Received: 1 July 2020 | Revised: 21 December 2020 | Accepted: 30 December 2020

DOI: 10.1002/sim.8880

RESEARCH ARTICLE

Statistics in Medicine WILEY

Learning latent heterogeneity for type 2 diabetes patients using longitudinal health markers in electronic health records

Jitong Lou[1] | **Yuanjia Wang**[2] | **Lang Li**[3] | **Donglin Zeng**[1]

[1] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[2] Department of Biostatistics, Columbia University, New York City, New York

[3] Department of Biomedical Informatics, Ohio State University, Columbus, Ohio

Correspondence
Donglin Zeng, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC.
Email: dzeng@email.unc.edu

Present address
Gillings School of Global Public Health, University of North Carolina, 3103B McGavran-Greenberg Hall, Chapel Hill, North Carolina, 27599.

Funding information
National Institute of General Medical Sciences, Grant/Award Number: GM124104; National Institute of Mental Health, Grant/Award Number: MH117458; National Institute of Neurological Disorders and Stroke, Grant/Award Number: NS073671

Electronic health records (EHRs) from type 2 diabetes (T2D) patients consist of longitudinally and sparsely measured health markers at clinical encounters. Our goal is to use such data to learn latent patterns that can inform patient's health status related to T2D while accounting for challenges in retrospectively collected EHRs. To handle challenges such as correlated longitudinal measurements, irregular and informative encounter times, and mixed marker types, we propose multivariate generalized linear models to learn latent patient subgroups. In our model, covariate effects were time-dependent and latent Gaussian processes were introduced to model between-marker correlations over time. Using inferred latent processes, we integrated the irregularly measured health markers of mixed types into composite scores and applied hierarchical clustering to learn latent subgroup structures among T2D patients. Application to an EHR dataset of T2D patients showed different trends of age, sex, and race effects on hypertension/high blood pressure, total cholesterol, glycated hemoglobin, high-density lipoprotein, and medications. The associations among these markers varied over time during the study window. Clustering results revealed four subgroups, each with distinct health status. The same patterns were further confirmed using new EHR records of the same cohort. We developed a novel latent model to integrate longitudinal health markers in EHRs and characterize patient latent heterogeneities. Analysis indicated that there were distinct subgroups of T2D patients, suggesting that effective healthcare managements for these patients should be performed separately for each subgroup.

KEYWORDS

electronic health records, generalized linear models, kernel smoothing, latent process, type 2 diabetes

1 | INTRODUCTION

In the modern era of precision medicine, one important source of patient's health data is electronic health records (EHRs). EHR data consist of longitudinal medical records from a large number of patients in one or more electronic healthcare systems that digitally capture measurements of patients health status through normal medical practices,[1-3] including

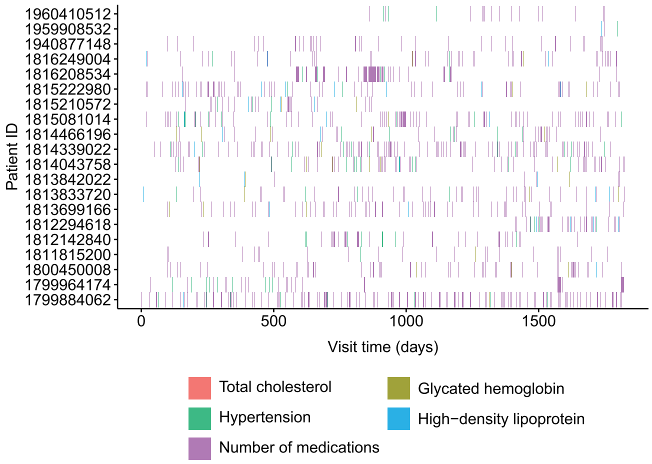


FIGURE 1 Observation time patterns of five health markers for 20 randomly selected T2D patients in the EHRs at the Ohio State University Wexner Medical Center. Each mark represents observing a measurement at the corresponding time point for a marker [Colour figure can be viewed at wileyonlinelibrary.com]

patient's vital signs, laboratory measurements, disease diagnosis codes, procedure codes, and medications. Benefits of EHRs include cost effectiveness, real time updates, and reflections on patients disease course and healthcare managements in realistic settings. Therefore, integrative analysis of this information over time provides a great opportunity to understand the heterogeneity of patient's disease progression and susceptibility in real-world settings, which is useful for monitoring disease prognosis and optimizing personalized healthcare management.

Due to the retrospective nature of EHRs, the analysis of EHRs is complicated by the following challenges: first, the health markers measured over time are multivariate and the measurements can be either continuous (eg, lab measures), binary (eg, disease diagnoses), or counts (eg, number of medications); second, for each patient, the health marker data are collected at each clinical encounter so the measurement times can be irregular, sparse, and heterogeneous across patients; third, the measurement times are often informative to patients health status or health care processes.

This work is motivated by the analyses of EHRs of type 2 diabetes (T2D) patients obtained from the Ohio State University Wexner Medical Center Information Warehouse (OSU-WMCIW). The data collection spanned a time period of 8 years (between 2011 and 2018) from a total of 58 490 patients. The data contained patients medical records of glycated hemoglobin, high-density lipoprotein, total cholesterol, hypertension, and all medications prescribed at each clinical encounter. Because these markers were of different types and were not measured at the same time across and within patients, directly combining the values from these markers is neither meaningful nor feasible. For example, Figure 1 gives a snapshot of the measurement time of several health markers from 20 randomly selected patients. Clearly, each marker was measured sparsely at irregular times for each patient, and the measurement time patterns vary significantly from patient to patient.

Joint models based on linear or generalized mixed effects models have been commonly used for analyzing multivariate longitudinal data.[4] In the joint models, various distribution families are used,[5-7] and subject-specific random effects are shared across all health markers to explain their dependence due to a finite number of latent variables. For example, Lambert and Vandenhende[8] jointly analyzed three repeatedly measured longitudinal outcomes using copula models in a dose titration safety study; Gueorguieva and Sanacora[9] proposed correlated probit models for joint analysis of repeated measurements with ordinal and continuous health markers. Some extensions allowed time-dependent effects,[10,11] but assumed constant between-marker dependence over time. However, assuming parametric patterns or attributing the dependence to a few time-invariant random effects is rather restrictive especially for modeling EHRs over a long period of time, since in EHRs, the trajectories of the health markers and their dependence may vary over time depending on the disease progression and medication usage for each patient. Moreover, it is computationally challenging to maximize a joint likelihood in the presence of a large number of patients and many health markers.

Machine learning approaches have been also proposed to perform EHR analysis, such as deep Poisson factor models,[12] tensor factorization and nonnegative matrix factorization,[13] and deep exponential families.[14] These approaches, although more flexible than aforementioned statistical models, are less interpretable and are highly computationally intensive, requiring substantial work for data engineering and model tuning. More importantly, none of these approaches can account for irregular but informative measurement patterns as seen in EHRs.

In this work, we seek to strike a balance between the complex statistical modeling and flexible machine learning methods, while accounting for the unique challenges in EHRs. To conduct an integrative analysis of EHRs, we extend the multivariate generalized linear models (GLMs) by assuming appropriate distribution and link functions depending on the marker type. We allow the effects of covariates on health markers to be time-varying. Moreover, to account for the time-varying dependence among health markers, we introduce latent Gaussian processes into the models, where the covariance matrix is assumed to vary over time. For estimation, we adopt kernel smoothing method to pool information across time points and patients and apply weights to account for the heterogeneous patterns of measurement times. The inferred latent processes represent patients underlying health status, so in order to integrate these mixed-type health markers, we use the inferred latent processes to calculate the distances between any two patients using the Mahalanobis distance.[15] Finally, we apply hierarchical clustering to identify patients health patterns and characterize between-group heterogeneities.

The remaining parts of this article are organized as follows. In Section 2, we propose our models and describe main ideas. We then provide inferences on estimating model parameters and procedures to perform numerical computations. In Section 3, we derive the asymptotic distributions of the estimators. We conduct simulation studies in Section 4. In Section 5, we apply our method to an integrative analysis on health markers for T2D patients using EHRs from the OSU-WMCIW.

