only disease-specific random intercepts are included in the log-odds model, we define $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^T$ as a q-dimensional random vector and $\mathbf{G_k}$ as the q-dimensional design vector with a one at the kth entry and zeros elsewhere.

Single-index modeling is a commonly used tool for constructing composite indices that summarize a pool of explanatory variables that are potential predictors of the outcomes of interest. In Model (1), different diseases are allowed to have different single-indices $\boldsymbol{\beta}_k^T \mathbf{X}$, accommodating the heterogeneous association between different diseases and the pool of risk factors. Inclusion of the nonparametric link functions $\{\psi_k(\cdot)\}$

is to make the model flexible. For model identifiability, the index coefficients are assumed to satisfy the conditions that $\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k = 1$ and the first element of $\boldsymbol{\beta}_k$ is positive, for each k.

In the literature, generalized single-index modeling for cases with a single binary response and independent observations has been extensively studied, with notable contributions by Carroll et al. (1997) and Cui et al. (2011). Extensions of these methods to longitudinal data with a single response have been explored within the framework of generalized estimating equations (GEE) in works such as Chowdhury and Sinha (2015) and Yi et al. (2009). However, estimating Model (1) is more complex, as it involves multiple mutually correlated binary responses and multiple single indices. If all link functions $\{\psi_k(\cdot)\}$ in Model (1) are identity functions, the model simplifies to a multivariate generalized linear mixed-effects model (GLMM) with a logit linkage function, which can be estimated using methods discussed in McCulloch (1997). In the current problem, the link functions are unknown smooth functions, making the model estimation significantly more challenging. In such cases, we suggest estimating the model using a Monte Carlo expectation-maximization (EM) algorithm combined with a local kernel smoothing procedure as described in Tian and Qiu (2024). The detailed methodology for this approach is summarized below.

For Model (1), its log-likelihood function has the following expression:

$$\begin{aligned} & l(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{b}) \\ = \; & \log\{f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\}) f(\mathbf{b}|\boldsymbol{\Sigma}_b)\} \\ = \; & \sum_{i=1}^{M}\sum_{j=1}^{m_i}\sum_{k=1}^{q}\left[Y_{ijk}\left\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) + \mathbf{G}_k^T\mathbf{b}_i\right\} - \log\left[1+\exp\left\{\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij}) + \mathbf{G}_k^T\mathbf{b}_i\right\}\right]\right] \\ & + \sum_{i=1}^{M}\left\{-\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_b| - \frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}_b^{-1}\mathbf{b}_i\right\}, \end{aligned}$$

where $\mathbf{Y} = (\mathbf{Y}_{\bullet\bullet 1}^T, \ldots, \mathbf{Y}_{\bullet\bullet q}^T)^T$ is a vector of all observations of the response variables, $\mathbf{Y}_{\bullet\bullet k} = (\mathbf{Y}_{1\bullet k}^T, \ldots, \mathbf{Y}_{M\bullet k}^T)^T$, $\mathbf{Y}_{i\bullet k} = (Y_{i1k}, \ldots, Y_{im_ik})^T$, for all i and k, $\mathbf{X}$ is the $(\sum_{i=1}^{M} m_i) \times p$ matrix of all observed data of the covariates whose $[\sum_{l=1}^{i-1}(l-1)m_l + j]$-th row is $\mathbf{X}_{ij}^T$, for all i and j, $\mathbf{X}_i$ is the $m_i \times p$ matrix whose jth row is $\mathbf{X}_{ij}^T$, and $\boldsymbol{\theta}$ denotes a collection of $\{\boldsymbol{\beta}_k\}$, $\boldsymbol{\Sigma}_b$, and the unknown link functions $\{\psi_k(\cdot)\}$ in Model (1).

For a given k, by the Taylor's expansion, when $\mathbf{X}_{i'j'}$ is chosen such that $\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}$ is close to $\boldsymbol{\beta}_k^T\mathbf{X}_{ij}$, $\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'}) + \psi_k'(\boldsymbol{\beta}_k^T\mathbf{X}_{i'j'})\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})$ would be a local linear approximation of $\psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij})$. Thus, if the values of ψ_k and ψ_k' evaluated at $\boldsymbol{\beta}_k^T\mathbf{X}_{ij}$ are denoted as a_{ijk} and c_{ijk}, respectively, the function $\log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\psi_k\}, \{\boldsymbol{\beta}_k\})$ can be approximated by the following kernel-weighted objective:

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

where

$$w_{iji'j'k} = \frac{K_{h_{0k}}[\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})]}{\sum_{i=1}^{M}\sum_{j=1}^{m_i} K_{h_{0k}}[\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})]},$$

$\eta_{iji'j'k} = a_{i'j'k} + c_{i'j'k}\boldsymbol{\beta}_k^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + \mathbf{G}_k^T\mathbf{b}_i$, $\mathbf{a}_k = (a_{11k}, a_{12k}, \ldots, a_{1m_1k}, a_{21k}, \ldots, a_{Mm_Mk})^T$, $\mathbf{c}_k = (c_{11k}, c_{12k}, \ldots, c_{1m_1k}, c_{21k}, \ldots, c_{Mm_Mk})^T$, $K_{h_{0k}}(\cdot) = K(\cdot/h_{0k})/h_{0k}$, for $k = 1, \ldots, q$, $K(\cdot)$ is a density kernel function, and $\{h_{0k}\}$ are q bandwidths for approximating the q link functions. The kernel function can be chosen to be the Epanechnikov kernel function, i.e., $K(u) = 0.75(1-u^2)I(|u| \leq 1)$, because of its good properties discussed in Epanechnikov (1969).

To obtain parameter estimates by the Monte Carlo EM algorithm, if $\{\widetilde{\boldsymbol{\beta}}_k\}$, $\widetilde{\boldsymbol{\Sigma}}_b$, $\{\widetilde{\mathbf{a}}_k\}$ and $\{\widetilde{\mathbf{c}}_k\}$ are the parameter estimates obtained in the previous iteration, the algorithm obtains the estimates $\{\widehat{\boldsymbol{\beta}}_k\}$, $\widehat{\boldsymbol{\Sigma}}_b$, $\{\widehat{\mathbf{a}}_k\}$ and $\{\widehat{\mathbf{c}}_k\}$ in the current iteration by the following updating equations: for $k = 1, \ldots, q$,

$$\begin{aligned}
(\widehat{\boldsymbol{\beta}}_k^T, \widehat{\mathbf{a}}_k^T, \widehat{\mathbf{c}}_k^T)^T &= \underset{\boldsymbol{\beta}_k, \mathbf{a}_k, \mathbf{c}_k}{\operatorname{argmax}} \mathbb{E}_{\mathbf{b}|\mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_b, \{\widetilde{\boldsymbol{\beta}}_k\}, \{\widetilde{\mathbf{a}}_k\}, \{\widetilde{\mathbf{c}}_k\}} \left\{ \log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}, \{\boldsymbol{\beta}_k\}, \{\mathbf{a}_k\}, \{\mathbf{c}_k\}) \middle| \mathbf{Y}, \widehat{\boldsymbol{\Sigma}}_b, \{\widetilde{\boldsymbol{\beta}}_k\}, \{\widetilde{\mathbf{a}}_k\}, \{\widetilde{\mathbf{c}}_k\} \right\} \\
&= \underset{\boldsymbol{\beta}_k, \mathbf{a}_k, \mathbf{c}_k}{\operatorname{argmax}} \frac{1}{B} \sum_{l=1}^{B} \log f(\mathbf{Y}|\mathbf{X}, \mathbf{b}^{(l)}, \{\widetilde{\boldsymbol{\beta}}_k\}, \{\widetilde{\mathbf{a}}_k\}, \{\widetilde{\mathbf{c}}_k\}), \\
\widehat{\boldsymbol{\Sigma}}_b &= \underset{\boldsymbol{\Sigma}_b}{\operatorname{argmax}} \mathbb{E}_{\mathbf{b}|\mathbf{Y}, \widetilde{\boldsymbol{\Sigma}}_b, \{\widehat{\boldsymbol{\beta}}_k\}, \{\widehat{\mathbf{a}}_k\}, \{\widehat{\mathbf{c}}_k\}} \left\{ \log f(\mathbf{b}|\boldsymbol{\Sigma}_b) \middle| \mathbf{Y}, \widetilde{\boldsymbol{\Sigma}}_b, \{\widehat{\boldsymbol{\beta}}_k\}, \{\widehat{\mathbf{a}}_k\}, \{\widehat{\mathbf{c}}_k\} \right\} \\
&= \underset{\boldsymbol{\Sigma}_b}{\operatorname{argmax}} \frac{1}{B} \sum_{l=1}^{B} \log f(\mathbf{b}^{(l)}|\widetilde{\boldsymbol{\Sigma}}_b),
\end{aligned}$$

where $B > 0$ is a large integer and $\{\mathbf{b}^{(l)}, l = 1, \ldots, B\}$ are random samples obtained from the posterior distribution of $\mathbf{b}$ given the current estimates of other parameters. In Appendix A, we provide the strategy for generating these posterior samples, guidelines for choosing bandwidths for estimating the link functions, and a detailed description of the iterative procedures for estimating the parameters in Model (1).

2.2 Multivariate disease risks and their regular longitudinal patterns

From Model (1), for the ith patient, the jth observation $\mathbf{X}_{ij}$ of the disease risk factors affects the likelihood of the kth disease through $r_{ik}(t_{ij}) = \psi_k(\boldsymbol{\beta}_k^T\mathbf{X}_{ij})$, for each i, j, k. Then, $r_{ik}(t_{ij})$ is defined to be the ith patient's *risk to the kth disease* based on the observed disease risk factors $\mathbf{X}_{ij}$, and $\mathbf{r}_i(t_{ij}) = (r_{i1}(t_{ij}), \ldots, r_{iq}(t_{ij}))^T$ is the ith patient's multivariate risk to the q diseases in concern at time t_{ij}. After Model (1) is estimated, the estimated multivariate risk for patient i at time t_{ij} is denoted as $\widehat{\mathbf{r}}_i(t_{ij}) = (\widehat{r}_{i1}(t_{ij}), \ldots, \widehat{r}_{iq}(t_{ij}))^T = (\widehat{\psi}_1(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}_{ij}), \ldots, \widehat{\psi}_q(\widehat{\boldsymbol{\beta}}_q^T\mathbf{X}_{ij}))^T$.

Without loss of generality, it is assumed that the first M_1 subjects in the training data are non-diseased (IC) individuals who do not have any targeted diseases in $[T_0, T_1]$, and the rest $M - M_1$ subjects are diseased individuals who have at least one disease in concern in $[T_0, T_1]$. Then, the estimated multivariate risks of the first M_1 subjects can be used to estimate the *regular longitudinal pattern* of the estimated multivariate risks, defined to be the longitudinal pattern of the disease risks for non-diseased people. To this end, the estimated risks of the non-diseased subjects are assumed to follow the multivariate nonparametric longitudinal model:

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

where $\boldsymbol{\mu}_r(t_{ij}) = (\mu_{r,1}(t_{ij}), \ldots, \mu_{r,q}(t_{ij}))^T$ is the mean of $\widehat{\mathbf{r}}_i(t_{ij})$, and $\boldsymbol{\varepsilon}_{r,i}(t_{ij}) = (\varepsilon_{r,i1}(t_{ij}), \ldots, \varepsilon_{r,iq}(t_{ij}))^T$ is the zero-mean error. For any $s, t \in [T_0, T_1]$, the covariance structure of the error terms is described by the covariance matrix function $\mathbf{V}_r(s,t) = \text{cov}(\boldsymbol{\varepsilon}_{r,i}(s), \boldsymbol{\varepsilon}_{r,i}(t))$. Then, a modified version of the model estimation procedure discussed in Xiang et al. (2013) is proposed to estimate the mean function $\boldsymbol{\mu}_r(t)$ and the covariance function $\mathbf{V}_r(s,t)$, which is described below.