2 | METHODOLOGIES

2.1 | Statistical models for integrative analysis

Suppose EHR data are obtained from n patients. For the ith patient, let $\boldsymbol{X}_i$ be m-dimensional baseline covariates. Among p health markers, let $Y_{ik}(t)$ denote the measurement of the kth health marker at time t. We suppose $Y_{ik}(t)$ is measured at time points $t_{ik1}, t_{ik2}, \dots, t_{ikn_{ik}}$, where n_{ik} is the total count of observations on the kth health marker for the ith patient. The total number of observations up to time t can be represented by a counting process $N_{ik}(t) \equiv \sum_{j=1}^{n_{ik}} I(t_{ikj} \leq t)$, where $I(\cdot)$ is the indicator function. Since the documentation times are patient's clinical encounters in the EHR system, patterns of these documentation/measurement time points may carry information on patients health status. Thus, we model the intensity of $N_{ik}(t)$ as

$$\mathbb{E}\left[dN_{ik}(t)|\boldsymbol{X}_i\right] = \lambda_k(t)\exp\{\boldsymbol{X}_i^T\boldsymbol{\gamma}_k\}dt, \tag{1}$$

where $\lambda_k(t)$ is a baseline intensity function, and γ_k is a vector of intensity parameters. By modeling the intensity of EHR measurement rates, one can adjust for the bias of informative measurement patterns and account for between patient heterogeneity.

We further assume $Y_{ik}(t)$ follows a distribution in an exponential family model as follows:

$$f_{ik}(y;\theta_{ik},\phi_{ik},t) = \exp\left\{\frac{y\theta_{ik}(t) - b_k(\theta_{ik}(t))}{a_k(\phi_{ik}(t))} + c_k(y,\phi_{ik}(t))\right\}, \tag{2}$$

where $\theta_{ik}(t)$ and $\phi_{ik}(t)$ are the canonical parameter and the dispersion parameter, respectively, specific to each patient and each health marker. $a_k(\cdot)$, $b_k(\cdot)$, and $c_k(\cdot)$ are known functions. Let $\theta_{ik}(t) = g_k(\mu_{ik}(t))$, where $g_k(\cdot)$ is the canonical link function, and $\mu_{ik}(t)$ is the mean of $Y_{ik}(t)$. To capture the patient heterogeneity and dependence, we assume, at time t,

$$\begin{gathered} g_k(\mu_{ik}(t)) = \boldsymbol{X}_i^T\boldsymbol{\beta}_k(t) + \epsilon_{ik}(t), \\ \boldsymbol{\epsilon}_i(t) \sim \mathcal{N}_p(\boldsymbol{0},\boldsymbol{\Omega}(t)), \end{gathered} \tag{3}$$

where $\boldsymbol{\beta}_k(t)$ is a vector of regression coefficients for covariates $\boldsymbol{X}_i$. $\epsilon_{ik}(t)$ is the kth element of the latent Gaussian process $\boldsymbol{\epsilon}_i(t) = \{\epsilon_{i1}(t), \epsilon_{i2}(t), \dots, \epsilon_{ip}(t)\}^T$. $\boldsymbol{\epsilon}_i(t)$ is independent of $\boldsymbol{X}_i$, and it follows a mean-zero multivariate Gaussian distribution with a covariance matrix $\boldsymbol{\Omega}(t)$. Estimating variances locally will requires dense measurements from the same biomarker, which is not the case for the EHRs. Moreover, in our empirical application the estimated variances do not vary much across time (Section 6). Thus, to ensure numerical stability in subsequent analysis, we assume each latent process to have a constant variance and the constant is estimated using historical records. Hence, in $\boldsymbol{\Omega}(t)$, only the correlations among health markers, that is, the off-diagonal elements, need to be estimated.

Under the proposed models (2) and (3), each measurement $Y_{ik}(t)$ can be uniquely represented by the latent process $\epsilon_{ik}(t)$. Since $\epsilon_{ik}(t)$ has the same scale for different k, one can integrate the latent processes $\{\epsilon_{ik}(t) : k = 1, 2, \dots , p\}$ as an alternative way to integrate the mixed-type health markers. The integration can use the Mahalanobis distance as follows,

$$D_{ij} = \left\{ \int_t \left[\boldsymbol{\epsilon}_i(t) - \boldsymbol{\epsilon}_j(t)\right]^T \boldsymbol{\Omega}^{-1}(t) \left[\boldsymbol{\epsilon}_i(t) - \boldsymbol{\epsilon}_j(t)\right] dt \right\}^{1/2}. \tag{4}$$

Thus, there are several important advantages of using the proposed models to perform an integrative analysis of mixed-type health markers. First of all, despite the health markers are irregularly measured and mixed-type, we can map them onto the same scale to align patients and characterize the between-patients heterogeneity. In addition, the dimension of latent processes can be further reduced to some lower dimensional subspaces than the number of health markers. Therefore, through the representation of latent processes, we achieve a dimension reduction.

2.2 | Model parameter estimation

First, we use marker-specific Anderson-Gill intensity models[16] to estimate $\boldsymbol{\gamma}_k$ in (1). With the estimator $\hat{\boldsymbol{\gamma}}_k$, we normalize the counting process $N_{ik}(t)$ by letting $\tilde{N}_{ik}(t) = N_{ik}(t)\exp\{-\boldsymbol{X}_i^T\hat{\boldsymbol{\gamma}}_k\}$. Thus, the normalized counting process is homogeneous across different patients and different health markers.

Next, to estimate $\boldsymbol{\beta}_k(t)$ for any fixed time point t, we solve the following kernel-weighted local estimating equation

$$U_{n,k}(\boldsymbol{\beta}_k(t)) \equiv \frac{1}{n}\sum_{i=1}^{n} \int \boldsymbol{X}_i \left[Y_{ik}(s) - \mathbb{E}\left[Y_{ik}(t)|\boldsymbol{X}_i\right]\right] K_{h_{1n}}(s-t)d\tilde{N}_{ik}(s) = \mathbf{0}, \tag{5}$$

where $K_h(z) = h^{-1}K(z/h)$ with $K(z)$ being a symmetric kernel function, and h_{1n} is the bandwidth of $K_h(z)$. Essentially, we assign weights to the observed measurements $Y_{ik}(s)$ near t, and we pool them together across all patients to estimate the mean (first moment) of $Y_{ik}(t)$. This pooling process relies on the kernel smoothing. Also, pooling information across observations nearby and across patients overcomes the difficulty in parameter estimations that some sparsely measured health markers do not have sufficient samples at some time points. Moreover, using $d\tilde{N}_{ik}(s)$ instead of $dN_{ik}(s)$, we remove the heterogeneity of informative measurement time points among patients in a similar spirit as inverse probability weighting.

Similarly, to estimate the correlation between two latent processes, $\sigma_{kl}(t) = \text{Cov}(\epsilon_{ik}(t), \epsilon_{il}(t))$, we propose to solve the following kernel-weighted local estimating equation, for $k \neq l$,

$$U_{n,k,l}(\sigma_{kl}(t)) \equiv \frac{1}{n^2}\sum_{i=1}^{n} \iint \left[Y_{ik}(s)Y_{il}(s') - \mathbb{E}\left[Y_{ik}(t)Y_{il}(t)|\boldsymbol{X}_i\right]\right] \tilde{K}_{h_{2n}}(s-t, s'-t)d\tilde{N}_{ik}(s)d\tilde{N}_{il}(s') = 0, \tag{6}$$

where $\tilde{K}_h(z_1, z_2)$ is a bivariate kernel function with bandwidth h_{2n}.

2.3 | Numerical computation

When the link functions in (3) take some simple forms, $\mathbb{E}\left[Y_{ik}(t)|\boldsymbol{X}_i\right]$ in (5) and $\mathbb{E}\left[Y_{ik}(t)Y_{il}(t)|\boldsymbol{X}_i\right]$ in (6) can be explicitly computed. Specifically, for $g_k(z) = g_l(z) = z$,

$$\mathbb{E}\left[Y_{ik}(t)|\boldsymbol{X}_i\right] = \boldsymbol{X}_i^T\boldsymbol{\beta}_k(t),$$

and

$$\mathbb{E}[Y_{ik}(t)Y_{il}(t)|\boldsymbol{X}_i] = \boldsymbol{X}_i^T\boldsymbol{\beta}_k(t)\boldsymbol{X}_i^T\boldsymbol{\beta}_l(t) + \sigma_{kl}(t).$$

When $g_k(z)$ takes a general form, we can compute the above expectations using the Gauss-Hermite quadrature method.[17]

Since $U_{n,k}(\boldsymbol{\beta}_k(t))$ is only related to the parameter $\boldsymbol{\beta}_k(t)$, we can solve (5) and obtain $\hat{\boldsymbol{\beta}}_k(t)$ for each health marker k, separately. Similarly, plugging $\hat{\boldsymbol{\beta}}_k(t)$ and $\hat{\boldsymbol{\beta}}_l(t)$ to (6), we can solve the equation and obtain $\hat{\sigma}_{kl}(t)$ for each pair of health markers, separately. Therefore, even with many health markers, that is, p is moderate or large, our algorithm can efficiently handle the computation burden by solving the estimating equations separately. Finally, we apply the above procedures for time grids $t_1, t_2, \dots, t_N$ to obtain the parameter estimators over the whole range of the follow-up.