The estimate of $\boldsymbol{\mu}_r(t)$ is obtained by computing the following local linear kernel smoothing estimates of $\mu_{r,1}(t), \ldots, \mu_{r,q}(t)$: for $k = 1, \ldots, q$,

$$\widehat{\mu}_{r,k}(t) = \mathbf{e}_1^T \left(\sum_{i=1}^{M_1} \mathbf{H}_i^T \mathbf{W}_{ik} \mathbf{H}_i \right)^{-1} \left(\sum_{i=1}^{M_1} \mathbf{H}_i^T \mathbf{W}_{ik} \widehat{\mathbf{r}}_{i\bullet k} \right), \tag{2}$$

where

$$\mathbf{H}_i = \begin{pmatrix} 1 & (t_{i1} - t) \\ \vdots & \vdots \\ 1 & (t_{im_i} - t) \end{pmatrix},$$

$\mathbf{e}_1 = (1, 0)^T$, $\mathbf{W}_{ik} = \text{diag}\{K_{h_{1k}}(t_{i1} - t), \ldots, K_{h_{1k}}(t_{im_i} - t)\}$, $K_{h_{1k}}(\cdot) = K(\cdot/h_{1k})/h_{1k}$, and $\widehat{\mathbf{r}}_{i\bullet k} = (\widehat{r}_{ik}(t_{i1}), \ldots, \widehat{r}_{ik}(t_{im_i}))^T$. Because the values of $\{\widehat{\mathbf{r}}_i(t_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, M_1\}$ could be serially correlated, the bandwidths selected by the conventional cross-validation (CV) procedure would not perform well (cf., Xie and Qiu (2023)). Instead, the modified CV (MCV) procedure originally suggested by De Brabanter et al. (2011) for univariate nonparametric regression of correlated data is used here for choosing the bandwidths $\{h_{1k}\}$, by which $\{h_{1k}\}$ are chosen by minimizing the following MCV scores:

$$MCV_{\mu_{r,k}}(h_{1k}) = \frac{1}{M_1} \sum_{i=1}^{M_1} \frac{1}{m_i} \sum_{j=1}^{m_i} \left\{ r_{ik}(t_{ij}) - \widehat{\mu}_{r,k}^{(-i,-j)}(t_{ij}) \right\}^2, \;\; k = 1, \ldots, q,$$

where $\widehat{\mu}_{r,k}^{(-i,-j)}(t_{ij})$ is the leave-one-out estimate of $\mu_{r,k}(t_{ij})$ computed by (2) with the observation $\widehat{r}_{ik}(t_{ij})$ being excluded from the computation and the kernel function $K(\cdot)$ being substituted by the following bi-

modal function:

$$K_\epsilon(u) = \frac{4}{4-3\epsilon-\epsilon^3}\begin{cases}\frac{3}{4}(1-u^2)\mathbf{1}(|u|\le 1), & \text{if } |u|\ge\epsilon,\\ \frac{3(1-\epsilon^2)}{4\epsilon}|u|, & \text{otherwise,}\end{cases} \tag{3}$$

where $\epsilon \in (0,1)$ is a pre-specified small constant.

After $\widehat{\boldsymbol{\mu}}_r(t)$ is obtained, we can compute the residuals $\widehat{\varepsilon}_{r,ik}(t_{ij}) = \widehat{r}_{ik}(t_{ij}) - \widehat{\boldsymbol{\mu}}_{r,k}(t_{ij})$, for $i = 1,\ldots,M_1$, $j = 1,\ldots,m_i$ and $k = 1,\ldots,q$. Denote $\widehat{\boldsymbol{\varepsilon}}_{r,i}(t_{ij}) = (\widehat{\varepsilon}_{r,i1}(t_{ij}),\ldots,\widehat{\varepsilon}_{r,iq}(t_{ij}))^T$. Then, the estimate of the covariance matrix $\mathbf{V}_r(s,t)$ can be obtained as follows. For $s \neq t$ or $k \neq k'$ with $k,k' = 1,\ldots,q$, the (k,k')-th entry of $\mathbf{V}_r(s,t)$ is estimated by

$$\widehat{\text{cov}}(\varepsilon_{r,ik}(s),\varepsilon_{r,ik'}(t)) = \frac{\sum_{i=1}^{M_1}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_i}\widehat{\varepsilon}_{r,ik}(t_{ij})\widehat{\varepsilon}_{r,ik'}(t_{ij'})K\left(\frac{t_{ij}-s}{h_{2k}}\right)K\left(\frac{t_{ij'}-t}{h_{2k'}}\right)}{\sum_{i=1}^{M_1}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_i}K\left(\frac{t_{ij}-s}{h_{2k}}\right)K\left(\frac{t_{ij'}-t}{h_{2k'}}\right)},$$

where $\{h_{2k}\}$ are bandwidths. When $s = t$ and $k = k'$, the kth diagonal term of $\mathbf{V}_r(s,t)$ is estimated by

$$\widehat{\text{var}}(\varepsilon_{r,ik}(t)) = \frac{\sum_{i=1}^{M_1}\sum_{j=1}^{m_i}\{\widehat{\varepsilon}_{r,ik}(t_{ij})\}^2K\left(\frac{t_{ij}-t}{h_{2k}}\right)}{\sum_{i=1}^{M_1}\sum_{j=1}^{m_i}K\left(\frac{t_{ij}-t}{h_{2k}}\right)}. \tag{4}$$

To choose the bandwidths $\{h_{2k}\}$, an MCV procedure similar to the one discussed above can be developed. Specifically, $\{h_{2k}\}$ can be chosen by minimizing the following MCV scores:

$$MCV_{\sigma_{r,k}}(h_{2k}) = \frac{1}{M}\sum_{i=1}^{M_1}\frac{1}{m_i}\sum_{j=1}^{m_i}\left[\{\widehat{\varepsilon}_{r,ik}(t_{ij})\}^2 - \widehat{\sigma}_{r,ik}^{(-i,-j)}(t_{ij})\right]^2,\ \ k = 1,\ldots,q,$$

where $\widehat{\sigma}_{r,ik}^{(-i,-j)}(t_{ij})$ is the leave-one-out estimate of var$(\varepsilon_{r,ik}(t_{ij}))$ computed by (4) with the observation $\widehat{\varepsilon}_{r,ik}(t_{ij})$ being excluded from the computation and the bimodal kernel function (3) used to accommodate serial data correlation. In all numerical studies in Section 4, $K(\cdot)$ is chosen to be the Epanechnikov kernel function, and ϵ (cf., (3)) is chosen to be 0.1 (suggested by De Brabanter et al. (2011)) when computing the MCV scores.

2.3 Online disease risk monitoring for disease early detection

To jointly monitor multiple diseases in concern for a new patient, we can sequentially monitor the patient's longitudinal pattern of the estimated disease risks computed from the observed disease risk factors. Let the new patient's estimated risks be $\widehat{\mathbf{r}}(t_j^*) = (\widehat{\psi}_1(\widehat{\boldsymbol{\beta}}_1^T\mathbf{X}(t_j^*)),\ldots,\widehat{\psi}_q(\widehat{\boldsymbol{\beta}}_q^T\mathbf{X}(t_j^*)))^T$, where $\{t_j^*, j \ge 1\}$ are observation times, $\mathbf{X}$ is the p-dimensional vector of disease risk factors, and $\{\widehat{\boldsymbol{\beta}}_k, k = 1,\ldots,q\}$ and $\{\widehat{\psi}_k, k =$

$1, \ldots, q\}$ are obtained from the training data by the procedure discussed in Section 2.1. It should be pointed out that the values of $\{\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}(t_j^*)\}$ could be beyond the range of $\{\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij}, i = 1, \ldots, M, j = 1, \ldots, m_i\}$, denoted as $[u_{min}, u_{max}]$, obtained in the training data. In such cases, the new patient's estimated risks may not be well defined since $\{\widehat{\psi}_k(\cdot)\}$ are defined in subsets of $[u_{min}, u_{max}]$ only. To solve this boundary problem, we suggest fitting a linear model for the data points $\{(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij}, r_{ik}(t_{ij})) : \widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} \in [u_{min}, u_{min} + h_{1k}]\}$ and $\{(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij}, r_{ik}(t_{ij})) : \widehat{\boldsymbol{\beta}}_k^T \mathbf{X}_{ij} \in [u_{max} - h_{1k}, u_{max}]\}$, respectively, and then obtain the values of $\widehat{\psi}_k(\widehat{\boldsymbol{\beta}}_k^T \mathbf{X}(t_j^*))$ by extrapolation. Our numerical experience shows that the above boundary problem is quite rare in practice, especially when the training dataset is sufficiently large. Additionally, alternative extrapolation strategies – such as fitting a quadratic curve to predict out-of-bound risk values – would minimally change the subsequent results.

To sequentially monitor the new patient's estimated risks $\{\widehat{\mathbf{r}}(t_j^*), j \geq 1\}$, a multivariate SPC chart is natural to consider. But, conventional control charts for sequential process monitoring are designed for cases when IC process observations are i.i.d. (cf., Qiu (2014)), which would not be valid here for the estimated disease risks. To address this issue, Li and Qiu (2017) proposed a sequential data decorrelation and standardization procedure for multivariate process observations. Next, we describe a modified version of this procedure, and will explain in Appendix B that this modified version is equivalent to the original procedure but has a more efficient computation.

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

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

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

  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 observation at time t_j^* is defined to be

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

  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 since $\widehat{\mathbf{D}}_j$ may not be a positive semi-definite matrix. In the case when $\widehat{\mathbf{D}}_j$ is not positive semi-definite, we suggest modifying it to be a positive semi-definite matrix using the algorithm proposed by Higham (1988).

After the above data standardization and decorrelation, the transformed disease risks $\{\widehat{\mathbf{e}}(t_j^*), j \geq 1\}$ would be asymptotically uncorrelated with each other and each has the asymptotic mean $\mathbf{0}$ and asymptotic identity covariance matrix. To online monitor these data, we suggest using a multivariate control chart modified from the one used in Qiu et al. (2018), which can accounts for possible unequally-spaced observation times in the observed data and provide an effective detection of upward mean shifts in the transformed multivariate risks. To proceed, let $\omega > 0$ be a basic time unit that all observation times are its integer multiples. Then, we have $\{t_j^* = n_j^*\omega, \text{ for } j \geq 1\}$, where n_j^* is a positive integer. At the current time point t_j^*, the proposed charting statistic is defined to be $\mathbf{E}_j^{+T}\mathbf{E}_j^+$, where $\mathbf{E}_j^+ = (E_{j1}^+, \ldots, E_{jq}^+)^T$,

$$E_{jk}^+ = \max\left[0, \{1 - \Lambda(t_j^*)\}E_{j-1,k}^+ + \Lambda(t_j^*)\widehat{e}_k(t_j^*)\right], \text{ for } j \geq 1, \ k = 1, \ldots, q, \tag{5}$$

$\mathbf{E}_0^+ = 0$, $\widehat{e}_k(t_j^*)$ is the kth 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 gives a signal of upward-shift in the transformed multivariate risks if

$$\mathbf{E}_j^{+T}\mathbf{E}_j^+ > \rho, \tag{6}$$

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

From (5), it can be seen that λ and $\{\Delta_j, j \geq 1\}$ control the weights assigned to the historical data when computing the the charting statistic values. The weights would be exponentially decay for older data and the unequally-spaced observation times have been accommodated by using $\{\Delta_j, j \geq 1\}$ in the weights. In addition, the proposed charting statistic $\mathbf{E}_j^{+T}\mathbf{E}_j^+$ is a composite of multiple one-sided EWMA charting statistics $\{E_{jk}^+, k = 1, \ldots, q\}$. Because we only care about upward shifts in all dimensions of disease risks, consideration of the one-sided EWMA charting statistics can rule out the possibility of a signal triggered by downward shifts in some dimensions of the disease risks.

In the SPC literature, performance of a control chart is usually measured by the IC average run length (ARL_0), defined to be the average number of observation times from the start of process monitoring to the signal time, and OC ARL (ARL_1), defined to be the average number of observation times from the occurrence of a shift to the signal time (Qiu, 2014). When the observation times of each subject are unequally-

spaced, however, the ARL values cannot provide a meaningful performance measure. To address this issue, Qiu and Xiang (2014) suggested using the average time to signal (ATS) in the basic time unit ω defined earlier to evaluate the performance of a control chart in such cases. Specifically, the IC ATS (ATS_0) is defined as the average time from the beginning of process monitoring to the signal time of the chart, and the OC ATS (ATS_1) is defined as the average time from the occurrence of a shift to the signal time. Usually, the ATS_0 is pre-specified at a given level, and a chart performs better in detecting a given shift if its ATS_1 is smaller.

In the proposed control chart (5)-(6), the smoothing parameter λ is usually pre-specified and the control limit ρ needs to be chosen so that a desired value of ATS_0 is achieved. Because the transformed disease risks $\{\widehat{\mathbf{e}}(t_1^*), \ldots, \widehat{\mathbf{e}}(t_j^*)\}$ may not be normally distributed, the bisection search procedures using Monte Carlo simulations as discussed in Qiu and Xiang (2014) and Li and Qiu (2017) cannot be used here to choose ρ. To overcome this difficulty, we recommend using the block bootstrap procedure described in Qiu and Xiang (2014). To this end, the training data with $2M$ subjects is first divided into two halves, with all diseased and non-diseased people splitted evenly. Then, the first half (with M subjects) of the training data is used to estimate Model (1) and the regular longitudinal pattern of the estimated multivariate disease risks as discussed in Sections 2.1 and 2.2, and the second half is used for determining ρ.

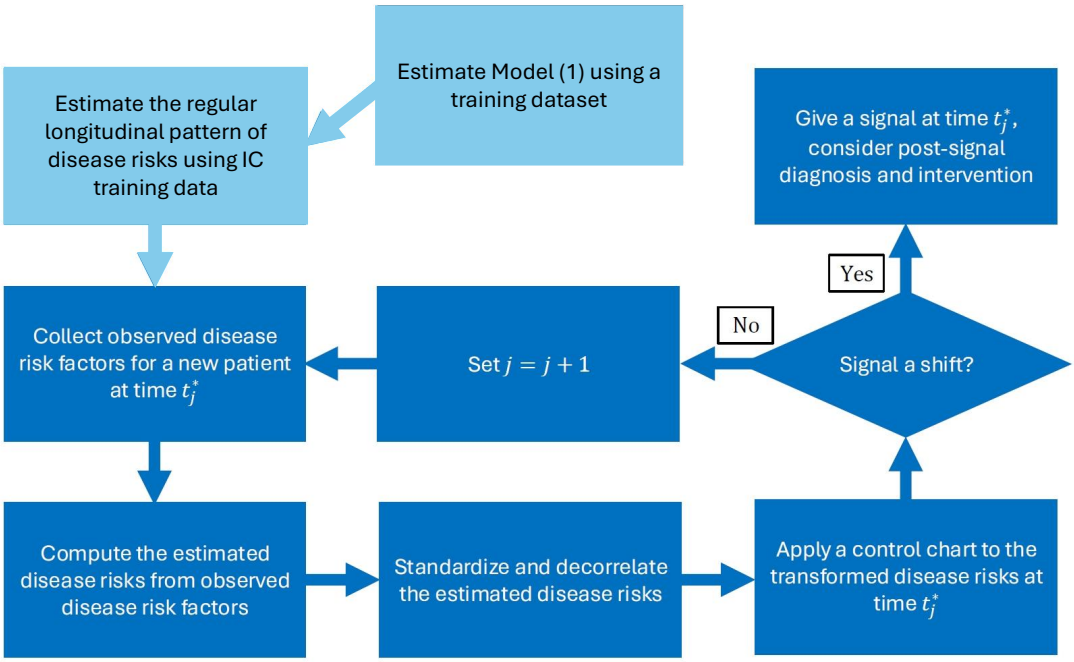
To conclude this section, some key components of the proposed DySS method are summarized in the diagram shown by Figure 1. The light blue boxes summarize the steps performed with the collected training dataset, while the dark blue boxes illustrate the online monitoring procedure for new patients, in which new observations of risk factors are continuously collected over time.

3 Simulation Studies

In this section, we present some simulation results about the numerical performance of the proposed method. The simulation setups are described below. First, observation times of all subjects are within the study time period $[0, 1]$ with the basic time unit of $\omega = 0.01$. For the ith subject, the number of observation times m_i is generated independently by rounding a random number from $N(\overline{m}, \overline{m}/30)$ to the nearest integer, where $\overline{m}$ is a pre-specified integer. That subject's jth observation time t_{ij} is defined to be $n_{ij}\omega$, where n_{ij} is the ceiling of a random number generated from Unif$(100(j-1)/m_i, 100j/m_i)$.

Second, it is assumed that there are $p = 4$ risk factors $\mathbf{X}_i(t) = (X_{i1}(t), X_{i2}(t), X_{i3}(t), X_{i4}(t))^T$

Figure 1: Diagram of the key components of the proposed DySS method for detecting multiple diseases.



with the IC mean function $\boldsymbol{\mu}_X(t) = (t, 1 + 0.2t + 0.3t^2, \cos(2\pi t)/2, 0)^T$, and the error term $\boldsymbol{\varepsilon}_X(t) = \text{diag}\{\exp(t)/2, 1/(1+t), \log(t+5)/2, 1\}\boldsymbol{\varepsilon}_X^*$, where $\boldsymbol{\varepsilon}_X^*$ is generated from the following vector AR(1) model $\boldsymbol{\varepsilon}_X^*(t) = \alpha\boldsymbol{\varepsilon}_X^*(t-\omega) + \mathbf{e}_X(t)$, and $\mathbf{e}_X(t)$ is a white noise process with mean 0 and covariance matrix

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \tau & \tau^2 & \tau^3 \\ \tau & 1 & \tau & \tau^2 \\ \tau^2 & \tau & 1 & \tau \\ \tau^3 & \tau^2 & \tau & 1 \end{pmatrix}.$$

In the simulation studies, we set $\alpha = 0.2$ or 0.8 and $\tau = 0.2$ or 0.8, to simulate small or large within-subject and between-variable variability, respectively. For the white noise process $\mathbf{e}_X(t)$, the distribution of each of its element is assumed to be either the standard normal distribution or a standardized version of the χ_3^2 distribution.

Third, it is assumed that there are $q = 2$ response variables, generated from the following IC multivariate single-index logistic regression model:

$$\begin{cases} \text{logit}\{\mathbb{E}(Y_{ij1}|\mathbf{X}_{ij}, \mathbf{b}_i)\} = \psi_1\{\boldsymbol{\beta}_1^T\mathbf{X}_i(t_{ij})\} + b_{i1} \\ \text{logit}\{\mathbb{E}(Y_{ij2}|\mathbf{X}_{ij}, \mathbf{b}_i)\} = \psi_2\{\boldsymbol{\beta}_2^T\mathbf{X}_i(t_{ij})\} + b_{i2}, \end{cases} \tag{7}$$

where the random-effects $\mathbf{b}_i = (b_{i1}, b_{i2})^T$ are generated from a zero-mean bivariate normal distribution with the covariance matrix

$$\mathbf{\Sigma}_b = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

In the single-index model (7), it is assumed that ψ_1 is an identity link function, and $\psi_2(u) = \exp(u)/2 - 3$. The true index coefficients are $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})^T = (1, -3, 0, 3)^T/\sqrt{19}$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24})^T = (2, -2, 3, 0)^T/\sqrt{17}$.

Fourth, in each simulation study, we first generate a training dataset with $2M$ subjects, of which two-thirds are generated from the IC model described above and the remaining one-third are generated from an OC model which is the same as the IC model except that the risk factors have the mean $\boldsymbol{\mu}_X(t) + (0.5, -0.5, 0.5, 0.5)^T\delta^*$, where $\boldsymbol{\mu}_X(t)$ is the IC mean function and δ^* is a random number generated from Unif$[0, 3]$. The training data are then divided into two halves with the IC and OC subjects divided evenly. The first half with M subjects is used to estimate Model (7) and the regular longitudinal pattern of the estimated risks, as discussed in Section 2.2. Observed data of $M_1 = 2M/3$ well-functioning (IC) subjects in the second half are used to determine the control limit ρ of the control chart (5)-(6), as discussed in Section 2.3. Then, the conditional ATS_0 or ATS_1 value given the training data is computed based on $1,000$ simulations of online process monitoring. To compute the (unconditional) actual ATS_0 or ATS_1 value, all above steps, from generation of the training data to computation of the conditional ATS_0 or ATS_1 value, are repeated for 100 times. The average of the 100 conditional ATS_0 or ATS_1 values is used as the final estimate of the actual ATS_0 or ATS_1 value. The standard error of the final estimate can also be computed. In all simulation studies, the nominal ATS_0 value is set to be 25.

In the subsequent sections, the numerical performance of the estimation procedure for Model (1) and the proposed control chart in (5)–(6) is evaluated. In addition to the proposed method, labeled PROPOSED, five alternative methods are considered for comparison. The first alternative method is the Multivariate Dynamic Screening System (MDySS) suggested by Li and Qiu (2017), which monitors the observed disease risk factors directly by a multivariate EWMA chart after serial data correlation is eliminated by a data decorrelation procedure. The second method, denoted UGSIM, estimates the risk for each disease following You and Qiu (2020) and then monitors the longitudinal disease-specific risk trajectories using multiple univariate EWMA charts. The third method, denoted MGSIM, applies a conventional multivariate EWMA chart with a constant weighting parameter to the decorrelated disease risks computed as described in Section 2. The fourth method, denoted PROPOSED-S, naively assumes that the conditional log-odds for each

disease is a linear function of the risk factors. It employs a generalized linear mixed model (GLMM) to estimate Model (1) and the corresponding disease risks, while all other components of the procedure remain the same as in PROPOSED. The control limits of control charts in the above four methods are obtained by the block bootstrap procedure discussed at the end of Section 2. The final alternative method, denoted GPBOOST, uses gradient boosting (Sigrist, 2022) to estimate nonlinear relationships between the risk factors and the conditional log-odds for each disease, while accounting for within-subject correlation through random intercepts. The fitted models are then used to compute estimated disease risks for new subjects. To ensure comparability with the other methods, pointwise confidence intervals are constructed for the risk of each disease over time based on well-functioning subjects in the training data, and a signal is triggered for a new subject if any component of the estimated disease risk exceeds the bounds of the $100 \times (1-\zeta)\%$ confidence interval, where $\zeta \in (0,1)$ is a prespecified constant.

3.1 Estimation performance of the longitudinal model

We first present some results about the estimation of Model (1). Table 1 presents the sum of squares of estimation error (SSE) for the index coefficient vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in various cases when the white noise process $\mathbf{e}_X(t)$ follows a normal distribution, together with the related standard errors (in parentheses), where SSE for $\boldsymbol{\beta}_k$ is defined to be $\sum_{l=1}^{p}(\widehat{\beta}_{kl}-\beta_{kl})^2$, for $k=1,2$. From the table, it can be seen that SSE and its standard error decrease as M and/or $\overline{m}$ increase, indicating the estimation becomes more accurate as the sample size increases. Corresponding results when the distribution of the white noise process $\mathbf{e}_X(t)$ is chi-square are presented in Table A.1 in Appendix C. Similar conclusions can be made from that table.

To further evaluate the prediction performance of the proposed multivariate single-index logistic regression model, we compare it with the GPBOOST method introduced earlier, which is a representative machine-learning approach for nonlinear disease-risk estimation with correlated longitudinal data. In the current problem, because GPBOOST is designed for univariate outcomes, separate models are fitted by this approach for different diseases. After model fitting, estimated disease risks are obtained from the fitted models and evaluated on independently generated test datasets with the same sample size as the training data. The corresponding mean area under the receiver operating characteristic curve (AUC) values and their standard errors (in parentheses), computed from 100 simulation replications under different settings, are presented in Tables 2 and 3 for outcomes Y_1 and Y_2, respectively, when the white noise process $\mathbf{e}_X(t)$ follows a normal distribution.

From the tables, it can be seen that the proposed method consistently achieves larger AUC values than GPBOOST in almost all cases considered. In addition, the standard errors of the estimated AUC values generally decrease as M and/or $\overline{m}$ increase, indicating improved estimation stability with larger sample sizes. While the mean AUC values of the proposed method remain relatively stable across different sample sizes, GPBOOST tends to achieve improved prediction accuracy as the sample size increases. Corresponding results when the white noise process $\mathbf{e}_X(t)$ follows a chi-square distribution show similar patterns and are therefore omitted. One possible explanation for the inferior performance of GPBOOST in this example is that the method is fully nonparametric and models each disease separately, without explicitly incorporating the correlation structure among multiple diseases. As a result, its prediction efficiency may be reduced in the current setting.

Table 1: Sum of squares of estimation error (SSE) of the estimated index coefficients of Model (1) in various cases when the white noise process $\mathbf{e}_X(t)$ has a normal distribution. Numbers in parentheses are standard errors in the unit of 10^{-3}.

		$\alpha=0.2,\tau=0.2$		$\alpha=0.2,\tau=0.8$		$\alpha=0.8,\tau=0.2$		$\alpha=0.8,\tau=0.8$	
$\overline{m}$	M	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
5	30	0.164 (17.083)	0.551 (94.016)	0.263 (26.208)	1.084 (113.102)	0.193 (40.380)	0.103 (6.822)	0.177 (14.740)	0.312 (55.033)
	75	0.083 (7.068)	0.119 (10.814)	0.091 (9.352)	0.272 (26.089)	0.045 (3.167)	0.049 (4.408)	0.073 (7.297)	0.139 (18.039)
	150	0.030 (2.837)	0.071 (5.049)	0.052 (6.554)	0.171 (12.700)	0.015 (1.502)	0.021 (1.081)	0.033 (2.346)	0.071 (8.915)
	300	0.011 (0.988)	0.034 (1.901)	0.038 (4.802)	0.072 (4.463)	0.009 (0.829)	0.010 (0.895)	0.012 (1.716)	0.026 (1.315)
	450	0.015 (1.294)	0.024 (1.630)	0.019 (1.308)	0.060 (9.308)	0.006 (0.347)	0.009 (0.425)	0.009 (0.930)	0.023 (2.290)
10	30	0.100 (6.208)	0.114 (9.855)	0.209 (18.732)	0.235 (20.592)	0.050 (3.958)	0.067 (4.135)	0.091 (8.007)	0.085 (4.909)
	75	0.034 (3.227)	0.056 (4.793)	0.129 (12.258)	0.115 (13.426)	0.012 (0.869)	0.017 (1.644)	0.038 (3.321)	0.037 (2.646)
	150	0.011 (0.593)	0.022 (1.367)	0.015 (1.409)	0.036 (2.902)	0.006 (0.504)	0.011 (0.663)	0.013 (1.076)	0.022 (1.690)
	300	0.006 (0.732)	0.017 (1.835)	0.013 (1.060)	0.026 (2.981)	0.003 (0.157)	0.005 (0.665)	0.004 (0.543)	0.015 (1.616)
	450	0.008 (0.510)	0.014 (1.176)	0.016 (1.283)	0.022 (1.617)	0.002 (0.175)	0.004 (0.397)	0.005 (0.561)	0.008 (0.554)
15	30	0.058 (7.932)	0.073 (8.077)	0.119 (11.950)	0.204 (20.636)	0.048 (5.770)	0.052 (5.655)	0.107 (8.378)	0.074 (4.377)
	75	0.030 (2.010)	0.035 (3.109)	0.034 (3.344)	0.078 (8.485)	0.008 (0.724)	0.010 (0.619)	0.018 (1.788)	0.026 (2.986)
	150	0.014 (0.671)	0.018 (2.045)	0.033 (2.435)	0.029 (3.022)	0.005 (0.352)	0.004 (0.261)	0.013 (0.783)	0.016 (0.999)
	300	0.005 (0.562)	0.015 (1.634)	0.015 (1.770)	0.030 (4.251)	0.003 (0.246)	0.003 (0.161)	0.009 (1.330)	0.005 (0.504)
	450	0.003 (0.207)	0.008 (0.766)	0.009 (0.806)	0.015 (1.072)	0.002 (0.149)	0.003 (0.181)	0.006 (0.619)	0.005 (0.635)

Table 2: AUC of the disease risks estimated by the proposed multivariate single-index logistic regression model and the GPBOOST method for outcome Y_1 in various cases when the white noise process $\mathbf{e}_X(t)$ has a normal distribution. Numbers in parentheses are standard errors in the unit of 10^{-3}.