A distance matrix D can be obtained by computing the Mahalanobis distance in (4) between each pair of patients. In particular, with the estimated latent processes, the distance is approximated by

$$D_{ij} = \left\{ \sum_{t=t_1}^{t_N} \left[\hat{\epsilon}_i(t) - \hat{\epsilon}_j(t)\right]^T \tilde{\boldsymbol{\Omega}}^{-1}(t) \left[\hat{\epsilon}_i(t) - \hat{\epsilon}_j(t)\right] \right\}^{1/2}, \tag{7}$$

and

$$\hat{\epsilon}_i(t) = \mathbb{E}\left[\boldsymbol{\epsilon}_i(t) \middle| \boldsymbol{Y}_i(t), \hat{\boldsymbol{\beta}}_k(t), \hat{\sigma}_{kl}(t)\right], \tag{8}$$

where $\tilde{\boldsymbol{\Omega}}(t)$ is the covariance matrix of $\hat{\epsilon}_i(t)$. In particular,

$$\mathbb{E}\left[\boldsymbol{\epsilon}_i(t) \middle| \boldsymbol{Y}_i(t), \hat{\boldsymbol{\beta}}_k(t), \hat{\sigma}_{kl}(t)\right] = \frac{\int P(\boldsymbol{Y}_i(t) \big| \boldsymbol{\epsilon}_i(t), \hat{\boldsymbol{\beta}}_k(t), \hat{\sigma}_{kl}(t)) P(\boldsymbol{\epsilon}_i(t) | \hat{\sigma}_{kl}(t)) \boldsymbol{\epsilon}_i(t) d\boldsymbol{\epsilon}_i(t)}{\int P(\boldsymbol{Y}_i(t) \big| \boldsymbol{\epsilon}_i(t), \hat{\boldsymbol{\beta}}_k(t), \hat{\sigma}_{kl}(t)) P(\boldsymbol{\epsilon}_i(t) | \hat{\sigma}_{kl}(t)) d\boldsymbol{\epsilon}_i(t)}.$$

The subsequent steps can be calculated using the Gauss-Hermite quadrature method as well, and the details are given in the supplementary material.

2.4 | Data-adaptive selection of bandwidths

Our asymptotic results in the supplementary material suggest the bandwidths h_{1n} and h_{2n} can be chosen, respectively, on the order of $n^{-1/3}$ and $n^{-1/4}$. However, for practical applications, we consider a data-adaptive method for selecting the bandwidths.[18] The key idea is using observed data to obtain the empirical bias and variability of the estimators in terms of the bandwidths. Consequently, we search for the bandwidths that minimize the empirical mean squared error of selecting them.

Specifically, to choose the optimal bandwidth h_{1n} for estimating $\hat{\boldsymbol{\beta}}_k(t)$, we first consider a reasonable range of bandwidths. For a fixed bandwidth h and a fixed time point t, we denote $\hat{\boldsymbol{\beta}}_{kh}(t)$ to the estimator for $\boldsymbol{\beta}_k(t)$. To estimate the bias of $\hat{\boldsymbol{\beta}}_{kh}(t)$, we fit a least squares regression by regressing $\hat{\boldsymbol{\beta}}_{kh}(t)$ on h^2. We denote the regression coefficient of h^2 as $\hat{\boldsymbol{C}}_k(t)$. Since the bias of $\hat{\boldsymbol{\beta}}_{kh}(t)$ is on the order of h^2, as shown in the asymptotic result, $\|\hat{\boldsymbol{C}}_k(t)\| h^2$ is an estimator for the bias of $\hat{\boldsymbol{\beta}}_{kh}(t)$. Next we investigate the variability of $\hat{\boldsymbol{\beta}}_{kh}(t)$. We randomly split the data into two equal parts. Using either one of the split data, we obtain $\hat{\boldsymbol{\beta}}^{*1}_{kh}(t)$ as the estimator for $\boldsymbol{\beta}_{kh}(t)$ in this case. Similarly, using the other half, we obtain $\hat{\boldsymbol{\beta}}^{*2}_{kh}(t)$. Thus, $\frac{1}{4}\|\hat{\boldsymbol{\beta}}^{*1}_{kh}(t) - \hat{\boldsymbol{\beta}}^{*2}_{kh}(t)\|^2$ can be used as an unbiased estimator of the variance of $\hat{\boldsymbol{\beta}}_{kh}(t)$. Finally, given all the time points, we select the optimal bandwidth as $\arg\min_h \sum_t \text{MSE}^h_\beta(t)$, where

$$\text{MSE}^h_\beta(t) = \sum_{k=1}^{p} \left\{ \widehat{\text{Var}}\left[\hat{\boldsymbol{\beta}}_{kh}(t)\right] + \left(\widehat{\text{Bias}}\left[\hat{\boldsymbol{\beta}}_{kh}(t)\right]\right)^2 \right\} = \sum_{k=1}^{p} \left\{ \frac{1}{4}\|\hat{\boldsymbol{\beta}}^{*1}_{kh}(t) - \hat{\boldsymbol{\beta}}^{*2}_{kh}(t)\|^2 + \|\hat{\boldsymbol{C}}_k(t)\|^2 h^4 \right\}. \tag{9}$$

We denote the optimal h_{1n} as H_1 and denote the corresponding estimators for $\boldsymbol{\beta}_k(t)$ as $\hat{\boldsymbol{\beta}}_{kH_1}(t)$. Next, given $h_{1n} = H_1$ and $\boldsymbol{\beta}_k(t) = \hat{\boldsymbol{\beta}}_{kH_1}(t)$, we select the optimal h_{2n}, the bandwidth for estimating $\sigma_{kl}(t)$'s, by minimizing the empirical mean squared error of the corresponding estimators, which is numerically calculated in the similar way to above.

3 | THEORETICAL RESULTS

We first state the following required conditions.

Condition 1. True parameters $\lambda_k^0(t)$, $\boldsymbol{\beta}_k^0(t)$, and $\sigma_{kl}^0(t)$ are continuously twice-differentiable for any $t \in [0, \tau]$, where $k, l = 1, 2, \dots, p$ and $k \neq l$. In addition, $\lambda_k^0(t)$ is strictly positive. Furthermore, second moments of $\text{Cov}(dN_{ik}(t), dN_{ik}(s)|\boldsymbol{X}_i)/dtds$ and temporal covariances $\text{Cov}(\epsilon_{ik}(t), \epsilon_{il}(s))$ are continuously twice-differentiable.

Condition 2. The vector of baseline covariate $\boldsymbol{X}$ is bounded. If there exists a vector $\boldsymbol{b}$ such that $\boldsymbol{X}^T\boldsymbol{b} = 0$, then $\boldsymbol{b} = 0$.

Condition 3. $h_{1n}, h_{2n} \to 0$ and $nh_{1n}, nh_{2n}^2 \to \infty$. Furthermore, $nh_{1n}^5, nh_{2n}^6 \to 0$.

Condition 4. The kernel function $K(z)$ is a symmetric density function satisfying $\int z^2K(z)dz < \infty$. Similarly, $\tilde{K}(z_1, z_2)$ is a symmetric bivariate density function with bounded fourth moments.

Condition 1 is used to give the asymptotic distribution for the parameter estimators in (1), and it assumes some smoothness properties of the time-varying coefficients and covariance matrices. From condition 3, the choice of h_{1n} and h_{2n} can be $n^{-1/3}$ and $n^{-1/4}$, respectively. A potential choice of the kernel satisfying condition 4 can be the Gaussian kernel or the Epanechnikov kernel. Theorem 1 states the asymptotic distribution of parameters $\hat{\boldsymbol{\beta}}_k(t)$, $k = 1, 2, \dots, p$. Theorem 2 establishes the asymptotic distribution of parameters $\hat{\sigma}_{kl}(t)$, $k, l = 1, 2, \dots, p$, and $k \neq l$.

Theorem 1 (asymptotic distribution of $\hat{\boldsymbol{\beta}}_k(t)$). *Under conditions 1 to 4, for any fixed t,*

$$(nh_{1n})^{1/2}A_k(t)\left[\hat{\boldsymbol{\beta}}_k(t) - \boldsymbol{\beta}_k^0(t)\right] \to_d \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_k(t)), \tag{10}$$

where

$$A_k(t) = \lambda_k^0(t)\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T \int \left[g_k^{-1}(\boldsymbol{X}^T\boldsymbol{\beta}_k^0(t) + \epsilon_k(t))\right]' f(\epsilon_k(t))d\epsilon_k(t)\right],$$

and the asymptotic variance

$$\boldsymbol{\Sigma}_k(t) = \lambda_k^0(t)\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\sigma^2(t, \boldsymbol{X}, \epsilon_k(t))\exp\{\boldsymbol{X}^T\gamma_k^0\}\right]\int_z K^2(z)dz,$$

where $\sigma^2(t, \boldsymbol{X}, \epsilon_k(t))$ is a function of $\epsilon_k(t)$. Its definition and the proof of theorem 1 are given in the supplementary material.