		$\alpha = 0.2, \tau = 0.2$		$\alpha = 0.2, \tau = 0.8$		$\alpha = 0.8, \tau = 0.2$		$\alpha = 0.8, \tau = 0.8$	
$\overline{m}$	M	PROPOSED	GPBOOST	PROPOSED	GPBOOST	PROPOSED	GPBOOST	PROPOSED	GPBOOST
5	30	0.710 (4.462)	0.618 (5.743)	0.678 (5.370)	0.568 (4.373)	0.778 (4.033)	0.711 (4.910)	0.722 (4.389)	0.621 (5.788)
	75	0.717 (3.106)	0.637 (2.946)	0.681 (2.991)	0.580 (3.720)	0.779 (2.347)	0.732 (2.932)	0.729 (2.971)	0.655 (3.039)
	150	0.717 (2.012)	0.650 (2.349)	0.679 (2.164)	0.592 (2.415)	0.783 (1.750)	0.741 (1.736)	0.729 (2.081)	0.668 (2.312)
	300	0.719 (1.509)	0.669 (1.539)	0.685 (1.864)	0.610 (1.784)	0.782 (1.260)	0.753 (1.358)	0.731 (1.639)	0.683 (1.565)
	450	0.717 (1.179)	0.671 (1.305)	0.682 (1.312)	0.615 (1.331)	0.782 (1.009)	0.756 (1.014)	0.729 (1.274)	0.684 (1.236)
10	30	0.704 (3.513)	0.637 (3.401)	0.663 (3.976)	0.574 (3.686)	0.774 (2.997)	0.728 (3.212)	0.719 (3.677)	0.650 (3.806)
	75	0.716 (2.346)	0.657 (1.854)	0.679 (2.368)	0.596 (1.881)	0.779 (1.776)	0.744 (1.908)	0.726 (1.928)	0.669 (1.841)
	150	0.716 (1.369)	0.667 (1.339)	0.681 (1.740)	0.607 (1.698)	0.782 (1.454)	0.751 (1.404)	0.728 (1.595)	0.679 (1.395)
	300	0.718 (0.993)	0.673 (1.032)	0.684 (1.072)	0.616 (0.986)	0.784 (0.868)	0.755 (0.999)	0.732 (1.069)	0.686 (1.096)
	450	0.716 (1.011)	0.675 (0.866)	0.683 (1.025)	0.621 (0.851)	0.783 (0.792)	0.759 (0.761)	0.730 (0.992)	0.690 (0.854)
15	30	0.715 (3.048)	0.648 (2.802)	0.676 (3.415)	0.585 (2.798)	0.780 (2.420)	0.734 (2.603)	0.726 (2.803)	0.657 (2.741)
	75	0.718 (1.819)	0.664 (1.795)	0.684 (2.174)	0.606 (1.700)	0.784 (1.451)	0.750 (1.594)	0.731 (1.903)	0.678 (1.806)
	150	0.719 (1.266)	0.673 (1.288)	0.682 (1.459)	0.615 (1.196)	0.782 (0.908)	0.756 (1.062)	0.729 (1.209)	0.685 (1.282)
	300	0.716 (0.960)	0.674 (0.862)	0.682 (1.227)	0.621 (0.793)	0.781 (0.807)	0.757 (0.808)	0.729 (0.945)	0.690 (0.842)
	450	0.717 (0.827)	0.676 (0.728)	0.683 (1.045)	0.624 (0.727)	0.782 (0.665)	0.760 (0.594)	0.730 (0.779)	0.690 (0.713)

Table 3: AUC of the disease risks estimated by the proposed multivariate single-index logistic regression model and the GPBOOST method for outcome Y_2 in various cases when the white noise process $\mathbf{e}_X(t)$ has a normal distribution. Numbers in parentheses are standard errors in the unit of 10^{-3}.

		$\alpha = 0.2, \tau = 0.2$		$\alpha = 0.2, \tau = 0.8$		$\alpha = 0.8, \tau = 0.2$		$\alpha = 0.8, \tau = 0.8$	
$\overline{m}$	M	PROPOSED	GPBOOST	PROPOSED	GPBOOST	PROPOSED	GPBOOST	PROPOSED	GPBOOST
	30	0.661 (7.219)	0.565 (5.138)	0.645 (8.022)	0.553 (5.277)	0.721 (5.914)	0.586 (6.626)	0.708 (6.555)	0.602 (7.109)
	75	0.667 (3.591)	0.549 (3.764)	0.659 (3.859)	0.546 (4.056)	0.723 (2.998)	0.614 (4.909)	0.710 (3.059)	0.600 (5.000)
5	150	0.671 (3.020)	0.555 (3.278)	0.664 (3.321)	0.544 (3.114)	0.726 (2.699)	0.629 (3.208)	0.710 (2.842)	0.615 (3.450)
	300	0.665 (1.923)	0.566 (2.448)	0.658 (1.916)	0.562 (2.215)	0.726 (1.687)	0.647 (2.222)	0.712 (1.725)	0.631 (2.192)
	450	0.665 (1.982)	0.569 (1.784)	0.662 (1.958)	0.567 (1.970)	0.725 (1.582)	0.650 (1.551)	0.711 (1.751)	0.635 (1.642)
	30	0.650 (5.181)	0.546 (4.293)	0.642 (5.464)	0.543 (4.015)	0.714 (4.332)	0.609 (4.613)	0.696 (4.730)	0.604 (4.957)
	75	0.656 (3.159)	0.555 (3.449)	0.648 (3.057)	0.557 (3.197)	0.718 (2.701)	0.630 (2.839)	0.703 (2.888)	0.611 (3.079)
10	150	0.663 (2.368)	0.561 (2.013)	0.654 (2.408)	0.558 (2.299)	0.724 (1.905)	0.643 (1.946)	0.709 (2.041)	0.629 (2.095)
	300	0.663 (1.561)	0.574 (1.665)	0.657 (1.786)	0.571 (1.618)	0.722 (1.318)	0.652 (1.498)	0.707 (1.526)	0.640 (1.683)
	450	0.663 (1.225)	0.581 (1.239)	0.657 (1.337)	0.576 (1.150)	0.725 (1.094)	0.656 (1.244)	0.710 (1.177)	0.645 (1.211)
	30	0.657 (4.769)	0.556 (3.695)	0.648 (5.010)	0.545 (3.641)	0.723 (3.688)	0.618 (4.466)	0.705 (3.719)	0.602 (4.314)
	75	0.664 (2.576)	0.561 (2.218)	0.659 (2.672)	0.553 (2.922)	0.726 (2.020)	0.638 (2.369)	0.712 (2.127)	0.625 (2.724)
15	150	0.665 (1.897)	0.573 (1.845)	0.658 (1.963)	0.566 (1.997)	0.723 (1.300)	0.648 (1.809)	0.709 (1.400)	0.633 (1.851)
	300	0.662 (1.203)	0.581 (1.244)	0.655 (1.285)	0.576 (1.461)	0.723 (0.991)	0.652 (1.182)	0.708 (0.996)	0.644 (1.374)
	450	0.662 (1.076)	0.584 (0.936)	0.654 (1.090)	0.579 (1.011)	0.723 (0.906)	0.655 (1.078)	0.708 (1.068)	0.644 (1.051)

3.2 IC performance of the proposed method

In cases when $\alpha = \tau = 0.2$ and the white noise process $\mathbf{e}_X(t)$ follows a normal distribution, the actual ATS_0 values of different methods are presented in Figure 2 in various cases considered. In the figure, MDySS0 denotes the ideal version of MDySS in which the IC process observations are assumed to be i.i.d. and have a multivariate standard normal distribution and its control limit is obtained by Monte Carlo simulations based on these assumptions (cf., Li and Qiu (2017)). For GPBOOST, we choose $\zeta = 0.08$ so that the actual ATS_0 values for this method under a relatively large M_1 and $\overline{m} = 5$ is close to the nominal level 25. From the figure, it can be seen that the actual ATS_0 values of the charts MDySS, UGSIM, MGSIM PROPOSED-S and PROPOSED all stay within 10% of the nominal ATS_0 value of 25 when $M_1 > 50$. As M_1 getting larger, their actual ATS_0 values get closer to the nominal ATS_0 value. These results confirm that the block bootstrap procedure discussed at the end of Section 2 performs well in choosing the control limits of the related control charts. As a comparison, the IC performance of MDySS0 is not good in some cases considered because its i.i.d. assumptions are violated. The IC performance of GPBOOST is also unstable: For $\overline{m} = 5$ with $M_1 < 200$, or when $\overline{m} = 10$ or 15, the computed ATS_0 can deviate substantially from its nominal level. Similar conclusions can be made from Figure 3 that presents the actual ATS_0 values of different methods in cases when the white noise process $\mathbf{e}_X(t)$ has a standardized chi-square distribution. Results from these two figures confirm that the block bootstrap procedure can effectively guarantee the IC performance of the related control charts given a sufficiently large M_1. As a comparison, the average number of observation times, $\overline{m}$, for individual subjects seems to have a small impact on the IC performance of different methods. The poor IC performance of GPBOOST also indicates that prespecified threshold values cannot reliably control false-alarm rates in sequential decision-making settings. Figures A.2-A.6 in Appendix C provide more simulation results about the actual ATS_0 values of different methods when α and τ take other values. Similar conclusions can be made there.

3.3 OC performance of the proposed method

Next, we evaluate the OC performance of different control charts considered when $M = 450$ (i.e., $M_1 = 300$) and $\overline{m} = 15$. In practice, different OC subjects may have different irregular longitudinal patterns of risk factors. The example here is designed for simulating such cases. Specifically, it is assumed that $100\gamma\%$ OC subjects have the risk factors defined by

$$\mathbf{X}^*(t) = \mathbf{X}(t) + \boldsymbol{\beta}_1\delta,$$

Figure 2: Actual ATS_0 values of the charts MDySS0, MDySS, UGSIM, MGSIM, GPBOOST, PROPOSED-S and PROPOSED in various cases when their nominal ATS_0 values are fixed at 25, $\alpha = \tau = 0.2$, and the white noise process $\mathbf{e}_X(t)$ has a normal distribution. The two gray dashed lines in each plot indicate the range within 5% of the nominal ATS_0 value.

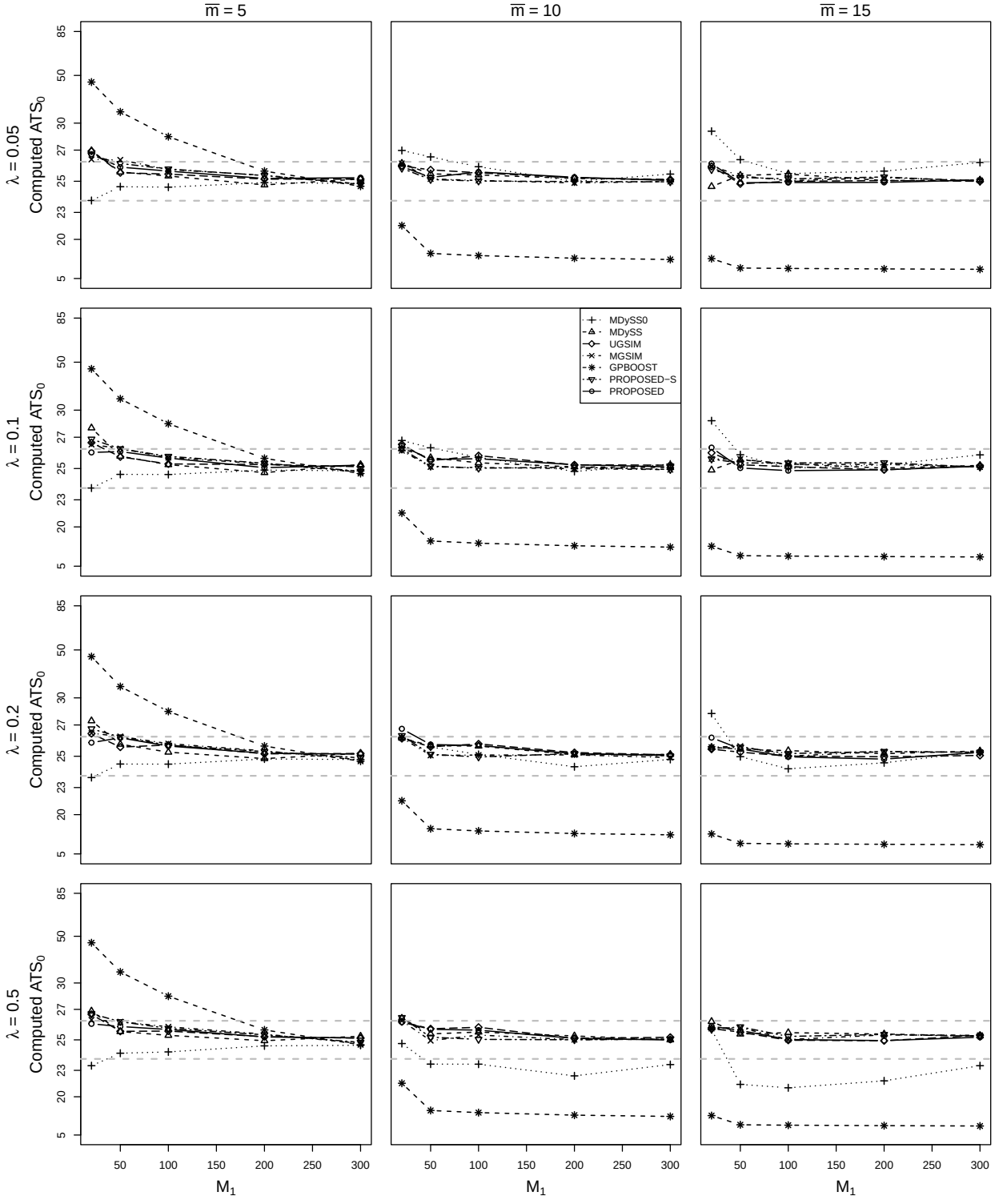
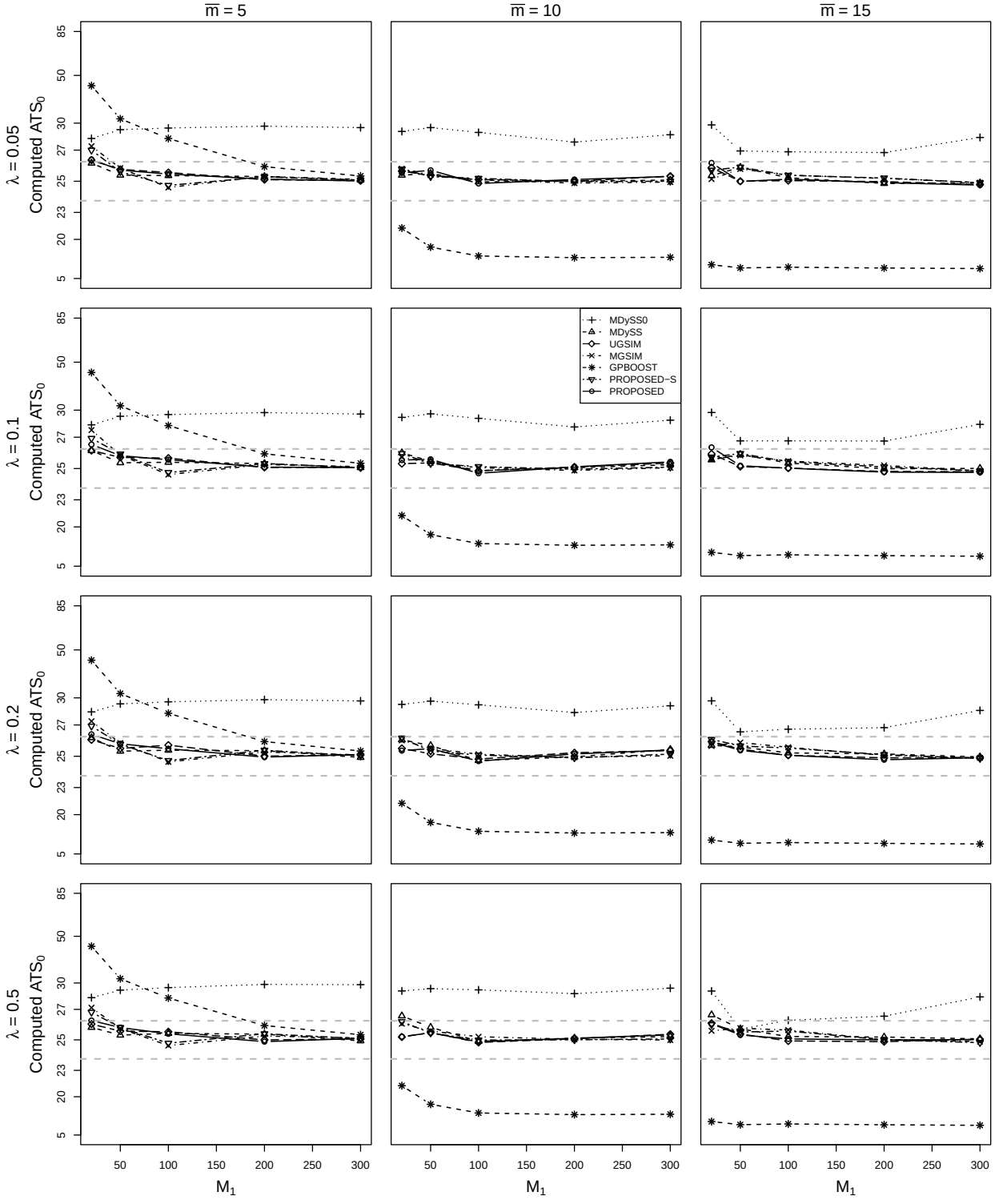


Figure 3: Actual ATS_0 values of the charts MDySS0, MDySS, UGSIM, MGSIM, GPBOOST, PROPOSED-S and PROPOSED in various cases when their nominal ATS_0 values are fixed at 25, $\alpha = \tau = 0.2$, and the white noise process $\mathbf{e}_X(t)$ has a standardized chi-square distribution. The two gray dashed lines in each plot indicate the range within 5% of the nominal ATS_0 value.



and the remaining $100(1-\gamma)\%$ OC subjects have the risk factors defined by

$$\mathbf{X}^*(t) = \mathbf{X}(t) + \boldsymbol{\beta}_2\delta,$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are defined as before, and δ controls the shift size and changes between 0 and 3. All other setups are the same as those in Section 3.2.

To have a fair comparison among different methods, their optimal ATS_1 values are considered here, where the optimal ATS_1 value of a chart is defined to be the minimal ATS_1 value for detecting a given shift when the charting parameter (e.g., λ in the chart (5)-(6)) varies and the nominal ATS_0 level is maintained. Figure 4 presents the optimal ATS_1 values of different methods in various cases when $\gamma \in \{1, 0.75, 0.5, 0.25, 0\}$ and the white noise process $\mathbf{e}_X(t)$ follows a normal distribution. From the figure, we have the following observations: i) the proposed method PROPOSED outperforms the other five methods in all cases considered, ii) the two simplified versions UGSIM and MGSIM perform similarly and both outperform MDySS, iii) PROPOSED-S performs slightly worse than PROPOSED in all settings, highlighting the negative impact of model misspecification on monitoring performance, iv) GPBOOST outperforms UGSIM and MGSIM when $\gamma < 0.5$, v) smaller within-subject correlation (i.e., α is smaller) or higher between-variable correlation (i.e., τ is larger) would lead to better OC performance of all six methods, and vi) optimal ATS_1 value of each chart decreases as γ decreases from 1 to 0. The last conclusion implies that mean shifts in the direction of $\boldsymbol{\beta}_2$ are easier to detect than those in the direction of $\boldsymbol{\beta}_1$. Figure A.7 in Appendix C shows the optimal ATS_1 values of the six charts when the white noise process $\mathbf{e}_X(t)$ has a standardized chi-square distribution. Similar conclusions can be made there.

While the above scenarios consider only abrupt mean shifts in the risk factors and the corresponding disease risks, gradual changes or drifts in risk factors are also common in disease progression. In the following example, we investigate the performance of the competing methods when OC subjects exhibit drifting risk factors over time. Specifically, the observed OC risk factors are assumed to follow

$$\mathbf{X}^*(t) = \mathbf{X}(t) + (0.5, -0.5, 0.5, 0.5)^T \phi t \times 100,$$

where ϕ controls the slope of the drift and change between 0 and 0.09. Figure 5 and Figure A.8 in Appendix C present the optimal ATS_1 values of different methods in various cases over different values of ϕ when the white noise process $\mathbf{e}_X(t)$ follows either a normal distribution or a standardized version of the χ^2_3 distribution. The proposed method outperforms the other five methods as shown in the abrupt mean shift scenarios.

Figure 4: Optimal ATS_1 values of the charts MDySS, UGSIM, MGSIM, GPBOOST, PROPOSED-S and PROPOSED in various cases when the nominal ATS_0 value is 25, the white noise process $\mathbf{e}_X(t)$ has a normal distribution, the observed risk factors of $100\gamma\%$ OC subjects have mean shifts of sizes $\boldsymbol{\beta}_1\delta$, and the observed risk factors of remaining OC subjects have mean shifts of sizes $\boldsymbol{\beta}_2\delta$, where δ changes from 0 to 3.