Theorem 2 (asymptotic distribution of $\hat{\sigma}_{kl}(t)$). *Under conditions 1 to 4, for any fixed t,*

$$(nh_{2n}^2)^{1/2}B_{kl}(t)\left[\hat{\sigma}_{kl}(t) - \sigma_{kl}^0(t)\right] \to_d \mathcal{N}(0, \Sigma_{kl}(t)), \tag{11}$$

where

$$B_{kl}(t) = \lambda_k^0(t)\lambda_l^0(t)\mathbb{E}\left[\iint g_k^{-1}(\boldsymbol{X}^T\boldsymbol{\beta}_k^0(t) + \epsilon_k(t))g_l^{-1}(\boldsymbol{X}^T\boldsymbol{\beta}_l^0(t) + \epsilon_l(t))\frac{\partial f(\epsilon_k(t), \epsilon_l(t); \sigma_{kl}(t))}{\partial\sigma_{kl}(t)}\Big|_{\sigma_{kl}(t)=\sigma_{kl}^0(t)} d\epsilon_k(t)d\epsilon_l(t)\right],$$

is assumed to be nonsingular, and the asymptotic variance

$$\Sigma_{kl}(t) = \lambda_k^0(t)\lambda_l^0(t)\mathbb{E}\left[\psi^2(t, t, \boldsymbol{X}, \epsilon_k(t), \epsilon_l(t))\exp\{\boldsymbol{X}^T\boldsymbol{\gamma}_k^0\}\exp\left\{\boldsymbol{X}^T\boldsymbol{\gamma}_l^0\right\}\right]\iint \tilde{K}^2(z_1, z_2)dz_1dz_2,$$

where $\psi^2(t, t, \boldsymbol{X}, \epsilon_k(t), \epsilon_l(t))$ is a function of $\epsilon_k(t)$ and $\epsilon_l(t)$. Its definition and the proof of theorem 2 are given in the supplementary material.

Since the asymptotic variances in Theorems 1 and 2 do not have simple expressions, we use the bootstrap method to estimate the asymptotic variances in practice.

4 | SIMULATION STUDIES

In the simulation studies, we simulated data of six health markers for 5000 subjects. For the ith subject, we generated two covariates $X_{i1} \sim \text{Uniform}(-1, 1)$ and $X_{i2} \sim \text{Bernoulli}(0.5) - 0.5$. Thus, $\boldsymbol{X}_i = (1, X_{i1}, X_{i2})^T$ was a three-dimensional vector of baseline variables. The maximum observation time T_i for each subject was set to 12. The measured time points for simulated markers were generated from a Poisson process whose intensity function was $\mathbb{E}\left[dN_{ik}(t)|\boldsymbol{X}_i\right] = 0.5\exp\{0.5X_{i1} + 0.25X_{i2}\}dt$. For the variances of latent processes, we assumed $c_k = 1$, $k = 1, 2, \dots, 6$. Suppose there were N_i unique measured time points $t_{i1}, t_{i2}, \dots, t_{iN_i}$ for all latent processes of the subject i, we sampled $\epsilon_i(t_{i1})$, $\epsilon_i(t_{i2})$, $\dots$, $\epsilon_i(t_{iN_i})$ from a mean-zero multivariate Gaussian distribution with a covariance matrix $\boldsymbol{\Omega}(\boldsymbol{t}_i) = \boldsymbol{\Sigma}_2(\boldsymbol{t}_i) \otimes \boldsymbol{\Sigma}_1$, where $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \dots, t_{iN_i})$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.34 & 0.48 & 0.58 & 0.03 & 0.05 \\ 0.34 & 1 & 0.80 & -0.49 & -0.78 & 0.80 \\ 0.48 & 0.80 & 1 & -0.16 & -0.36 & 0.53 \\ 0.58 & -0.49 & -0.16 & 1 & 0.80 & -0.69 \\ 0.03 & -0.78 & -0.36 & 0.80 & 1 & -0.85 \\ 0.05 & 0.80 & 0.53 & -0.69 & -0.85 & 1 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_2(\boldsymbol{t}_i) = \begin{pmatrix} 1 & e_{12} & \dots & e_{1N_i} \\ e_{21} & 1 & \dots & e_{2N_i} \\ \vdots & \vdots & \ddots & \vdots \\ e_{N_i1} & e_{N_i2} & \dots & 1 \end{pmatrix},$$

where $e_{kl} = \exp\{-(t_{ik} - t_{il})^2\}$, $k, l = 1, 2, \dots, N_i$. Thus, at each measured time point, $\boldsymbol{\Omega}(t)$ is constant and equals to $\boldsymbol{\Sigma}_1$, but there exist underlying dependencies in the time intervals between these time points.

The values of simulated markers were generated according to (2) and (3). To assess the ability of our models in Section 2.1 to handle mixed-type markers, we assumed $Y_{i1}(t)$ and $Y_{i4}(t)$ were Gaussian distributed. $Y_{i2}(t)$ was Poisson distributed. $Y_{i3}(t)$, $Y_{i5}(t)$, and $Y_{i6}(t)$ were Bernoulli distributed. Thus, $g_1^{-1}(z) = g_4^{-1}(z) = z$, $g_2^{-1}(z) = e^z$, and $g_3^{-1}(z) = g_5^{-1}(z) = g_6^{-1}(z) = e^z/(1 + e^z)$. Furthermore, since the distributions of $Y_{i1}(t)$ and $Y_{i4}(t)$ have dispersion parameters, we set $\phi_{i1}(t) = \phi_{i4}(t) = 0.5$. The true values of $\boldsymbol{\beta}_k(t)$ were assumed to be

$$\begin{pmatrix} \boldsymbol{\beta}_1^T(t) \\ \boldsymbol{\beta}_2^T(t) \\ \boldsymbol{\beta}_3^T(t) \\ \boldsymbol{\beta}_4^T(t) \\ \boldsymbol{\beta}_5^T(t) \\ \boldsymbol{\beta}_6^T(t) \end{pmatrix} = \begin{pmatrix} -0.44 - \frac{t}{8} & 0.6 + \frac{\sqrt{t}}{3} & -0.5 + \frac{\sqrt[3]{t}}{2} \\ -0.93 + \frac{t}{9} & -0.53 - \frac{\sqrt{t}}{2} & 0.4 + \frac{\sqrt[3]{t}}{2} \\ 0.35 + \frac{t}{10} & -2 + \sqrt{t} & 1.9 - \sqrt[3]{t} \\ -1.36 + \frac{t}{10} & \sin(0.76 + t) & \cos(-0.3 + t) \\ \cos(-0.25 + t) & 0.37 + \frac{t}{10} & \sin(-0.68 + t) \\ 0.91 + \frac{(t-6)^3}{216} & \frac{t}{10} & 1.23 + \frac{(t-6)^2}{36} \end{pmatrix}.$$

The scaled Epanechnikov kernel was chosen as the kernel function in (5), that is,

$$K_{h_{1n}}(z) = \frac{3}{4h_{1n}}\left[1 - \left(\frac{z}{h_{1n}}\right)^2\right]_+. \tag{12}$$

Furthermore, the kernel function in (6) was set to the product of two scaled univariate Epanechnikov kernels, that is,

$$\tilde{K}_{h_{2n}}(z_1, z_2) = \frac{9}{16h_{2n}^2}\left[1 - \left(\frac{z_1}{h_{2n}}\right)^2\right]_+\left[1 - \left(\frac{z_2}{h_{2n}}\right)^2\right]_+. \tag{13}$$

Since the data-adaptive method for selecting bandwidths was computationally intensive, we first conducted a preliminary study on the simulated data. We used the method in Section 2.4 and selected the optimal bandwidths among $h = cn^{-1/z}$, where $n = 5000$, $c = \{5, 10, 20, 30\}$, and $z = 1, 2, \dots, 10$. Hence, the potential bandwidths ranged from 0.001 to

TABLE 1 Summary statistics for $\boldsymbol{\beta}_k(t)$ at $t = 1$ based on 100 simulations

Marker	Parameter	True value	Bias	SD	SE	CP
Y_1	β_{10}	−0.565	0.002	0.035	0.039	0.98
Continuous	β_{11}	0.933	0.001	0.059	0.067	0.98
	β_{12}	0.000	−0.002	0.085	0.078	0.94
Y_2	β_{20}	−0.819	0.007	0.050	0.058	0.98
Count	β_{21}	−1.030	0.026	0.112	0.112	0.95
	β_{22}	0.900	−0.010	0.117	0.132	0.97
Y_3	β_{30}	0.450	−0.006	0.074	0.077	0.94
Binary	β_{31}	−1.000	0.013	0.112	0.136	0.99
	β_{32}	0.900	0.011	0.157	0.151	0.93
Y_4	β_{40}	−1.260	−0.006	0.038	0.039	0.93
Continuous	β_{41}	0.982	−0.010	0.063	0.068	0.97
	β_{42}	0.765	−0.005	0.078	0.077	0.93
Y_5	β_{50}	0.732	0.001	0.077	0.074	0.95
Binary	β_{51}	0.470	0.001	0.149	0.134	0.92
	β_{52}	0.315	−0.014	0.163	0.150	0.92
Y_6	β_{60}	0.331	−0.018	0.085	0.077	0.90
Binary	β_{61}	0.100	0.004	0.144	0.136	0.95
	β_{62}	1.924	−0.021	0.163	0.156	0.94

Note: "Bias" is the bias of the average estimates; "SD" is the sample standard deviation of the estimates; "SE" is the average of the estimated standard errors based on 100 bootstrap samples; "CP" is the coverage probability of the 95% confidence intervals.

12.800. We found $h_{1n} = 5n^{-1/3} = 0.292$ and $h_{2n} = 10n^{-1/3} = 0.585$ were close to the optimal. This set of h_{1n} and h_{2n} was used in all subsequent simulations.

For time points $t = 0, 1, \ldots, 12$, we solved (5) and (6), and we obtained $\hat{\boldsymbol{\beta}}_k(t)$ and $\hat{\sigma}_{kl}(t)$. We evaluated accuracies of the asymptotic approximations by calculating the average bias and the sample standard deviation of $\hat{\boldsymbol{\beta}}_k(t)$ and $\hat{\sigma}_{kl}(t)$, respectively. In addition, using the bootstrap method, we calculated the bootstrap estimators for standard errors of $\hat{\boldsymbol{\beta}}_k(t)$ and $\hat{\sigma}_{kl}(t)$. Specifically, for each dataset, we resampled 5000 observations with replacement from $\boldsymbol{X}$ to produce a bootstrap dataset $\boldsymbol{X}^{*1}$. We could use $\boldsymbol{X}^{*1}$ to produce a new bootstrap estimator for $\boldsymbol{\beta}_k(t)$, which we called $\hat{\boldsymbol{\beta}}_k^{*1}(t)$. This procedure was repeated B times in order to produce B different bootstrap datasets, $\boldsymbol{X}^{*1}, \boldsymbol{X}^{*2}, \ldots, \boldsymbol{X}^{*B}$, and B corresponding $\boldsymbol{\beta}_k(t)$ estimators, $\hat{\boldsymbol{\beta}}_k^{*1}(t), \hat{\boldsymbol{\beta}}_k^{*2}(t), \ldots, \hat{\boldsymbol{\beta}}_k^{*B}(t)$. Next we computed the sample variance of these bootstrap estimators and treated it as the estimated variance. Similar procedures were also applicable to $\hat{\sigma}_{kl}(t)$. Afterward, 95% confidence intervals of each parameter were constructed. Finally, we counted how many times true parameters $\boldsymbol{\beta}_k(t)$ and $\sigma_{kl}(t)$ fell in their confidence intervals to obtain coverage probabilities.

Tables 1 and 2 summarize the main results over 100 simulations at $t = 1$. From Tables 1 and 2, we can conclude that, at $t = 1$, our method yields estimators $\hat{\boldsymbol{\beta}}_k(t)$ which are close to the true parameters. All the estimators deviate from true parameters by less than 0.03. On the other hand, the absolute values of biases between estimators $\hat{\sigma}_{kl}(t)$ and true parameters become a little greater, but most of them are still less than 0.1. In addition, the bootstrap based standard errors are reasonable estimators for the standard deviations of $\hat{\boldsymbol{\beta}}_k(t)$ and $\hat{\sigma}_{kl}(t)$. Almost all the differences between SD and SE are smaller than 0.03, except for $\hat{\sigma}_{34}(t)$. Also, excluding $\hat{\sigma}_{13}(t)$, all the coverage probabilities are greater than or equal to 0.9, and the majority of them are around 0.95.

After examining the estimators at a fixed time point, we also investigated the estimation performance as time changes. For instance, Figure 2 presents true parameters vs estimators across the 13 time points for $\beta_{52}(t)$ and $\sigma_{34}(t)$, respectively. From Figure 2, we can conclude $\hat{\beta}_{52}(t)$ is very close to the true parameter at each time point, and it well captures the underlying smooth function of $\beta_{52}(t)$ across time. Although the bias between $\sigma_{34}(t)$ and $\hat{\sigma}_{34}(t)$ is greater than that between

TABLE 2 Summary statistics for $\sigma_{kl}(t)$ at $t = 1$ based on 100 simulations

Parameter	True value	Bias	SD	SE	CP
σ_{12}	0.342	−0.061	0.149	0.136	0.91
σ_{13}	0.484	−0.058	0.202	0.213	0.98
σ_{14}	0.578	−0.086	0.127	0.121	0.87
σ_{15}	0.034	0.030	0.216	0.218	0.95
σ_{16}	0.047	−0.009	0.210	0.207	0.96
σ_{23}	0.799	−0.150	0.388	0.382	0.90
σ_{24}	−0.493	0.065	0.232	0.233	0.95
σ_{25}	−0.779	0.078	0.241	0.242	0.94
σ_{26}	0.796	−0.143	0.371	0.366	0.91
σ_{34}	−0.163	0.048	0.216	0.252	0.97
σ_{35}	−0.363	−0.024	0.252	0.261	0.97
σ_{36}	0.530	−0.024	0.257	0.249	0.95
σ_{45}	0.802	−0.076	0.212	0.219	0.94
σ_{46}	−0.686	0.089	0.228	0.244	0.94
σ_{56}	−0.846	−0.019	0.160	0.181	0.97

Note: "Bias" is the bias of the average estimates; "SD" is the sample standard deviation of the estimates; "SE" is the average of the estimated standard errors based on 100 bootstrap samples; "CP" is the coverage probability of the 95% confidence intervals.

$\beta_{52}(t)$ and $\hat{\beta}_{52}(t)$, all of $\sigma_{34}(t)$ are in the interquartile range of $\hat{\sigma}_{34}(t)$. Thus, the estimators perform consistently and the deviations are reasonable.

5 | REAL DATA APPLICATION

5.1 | Data prepocessing

We applied the proposed method to analyze EHRs of T2D patients from the OSU-WMCIW. In our application, we included three baseline variables $\boldsymbol{X}_i$: baseline age, race (1: white; 0: nonwhite), and sex (1: male; 0: female). Besides, there were five health markers $Y_{ik}(t)$ related to T2D: hypertension/high blood pressure (HBP), total cholesterol (TC), glycated hemoglobin (HbA1c), high-density lipoprotein (HDL), and medications prescribed at each clinical encounter. Here, we dichotomized HBP as HBP=1 if a patient's systolic blood pressure is higher than 140 mm Hg and 0, otherwise. The medications served as one strong indicator of patient's comorbidity and they could be T2D related or not. Thus, the health markers in the analysis consisted of three continuous markers (TC, HbA1c, HDL), one binary marker (HBP), and one count marker (number of medications).

For analysis, we split the data into three parts for different purposes. The first data consisted of the records collected between 2011 and 2012 and was used to estimate the variances of individual latent processes by fitting univariate generalized linear mixed models. The second part included the records from 24 975 patients between 2013 and 2017 who had at least one marker measurement. This part of the data was used for training our models and learning latent groups among the patients. The third part was the data collected in 2018 and would be used for validation purpose. The flow-chart for this work is illustrated in Figure 3.