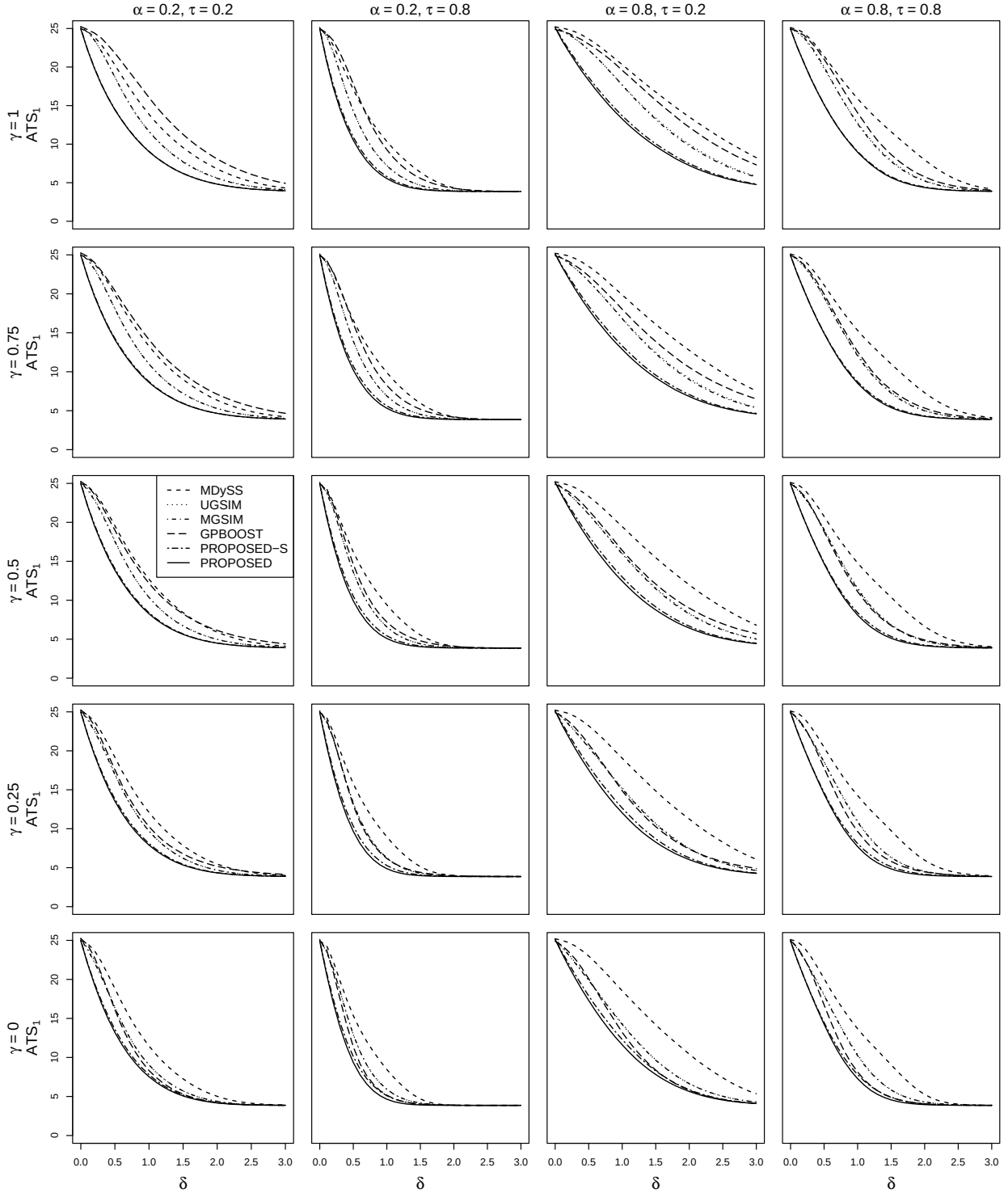
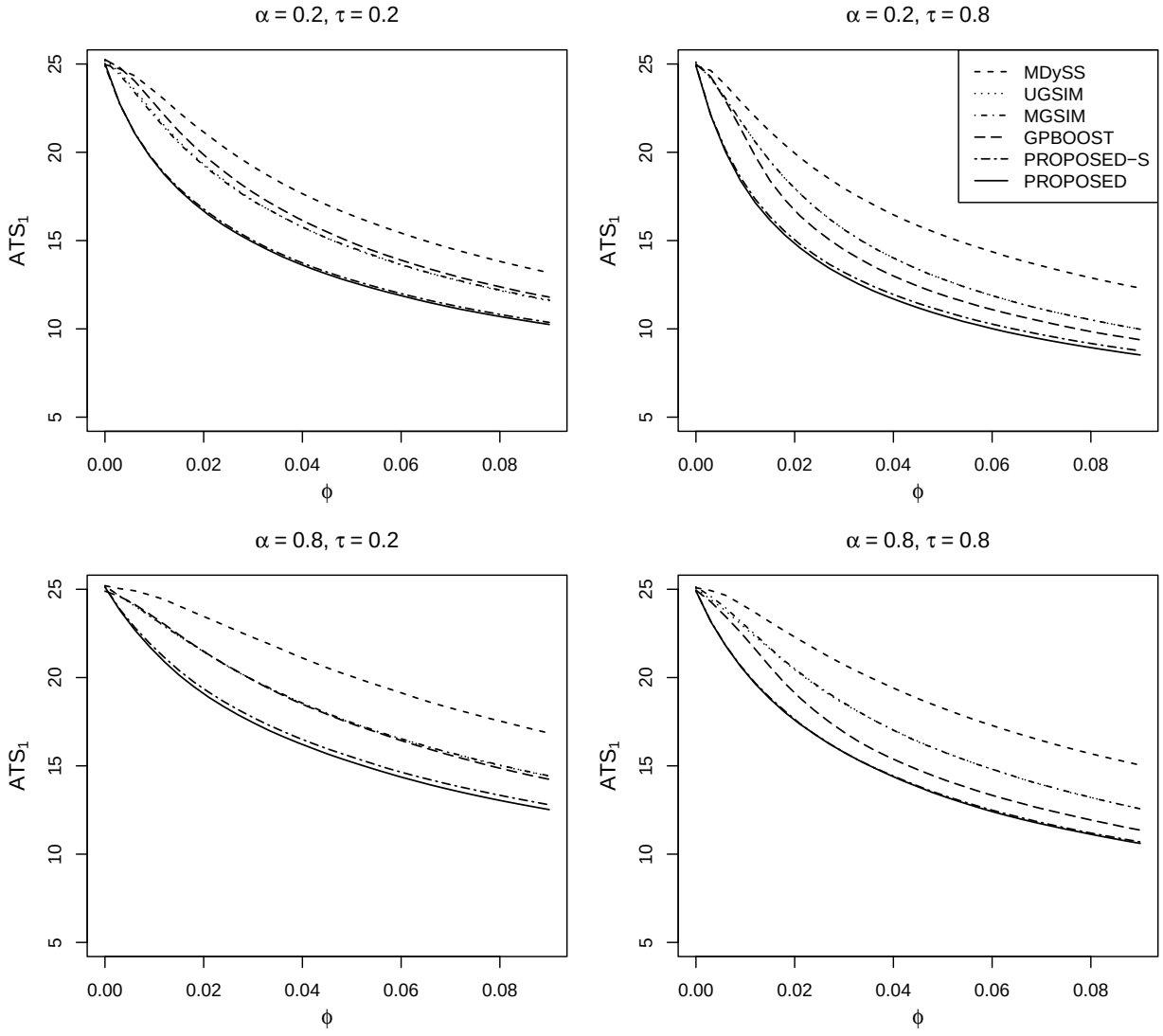


Figure 5: Optimal ATS_1 values of the charts MDySS, UGSIM, MGSIM, GPBOOST, PROPOSED-S and PROPOSED in various cases when the nominal ATS_0 value is 25, the white noise process $\mathbf{e}_X(t)$ has a normal distribution, the observed risk factors of the OC subjects have a mean drift over time with the slope controlled by ϕ where ϕ changes from 0 to 0.09.

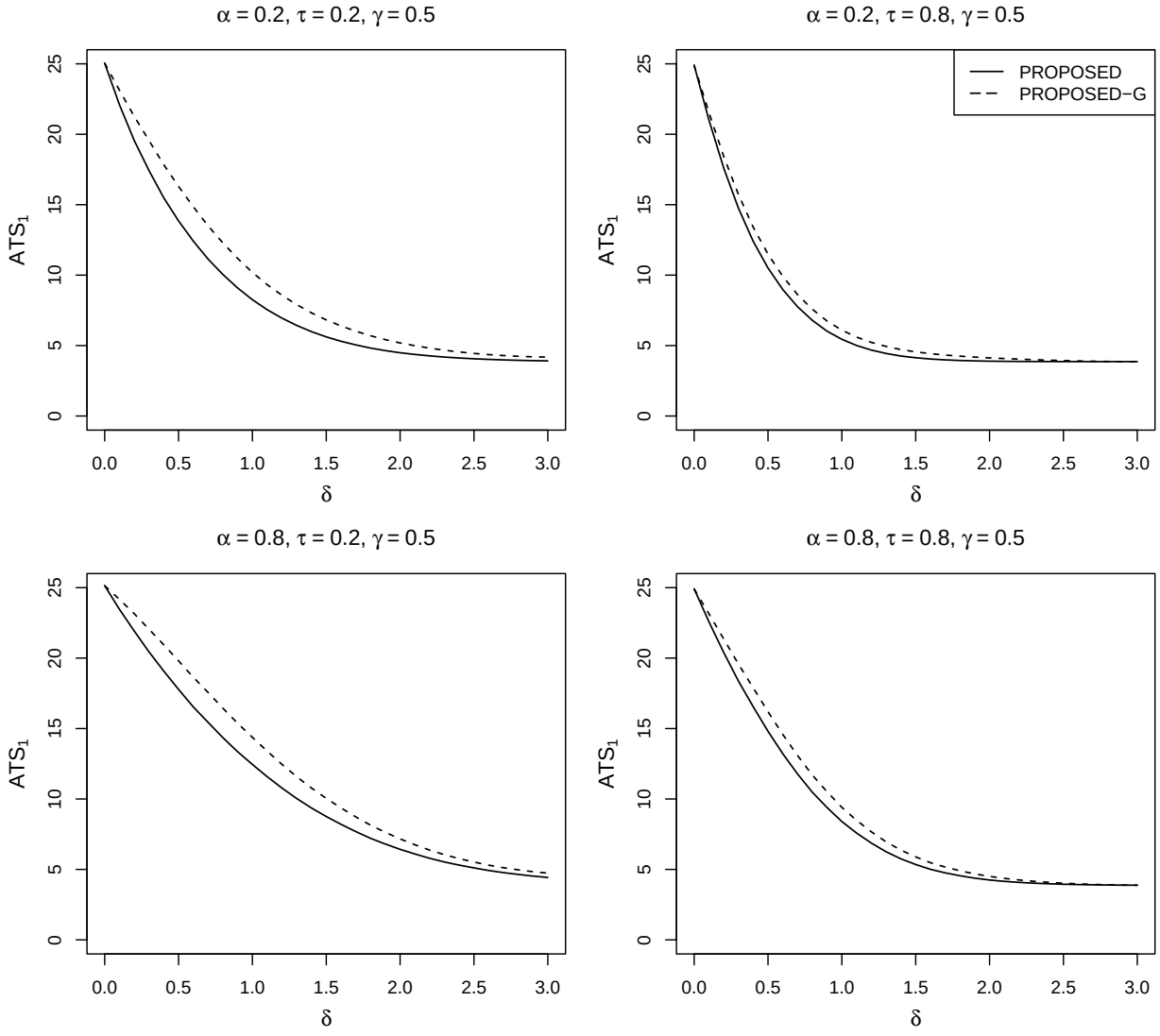


To further investigate the contribution of the disease-risk estimation step to the overall monitoring performance, we additionally compare the proposed method with a modified version, denoted by PROPOSED-G. The PROPOSED-G method follows the same two-stage monitoring framework as PROPOSED, including identical data decorrelation, standardization, and sequential monitoring procedures, but replaces the multivariate single-index logistic regression model in the first stage with the GPBOOST method for disease-risk estimation. Figure 6 presents the corresponding optimal ATS_1 values of PROPOSED and PROPOSED-G under various settings in which the white-noise process $\mathbf{e}_X(t)$ follows a normal distribution. In these settings, the observed risk factors of half OC subjects have mean shifts of size $\boldsymbol{\beta}_1\delta$, while those of the remaining OC subjects have mean shifts of size $\boldsymbol{\beta}_2\delta$, where δ varies from 0 to 3. The results show that PROPOSED consistently achieves smaller optimal ATS_1 values than PROPOSED-G across all cases, indicating superior OC performance. The performance gap is particularly pronounced when the shift size δ is relatively small, a setting in which accurate disease-risk estimation becomes especially important for early detection. These findings suggest that improved disease-risk estimation in the first stage can directly enhance the effectiveness of the subsequent sequential monitoring procedure.

3.4 Practical Guidelines

In this subsection, we provide some practical guidances on selecting the number of IC subjects used for training and the weighting parameter λ used in the proposed method. Tables A.2 and A.3 in Appendix C report the estimated ATS_0 and the corresponding standard errors (SEs) under combinations of $M_1 = 20, 50, 100, 200, 300$, $\overline{m} = 5, 10, 15$, and $\lambda = 0.05, 0.1, 0.2, 0.5$, with $\alpha = 0.2$ and $\tau = 0.8$. Results for other values of α and τ exhibit similar patterns and are therefore omitted for brevity. The white noise term used to generate observations of the risk factors follows either a normal distribution or a standardized version of the χ^2_3 distribution. Overall, the choice of λ has a negligible impact on ATS_0 and its associated SE, indicating that the proposed monitoring procedure is robust to the selection of the weighting parameter in terms of IC performance. As expected, the SE decreases as the number of IC subjects increases, reflecting improved estimation stability with larger training samples. In contrast, the average number of observation times per subject, $\overline{m}$, has little effect on the variability of ATS_0. For scenarios involving a small number of target diseases, such as the two-disease settings considered in our simulation studies, a moderate training sample size (e.g., on the order of a few hundred IC subjects) appears sufficient to achieve stable IC performance.

Figure 6: Optimal ATS_1 values of the charts PROPOSED and PROPOSED-G in various cases when the nominal ATS_0 value is 25, the white noise process $\mathbf{e}_X(t)$ has a normal distribution, $\gamma = 0.5$, the observed risk factors of $100\gamma\%$ OC subjects have mean shifts of sizes $\boldsymbol{\beta}_1\delta$, and the observed risk factors of remaining OC subjects have mean shifts of sizes $\boldsymbol{\beta}_2\delta$, where δ changes from 0 to 3.



Notably, the discussion above focuses on the stability of the monitoring performance, as measured by ATS_0, rather than on the accuracy of longitudinal model estimation. Although the quality of model estimation is influenced by factors such as the number of risk factors, the number of diseases, and the available sample size, the results in Tables A.2 and A.3 indicate that the stability of ATS_0 is primarily driven by the number of IC subjects M_1, and is much less sensitive to the average number of observation times per subject $\bar{m}$. These findings suggest that, from a monitoring perspective, collecting data from a sufficiently large number of IC subjects is more important for achieving stable false-alarm control than including individual subjects with more frequent measurements.

Figures A.9 and A.10 present the computed ATS_1 values of PROPOSED under a wide range of λ, across different choices of α and τ. In these experiments, the nominal ATS_0 value is fixed as 25, the white noise process $\mathbf{e}_X(t)$ has a normal distribution, and the OC subjects exhibit either abrupt mean shifts or gradual mean drifts in the risk factors, as described in Section 3.3. The results indicate that smaller (larger) values of λ are more effective for detecting smaller (larger) mean shifts or more slowly (more rapidly) increasing drifts. However, the overall differences in ATS_1 across a wide range of λ values are relatively modest. In practice, we recommend that users begin with a relatively small weighting parameter, like $\lambda = 0.1$, which provides good sensitivity to small to moderate shifts or gradual drifts while maintaining stable IC performance. Subsequent sensitivity analyses can be conducted if prior knowledge suggests the presence of larger or more abrupt changes in the risk factors and hence disease risks.

4 Early Detection of Myocardial Infarction and Stroke in the Framingham Heart Study

Myocardial infarction (MI) and stroke are two of the most significant and life-threatening manifestations of cardiovascular diseases (CVD) (Duprez, 2006; Kim et al., 2023). Both conditions result from disruptions in blood flow: MI is caused by blockages in the coronary arteries that supply blood to the heart, whereas stroke arises from ischemic or hemorrhagic events affecting the brain's blood vessels. The Framingham Heart Study (FHS) has been one of the most influential resources for studying MI, stroke, and other CVD manifestations, as well as for understanding their associated risk factors. Since its inception in 1948, the FHS has tracked multiple generations of participants in the city of Framingham in MA, providing invaluable insights into the development of CVD and other chronic conditions. By analyzing a broad range of clinical, environmental, and genetic factors, the study has identified key risk factors for cardiovascular con-

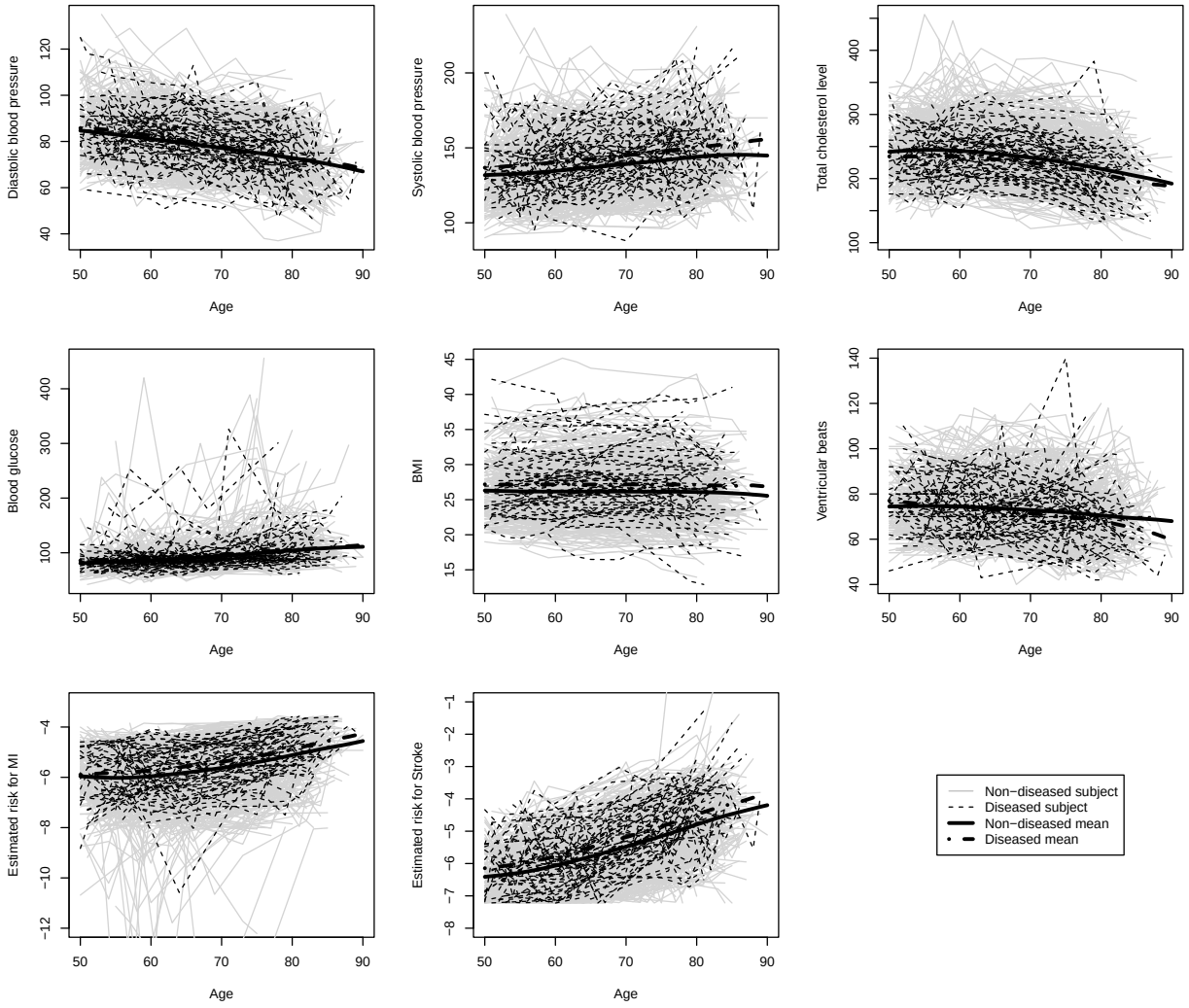
ditions, including hypertension (high blood pressure), diabetes (elevated blood glucose), high cholesterol, obesity (high BMI), and irregular heart rate (D'Agostino et al., 1994; Mahmood et al., 2014; Wilson et al., 2002). For a comprehensive overview of the FHS, covering its study design, participant follow-up protocols, examination cycles, and phenotype and outcome measures, refer to Tsao and Vasan (2015).

In this section, we use data from the FHS to demonstrate the application of the proposed method for early detection of MI and stroke. The analysis focuses on six key risk factors: systolic blood pressure (mmHg), diastolic blood pressure (mmHg), total cholesterol level (mg/dL), blood glucose (mg/dL), body mass index (BMI, kg/m^2), and ventricular beats (beats per minute). The dataset includes longitudinal data from 1,078 patients aged 50 and older, with each patient having seven follow-ups for the selected risk factors and disease diagnostic status. Among these patients, 945 reported no occurrence of MI or stroke, while 133 reported at least one occurrence of these diseases.

The dataset is then divided into a training dataset and a test dataset, ensuring approximately equal proportions of non-diseased and diseased subjects. The training dataset comprises 473 non-diseased subjects and 67 diseased subjects, while the test dataset includes 472 non-diseased subjects and 66 diseased subjects. The training dataset is used to estimate the multivariate single-index logistic regression model (1) and the regular longitudinal patterns of the risk factors and the estimated disease risks, as discussed in Sections 2.1 and 2.2. Using the estimated model, disease risks are computed for all patients. Figure 7 illustrates these results. The upper six panels depict the longitudinal observations of the six risk factors for the patients in the test dataset, while the lower two panels show the estimated risks of MI and stroke for each patient. In each panel, the corresponding mean curves for non-diseased and diseased subjects are also presented. Visual inspection reveals that the differences between the mean curves for non-diseased and diseased subjects are relatively small for diastolic blood pressure, blood glucose, and ventricular beats. However, the mean curves for the estimated disease risks to MI and stroke are well-separated, with an increasing trend observed for diseased patients compared to non-diseased patients.

Next, the related control charts are used to monitor either the risk factors or the computed disease risks of patients in the test dataset for the early detection of MI and stroke. To provide a more comprehensive comparison, Figure 8 displays the ATS_1 values of the six methods under various nominal ATS_0 levels and weighting parameter values. For comparability, the choices of threshold value ζ in GPBOOST are adjusted accordingly so that the given nominal ATS_0 can be achieved. The figure highlights that the proposed method PROPOSED and its simplified version PROPOSED-S are the best-performing approach across most of the cases considered. Additionally, Figure 9 presents the receiver operating characteristic (ROC) curves for the

Figure 7: Longitudinal observations of the 6 risk factors of MI and stroke, and estimated risks to MI and stroke over time. Solid gray lines denote longitudinal observations of non-diseased subjects, and black dashed lines denote longitudinal observations of diseased subjects. The two bold lines denote the sample means of the two groups of subjects.



six methods, obtained by varying their control limit or threshold values for the corresponding method and computing the resulting true positive and false positive rates (TPR and FPR) across the test dataset. The analysis considers four different weighting parameter values, 0.05, 0.1, 0.2, and 0.5. When $\lambda = 0.05$ or 0.1, the PROPOSED achieves the largest area under the curve (AUC). When $\lambda = 0.2$ or 0.5, PROPOSED and PROPOSED-S yield comparable AUC values, both of which are higher than those of the competing methods. These results indicate the superior accuracy of the proposed approach in distinguishing between diseased and non-diseased patients. Collectively, based on the results in Figures 8 and 9, we conclude that

the proposed method detects MI and stroke more promptly and accurately than the three competing methods. It is worth emphasizing that, although the AUC values of the proposed method are relatively modest in the cases considered in Figure 9, it achieves notable improvements over the best AUC values of competing methods MDySS, UGSIM, MGSIM and GPBOOST, by 7.4%, 6.1%, 3.2%, and 2.9% when λ equals 0.05, 0.1, 0.2, and 0.5, respectively. These improvements are practically significant, particularly in high-stakes applications like the early detection of MI and stroke (Latha and Jeeva, 2019).

Figure 8: Computed ATS_1 values of the six methods MDySS, UGSIM, MGSIM, GPBOOST, PROPOSED-S and PROPOSED under different nominal levels of ATS_0 when they are used to detect diseases for the 538 subjects in the test dataset. In each method, the weighting parameter λ changes among $\{0.05, 0.1, 0.2, 0.5\}$ and the nominal ATS_0 value changes among $\{10, 12, 14, 16, 18, 20, 22, 24\}$.

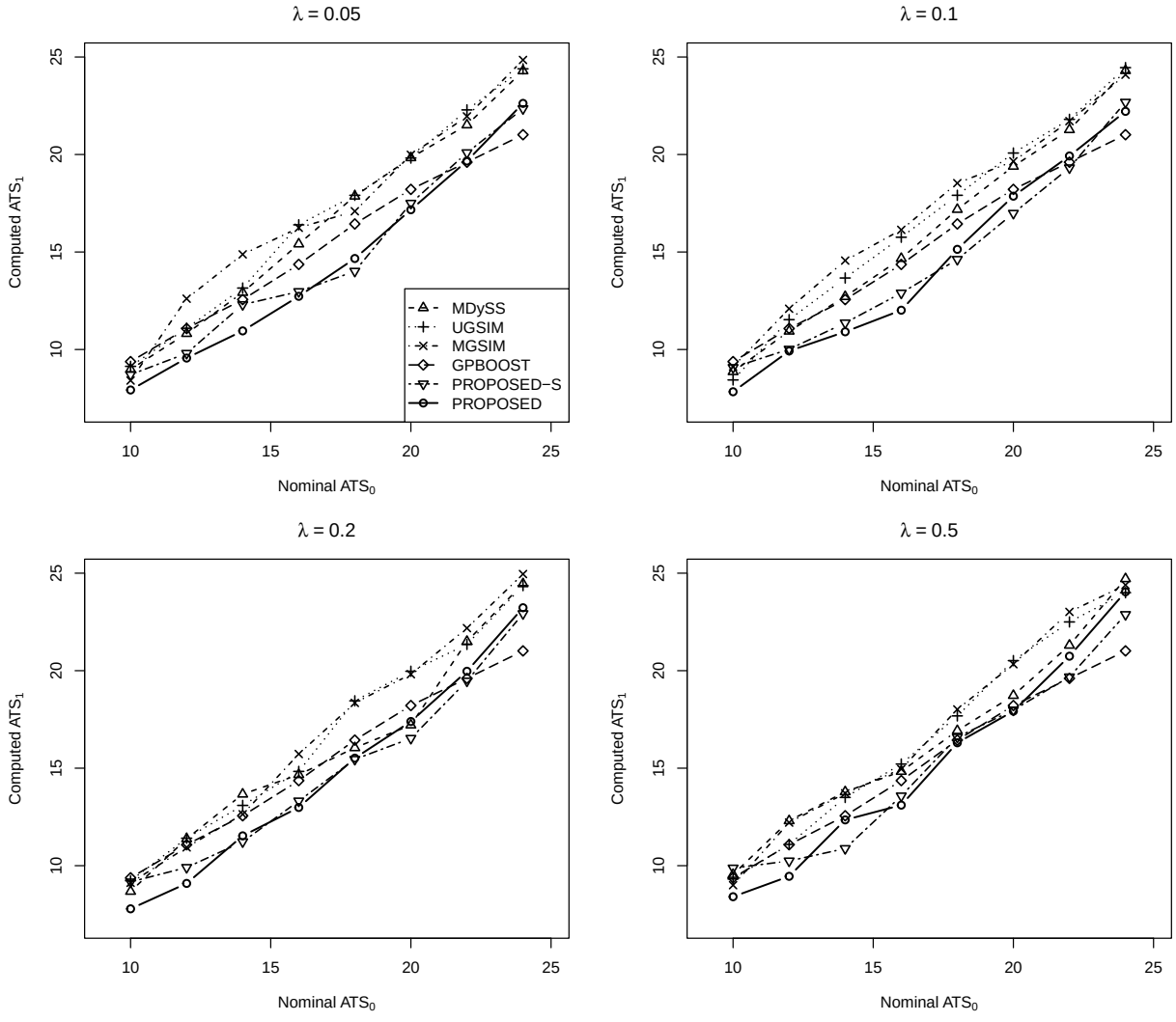
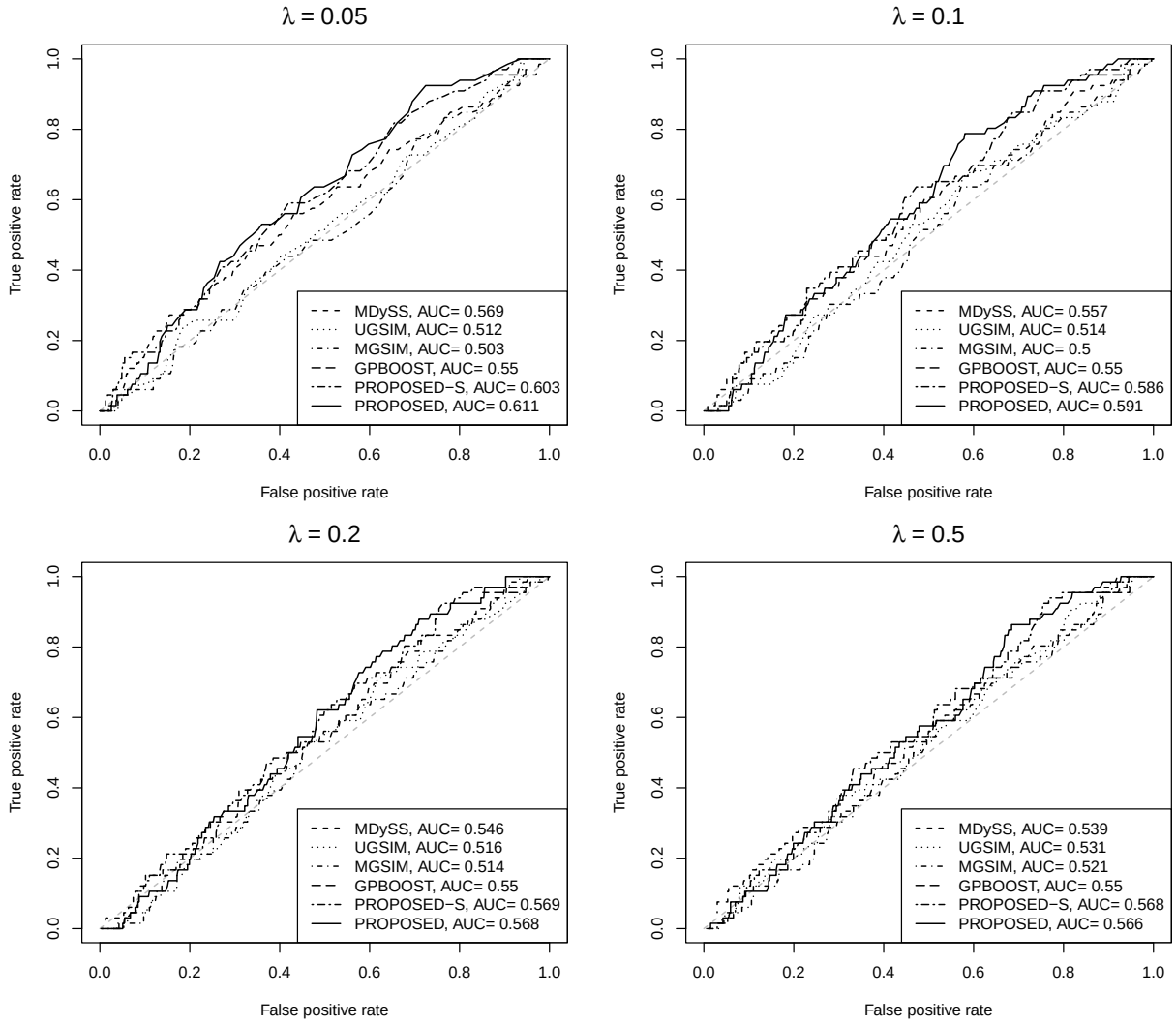