In our model fitting using the second part of the data, after checking normal ranges for the health markers,[19-21] we removed extreme records such as TC ≤ 0 or ≥ 500 mg/dL, HbA1c ≤ 3 or $\geq 20\%$, and HDL ≤ 0 or ≥ 120 mg/dL. This led to a deletion of 1% of the data and a total number of 24 655 patients for analysis. Among these patients, 52.08% were female, 63.42% were white, and their ages in years ranged from 18.30 to 97.67 with a mean of 56.06. All of them had at least one observation for at least one health marker in the 5 years, but not necessarily for other health markers. Specifically, the average numbers of records for HBP, TC, HbA1c, HDL, and the number of medications per patient during these 5 years

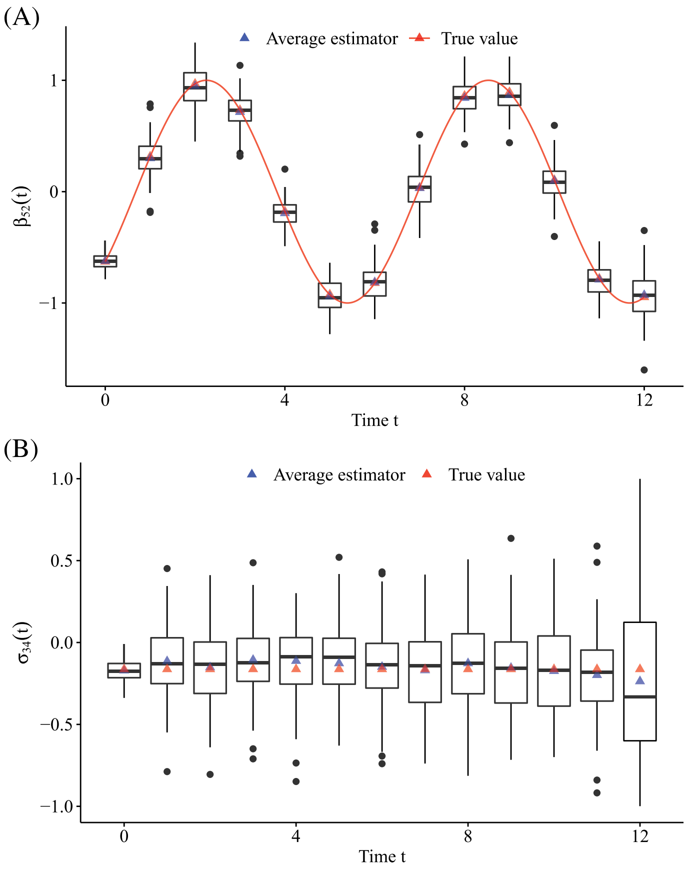


FIGURE 2 Top panel: true $\beta_{52}(t)$ vs $\hat{\beta}_{52}(t)$ across 13 time points based on 100 simulations. Bottom panel: true $\sigma_{34}(t)$ vs $\hat{\sigma}_{34}(t)$ across 13 time points based on 100 simulations. Red triangles: true values of the parameter. Blue triangles: average estimators of the parameter. Red curve: the true function of the parameter [Colour figure can be viewed at wileyonlinelibrary.com]

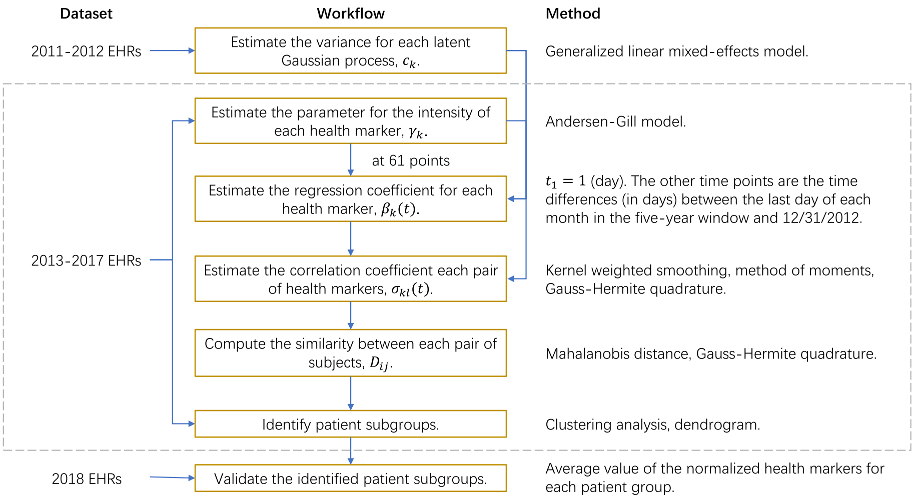

Dataset
Workflow
Method
2011-2012 EHRs
Estimate the variance for each latent Gaussian process, c_k.
Generalized linear mixed-effects model.
Estimate the parameter for the intensity of each health marker, γ_k.
Andersen-Gill model.
at 61 points
Estimate the regression coefficient for each health marker, $\beta_k(t)$.
$t_1 = 1$ (day). The other time points are the time differences (in days) between the last day of each month in the five-year window and 12/31/2012.
2013-2017 EHRs
Estimate the correlation coefficient each pair of health markers, $\sigma_{kl}(t)$.
Kernel weighted smoothing, method of moments, Gauss-Hermite quadrature.
Compute the similarity between each pair of subjects, D_{ij}.
Mahalanobis distance, Gauss-Hermite quadrature.
Identify patient subgroups.
Clustering analysis, dendrogram.
2018 EHRs
Validate the identified patient subgroups.
Average value of the normalized health markers for each patient group.

FIGURE 3 Flow-chart of the proposed analysis framework of EHRs to dissect patient heterogeneity using a diverse set of health markers. [Colour figure can be viewed at wileyonlinelibrary.com]

were 17.50, 4.01, 5.95, 3.64, and 53.21, respectively. In order to minimize the influence of different scales on the numeric stability, we normalized all continuous variables before identifying patient subgroups. Each of them has zero mean and unit variance.

5.2 | Results

Table 3 shows the effect of each demographic variable on the pattern of the measurement times for each marker. From Table 3, we conclude that elder patients tend to have more observations for all health markers and females appeared to have more observations for HBP, HbA1c, and the number of medications, while males tend to have more TC

TABLE 3 Effects of demographic variables on the frequency of health marker measurements

Marker	Demographic	Est	HR	SE	Z	*P*-value
HBP	Age	0.065	1.067	0.006	10.552	<.001
	Sex	0.064	1.066	0.013	4.813	<.001
	Race	−0.129	0.879	0.014	−9.425	<.001
TC	Age	0.035	1.035	0.006	6.238	<.001
	Sex	−0.035	0.965	0.012	−3.000	.003
	Race	−0.012	0.988	0.013	−0.968	.333
HbA1c	Age	0.008	1.008	0.005	1.721	.085
	Sex	0.034	1.034	0.009	3.678	<.001
	Race	−0.044	0.957	0.009	−4.650	<.001
HDL	Age	0.047	1.048	0.005	10.090	<.001
	Sex	−0.010	0.990	0.010	−1.007	.314
	Race	−0.007	0.993	0.010	−0.715	.475
Medications	Age	0.042	1.043	0.006	7.262	<.001
	Sex	0.086	1.090	0.012	7.069	<.001
	Race	−0.113	0.893	0.013	−8.988	<.001

Note: "Est" is the regression coefficient estimator; "HR" is the hazard ratio; "SE" is the standard error of the coefficient estimator; "Z" is the statistic for a z-test; "*P*-value" is the *P*-value for the z-test.

measurements. Finally, whites have significantly less observations for HBP, HbA1c, and the number of medications than nonwhites.

To estimate the parameters in the joint models, we first implemented the adaptive method of bandwidth selection as stated in Section 2.4, and results are shown in the supplementary Figure 1. We ended up to choose $h_{1n} = 564.112$ days and $h_{2n} = 494.687$ days as the optimal bandwidths. Using the optimal bandwidths, we estimated $\boldsymbol{\beta}_k(t)$ and $\sigma_{kl}(t)$ at 61 time points. The results are presented in Figures 4 and 5, respectively. The salmon-colored ribbons in these two figures are 95% confidence intervals for the parameters based on 100 bootstrap datasets.

Figure 4 presents the relationships between each pair of health markers and covariates. In general, all health markers exhibit changes over time. Mean HbA1c ($\hat{\beta}_{30}(t)$) decreases during the first 1.5 years and has an increasing trend afterward, which may suggest the difficulty to achieve long-term control of glycemic levels in a chronically ill patient population. Mean HDL ($\hat{\beta}_{40}(t)$) shows a similar quadratic pattern over time, suggesting difficulty of long-term cholesterol control. The estimated regression coefficients for covariates, that is, the estimated effects of covariates on health markers, do not show any pattern of drastic changes over time. Instead, the estimated values across time fluctuate around mean values. However, we can observe decreasing trends for $\hat{\beta}_{20}(t)$ and $\hat{\beta}_{50}(t)$, suggesting that as time increases, the expected means of cholesterol and the number of medications decrease. $\hat{\beta}_{11}(t)$ and $\hat{\beta}_{41}(t)$ are positive across time, while $\hat{\beta}_{21}(t)$ and $\hat{\beta}_{31}(t)$ are negative. $\hat{\beta}_{51}(t)$ is negative but close to 0. Hence, estimators $\hat{\beta}_{\cdot 1}(t)$ suggest that elder subjects on average have higher HBP and HDL, but they have lower cholesterol and HbA1c. There is no apparent difference in the average number of medications between elder subjects and younger subjects. Similarly, estimators of sex effect, $\hat{\beta}_{\cdot 2}(t)$, suggest that compared with men, women tend to have higher expected means of cholesterol and HDL, but they have lower values of HBP and the number of medications. Although women have slightly lower expected means of HbA1c than men, the difference is inapparent. For race, the estimators of $\hat{\beta}_{\cdot 3}(t)$ indicate that white people have lower or equal expected means than nonwhite people in almost all five health markers.

Figure 5 presents the correlations between each pair of health markers. The results suggest the concurrent correlations between HBP and cholesterol, HBP and medications, cholesterol and HbA1c, cholesterol and HDL are positive and moderate. Moreover, there exists negative and observable concurrent correlations between HbA1c and HDL, HDL and medications. The correlation between HbA1c and HDL decreases as time increases. On the opposite, the positive correlation between cholesterol and HDL decreases at the beginning, but increases after about 1 year. The positive correlation between cholesterol and HbA1c has a similar pattern as it decreases at first and increases after 1000 days. The correlations of HBP and cholesterol, HbA1c and number of medications increase in first 500 days, but they start to decrease during

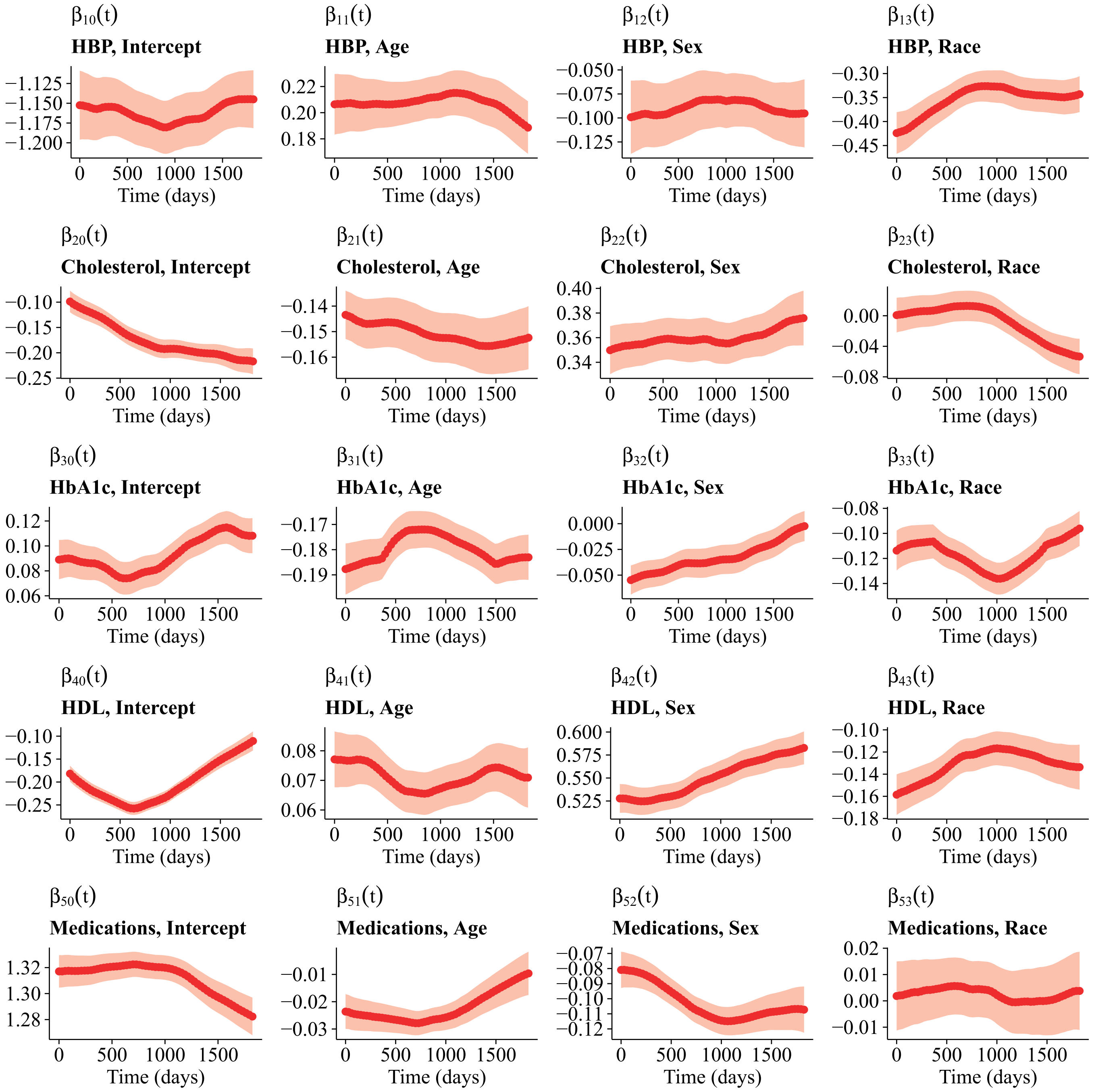


FIGURE 4 Estimated regression coefficients $\hat{\beta}_k(t)$ across 61 time points with $h_{1n} = 564.112$ days and $h_{2n} = 494.687$ days from EHRs at the Ohio State University Wexner Medical Center. Salmon-colored ribbons: 95% confidence intervals for the estimators [Colour figure can be viewed at wileyonlinelibrary.com]

500 to 1000 days, and bounce back afterward. The correlations of HBP and HbA1c, HBP and HDL, HDL, and number of medications decrease in first 500 days, and then increase, but decrease again after 1000 days.

One interesting observation from Figure 5 is that the estimated correlation between the number of medications and HBP is as high as 0.6 but its correlations with cholesterol and HDL are both negative, fluctuating around −0.30. However, there does not appear to be a strong association between the number of medications and HbAc1 over time. This may suggest that the patients in this cohort were most likely to take medications that aimed to control the levels of cholesterol and HDL, but not necessarily for controlling the level of HbA1c. The latter is consistent with the fact that over 90% of drugs recorded this database are nondiabetic drugs. One possible interpretation of the observed time-dependent correlation