Figure 9: ROC curves for the six methods MDySS, UGSIM, MGSIM, GPBOOST, PROPOSED-S and PROPOSED are presented for detecting diseases among the 538 subjects in the test dataset. The weighting parameter λ of each method varies across $0.05, 0.1, 0.2, 0.5$. Corresponding AUC values for the six methods are also provided. In each plot, a dashed gray diagonal line serves as a reference, representing random guessing.



5 Concluding Remarks

In this paper, we introduce a novel method for the early detection of multiple diseases, emphasizing the quantification and sequential monitoring of disease risks. Numerical results highlight its robust performance across diverse scenarios. The proposed method addresses a critical gap in the detection of life-threatening conditions such as MI and stroke, particularly after identifying key risk factors like hypertension, elevated

cholesterol, diabetes, and obesity. It offers a systematic and efficient approach by enabling real-time monitoring of estimated disease risks for individuals and generating timely alerts when the cumulative deviation from regular longitudinal patterns exceeds a critical threshold. This proactive strategy supports earlier interventions, potentially preventing disease onset and enhancing patient outcomes.

However, there are still some challenges with the proposed method that need to be addressed in future research. For example, the method is primarily designed for timely and accurate detection of abnormal multivariate disease-risk patterns, rather than direct identification of which specific disease is primarily responsible for an OC signal. In the proposed framework, the decorrelation and standardization procedure is introduced mainly to stabilize the distribution of the charting statistics and improve the reliability of sequential monitoring under serial and cross-sectional dependence. While this transformation may reduce some direct interpretability of the raw disease-risk trajectories, post-signal diagnosis remains an important practical issue because identifying the disease(s) contributing to the signal is crucial for tailoring medical treatments to individual patients.

The simplified UGSIM approach offers a partial solution by employing multiple univariate control charts to monitor estimated disease risks separately, allowing identification of the specific chart that triggers a signal. However, this monitoring scheme may lack efficiency compared with the proposed joint monitoring framework. In addition, post-signal diagnostic analysis based on the original (non-transformed) disease-risk trajectories, such as disease-specific distributional comparisons with the estimated IC longitudinal patterns, may provide a more direct interpretation of the abnormal disease-risk dimensions. Further research is needed to develop and systematically evaluate efficient and robust post-signal diagnostic procedures for multiple-disease monitoring problems.

Moreover, since the charting statistic in (5) aggregates multiple disease-risk dimensions, incorporating a sparsity-inducing variable-selection step before forming this statistic may further improve detection efficiency when only a subset of transformed risk components becomes out of control (Zou and Qiu, 2009; Xie, 2025), which is another interesting topic for future research. Furthermore, while the current framework can be implemented with a moderately large number of risk factors, estimation of the single-index coefficients and nonparametric link functions may become less efficient in high-dimensional settings. Developing regularized estimation procedures, such as incorporating penalization or shrinkage techniques into the estimation of the index coefficients, could therefore further improve the scalability and numerical performance of the proposed method in applications involving high-dimensional clinical risk factors.

Another promising direction for future work concerns the explicit modeling of heterogeneity across patients. Although the proposed framework has the flexibility to use multiple disease risks for describing the association between different diseases and their observed risk factors, it currently assumes that all IC or well-functioning subjects share a common longitudinal risk structure, characterized by a unified mean function and covariance structure. In real-world clinical settings, however, individuals may have fundamentally different baseline health conditions, physiological resilience, or aging trajectories. As a result, the same level or change in estimated disease risk may carry different clinical implications for different patients. Future extensions of the proposed framework would incorporate subject-specific longitudinal models, which make the regular risk patterns depend not only on time but also on baseline and time-varying covariates, or even latent health states. Such extensions could allow the definition of personalized health standards and improve the precision, interpretability and clinical relevance of monitoring signals in heterogeneous populations.

Acknowledgments

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

Data Availability Statement

The FHS datasets analyzed during this study are available in the Biologic Specimen and Data Repository Information Coordinating Center, https://biolincc.nhlbi.nih.gov/studies/framcohort/.

Supplementary Materials

To save some space in the paper with the above title, the Appendices A to C mentioned in the main article are provided in this supplementary file.

References

Andersson, C., Johnson, A. D., Benjamin, E. J., Levy, D., and Vasan, R. S. (2019). 70-year legacy of the Framingham Heart Study. *Nature Reviews Cardiology*, **16**(11), 687–698.

Bauer, U. E., Briss, P. A., Goodman, R. A., and Bowman, B. A. (2014). Prevention of chronic disease in the 21st century: elimination of the leading preventable causes of premature death and disability in the USA. *The Lancet*, **384**(9937), 45–52.

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.

Christensen, K., Doblhammer, G., Rau, R., and Vaupel, J. W. (2009). Ageing populations: the challenges ahead. *The Lancet*, **374**(9696), 1196–1208.

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

D'Agostino, R. B., Wolf, P. A., Belanger, A. J., and Kannel, W. B. (1994). Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke*, **25**(1), 40–43.

De Brabanter, K., De Brabanter, J., Suykens, J. A., and De Moor, B. (2011). Kernel Regression in the Presence of Correlated Errors. *Journal of Machine Learning Research*, **12**, 1955–1976.

Duprez, D. (2006). Early detection of cardiovascular disease-the future of cardiology. *E-Journal of the ESC Council for Cardiology Practice*, **4**(19), 1.

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

Fox, C. S. (2010). Cardiovascular disease risk factors, type 2 diabetes mellitus, and the Framingham Heart Study. *Trends in Cardiovascular Medicine*, **20**(3), 90–95.

Greenberg, H. and Pi-Sunyer, F. X. (2019). Preventing preventable chronic disease: an essential goal. *Progress in Cardiovascular Diseases*, **62**(4), 303–305.

Hacker, K. (2024). The burden of chronic disease. *Mayo Clinic Proceedings: Innovations, Quality and Outcomes*, **8**(1), 112–119.

Hajar, R. (2017). Risk factors for coronary artery disease: historical perspectives. *Heart Views*, **18**(3), 109–114.

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

Kennedy, B. K., Berger, S. L., Brunet, A., Campisi, J., Cuervo, A. M., Epel, E. S., Franceschi, C., Lithgow, G. J., Morimoto, R. I., Pessin, J. E., et al. (2014). Geroscience: linking aging to chronic disease. *Cell*, **159**, 709–713.

Kim, G., Jang, H., Kwon, S., Lee, B., Jang, S. Y., Chae, W., and Jang, S. I. (2023). Engaging social activities prevent stroke and myocardial infarction by raising awareness of warning symptoms: A cross-sectional survey study. *Frontiers in Public Health*, **10**, 1043875.

Latha, C.B.C. and Jeeva, S.C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, **16**, 100203.

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.

Lloyd-Jones, D. M., Wilson, P. W., Larson, M. G., Beiser, A., Leip, E. P., D'Agostino, R. B., and Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary heart disease. *The American Journal of Cardiology*, **94**(1), 20–24.

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

Mahmood, S. S., Levy, D., Vasan, R. S., and Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, **383**(9921), 999–1008.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**(437), 162–170.

Perel, P., Avezum, A., Huffman, M., Pais, P., Rodgers, A., Vedanthan, R., ... and Yusuf, S. (2015). Reducing premature cardiovascular morbidity and mortality in people with atherosclerotic vascular disease. The WHF roadmap for secondary prevention of cardiovascular disease. *Global Heart*, **10**, 99–110.

Perry, M. B. (2024). Joint monitoring of location and scale for modern univariate processes. *Journal of Quality Technology*, 56(5), 409-427.

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.

Qiu, P., Zi, X., and Zou, C. (2018). Nonparametric dynamic curve monitoring. *Technometrics*, **60**(3), 386–397.

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Babbour, L. M., ... (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American College of Cardiology*, **76**(25), 2982–3021.

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.

Sigrist, F. (2022). Gaussian process boosting. *Journal of Machine Learning Research*, **23**(232), 1–46.

Sinnige, J., Braspenning, J., Schellevis, F., Stirbu-Wagner, I., Westert, G., and Korevaar, J. (2013). The prevalence of disease clusters in older adults with multiple chronic diseases–a systematic literature review. *PLoS ONE*, **8**(11), e79641.

Spanbauer, C., and Sparapani, R. (2021). Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. *Statistics in Medicine*, **40**(11), 2665–2691.

Staerk, L., Preis, S. R., Lin, H., Lubitz, S. A., Ellinor, P. T., Levy, D., ... and Trinquart, L. (2020). Protein biomarkers and risk of atrial fibrillation: the FHS. *Circulation: Arrhythmia and Electrophysiology*, **13**(2), e007607.

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

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.

Wilson, P. W., D'Agostino, R. B., Sullivan, L., Parise, H., and Kannel, W. B. (2002). Overweight and obesity as determinants of cardiovascular risk: the Framingham experience. *Archives of Internal Medicine*, **162**(16), 1867–1872.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**(6), 1112–1137.

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

Xie, X. (2025), "Adaptive LASSO-based robust multivariate process monitoring and diagnostics," *Quality and Reliability Engineering International*, In press.

Xie, X., and Qiu, P. (2023). Control charts for dynamic process monitoring with an application to air pollution surveillance. *The Annals of Applied Statistics*, **17**(1), 47–66.

Xue, L., and Qiu, P. (2021). A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology*, **53**(4), 396—409.

Yang, Z., Mitra, A., Liu, W., Berlowitz, D., and Yu, H. (2023). TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, **14**(1), 7857.

Yao, Y., Chakraborti, S., Yang, X., Parton, J., Lewis Jr, D., and Hudnall, M. (2023). Phase I control chart for individual autocorrelated data: application to prescription opioid monitoring. *Journal of Quality Technology*, **55**(3), 302-317.

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.

Zhao, Z. and Wu, W. B. (2008). Confidence bands in nonparametric time series regression. *Annals of Applied Statistics*, **36**(4), 1854–1878.

Zou, C., and Qiu, P. (2009). Multivariate statistical process control using LASSO. *Journal of the American Statistical Association*, 104, 1586–1596.