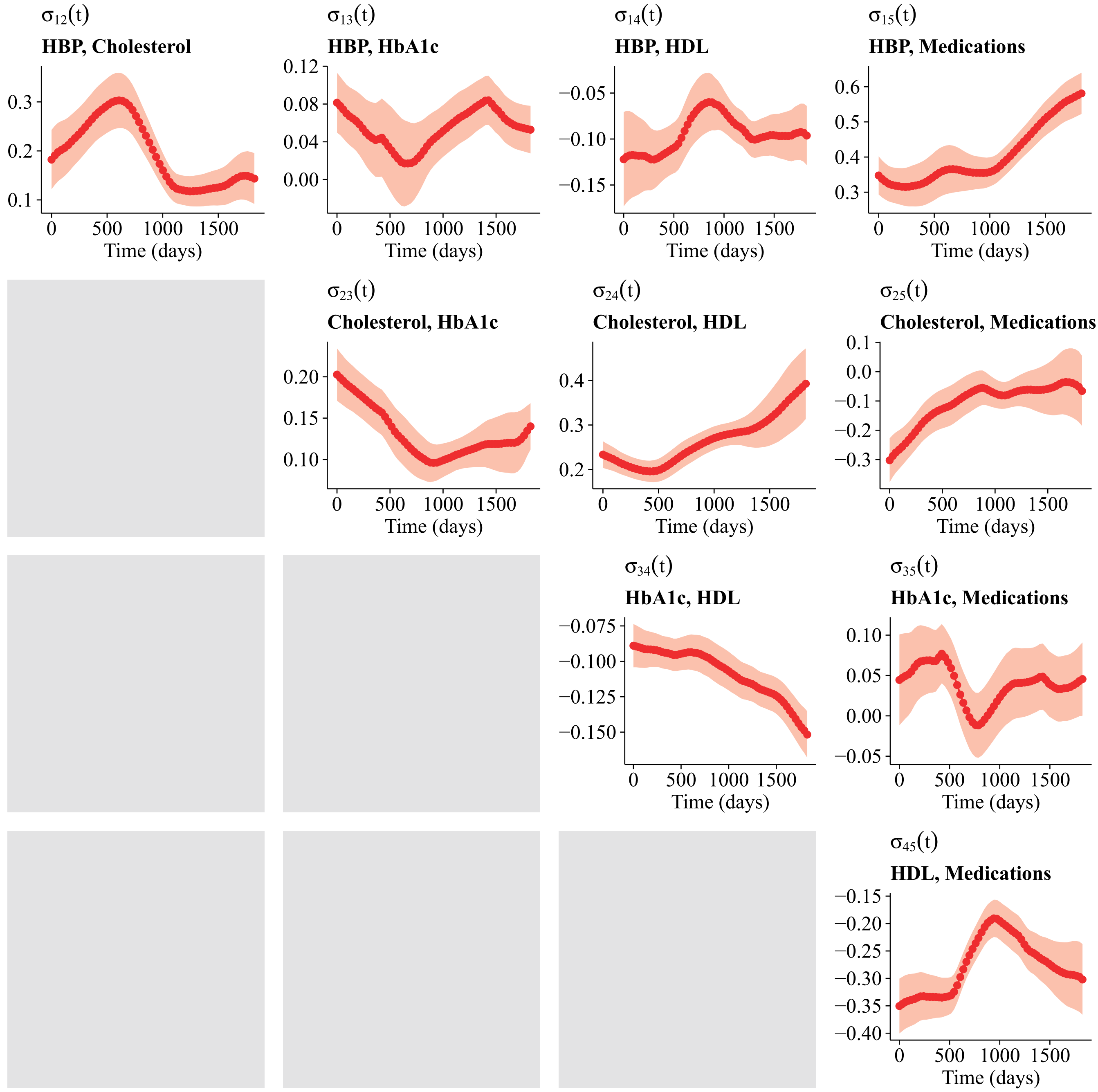


FIGURE 5 Estimated correlations $\hat{\sigma}_{kl}(t)$ across 61 time points using $h_{1n} = 564.112$ days and $h_{2n} = 494.687$ days from EHRs at the Ohio State University Wexner Medical Center. Salmon-colored ribbons: 95% confidence intervals for the estimators [Colour figure can be viewed at wileyonlinelibrary.com]

pattern is that there might exists another unobserved disease health marker that influences the two observed markers temporally. Thus, the estimated correlation pattern could be potentially useful to identify such "common cause" health marker so as to better understand the mechanism of disease progression.

Finally, we computed the similarity between each pair of patients using the distance defined in (7). To compute $\hat{\epsilon}_i(t)$ as (8), we substituted $\hat{Y}_i(t)$ with the nearest neighbor observation of time t for patient i. Using the between-patient similarity matrix, we performed a cluster analysis on the 24 655 patients, and the results are given in Figure 6. We observe four clusters within which patients had similar health marker profiles.

To better understand the health patterns of patients in each subgroup, we calculated the average of normalized measurements for each health marker in each group, as shown in Figure 7. In the top panel of Figure 7, the value in each

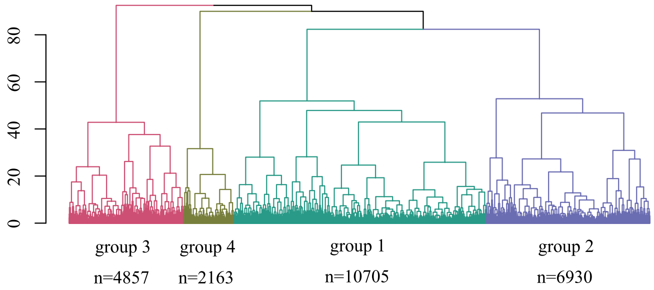
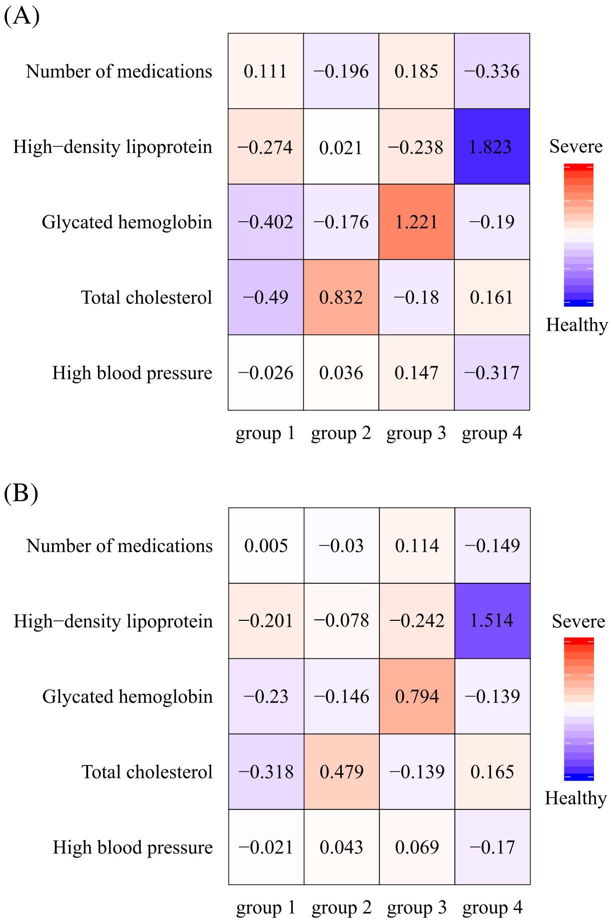


FIGURE 6 Dendrogram of Mahalanobis distances for 24 655 patients at the Ohio State University Wexner Medical Center. Group index numbers are assigned according to group sizes [Colour figure can be viewed at wileyonlinelibrary.com]



(A)

	group 1	group 2	group 3	group 4
Number of medications	0.111	−0.196	0.185	−0.336
High−density lipoprotein	−0.274	0.021	−0.238	1.823
Glycated hemoglobin	−0.402	−0.176	1.221	−0.19
Total cholesterol	−0.49	0.832	−0.18	0.161
High blood pressure	−0.026	0.036	0.147	−0.317

(B)

	group 1	group 2	group 3	group 4
Number of medications	0.005	−0.03	0.114	−0.149
High−density lipoprotein	−0.201	−0.078	−0.242	1.514
Glycated hemoglobin	−0.23	−0.146	0.794	−0.139
Total cholesterol	−0.318	0.479	−0.139	0.165
High blood pressure	−0.021	0.043	0.069	−0.17

FIGURE 7 Averages of normalized measurements by health markers and patient subgroups. A, Using data from 1/1/2013 to 12/31/2017. B, Using data after 1/1/2018. Red: more severe status than the overall sample average in terms of a health marker; blue: healthier status than the overall sample average in terms of a health marker; white: overall sample average status in terms of a health marker [Colour figure can be viewed at wileyonlinelibrary.com]

cell is averaged over all patients and all clinical encounters between 1 January 2013 and 31 December 2017. We compare these values to the average of each health marker in the entire study sample. A higher value of HDL and a lower value of HBP, cholesterol, and HbA1c represent healthier T2D status. The number of medications prescribed at each clinical encounter does not directly reflect the disease status, but a lower count usually indicates a less severe state. Group 4 contains 2163 patients, whose cholesterol was slightly higher than the overall average. Their HBP, HDL, and the number of medications were lower than the overall averages. In addition, they had the highest HDL and it was substantially higher than the overall average. Thus, group 4 is the relatively healthy group in which patients did not take many medications. Group 1 contains 10 705 patients who were less healthy since they had lower-than-average HDL, but other health markers were favorable or roughly neutral. The cholesterol of 6930 patients in group 2 was higher than the overall average, while other health markers were lower or around the averages. We conclude that group 2 is a moderately ill group. For the 4857 patients in group 3, their cholesterol levels were slightly lower than the overall average; however, they had the highest HbA1c. Also, other markers indicated bad health status. Therefore, group 3 patients were in the most severe state of T2D.

To examine whether the subgroups inferred by the clustering truly represent patients health profiles, we validated the detected patterns using the third fold of the split data that consisted of the EHR data collected after 1 January 2018. These data were not used in any other analyses of this application. The average values of normalized measurements for each health marker in each group are shown in the bottom panel of Figure 7. We conclude that the patients health patterns identified prior to year 2018 are consistent with those patterns afterward. Therefore, the patient groups are not only meaningful, but also represent some true underlying patient patterns over time. This robustness is particularly important to the long-term health management of T2D patients.

6 | DISCUSSION

In this work, we proposed a latent temporal process model to integrate health markers in EHRs and characterize patient heterogeneities. The proposed method is capable of handling unbalanced records and informative visits, that is, patients can have missing health markers at some encounters or with visit times depending on their health status. Additionally, our model can both fit different types of health marker, capture the dependence structures among health markers, and takes into account informative patterns of visit times, via the intensity function of health markers. The real data application shows the capability of the proposed method on addressing the data challenges of EHRs, integrating different types of health markers, and identifying meaningful and robust patient subgroups. Therefore, the proposed method may shed lights on the detection of patient homogeneities and heterogeneities, and serve as a step toward applications of personalized medicine.

In the parameter estimation process, we assumed that variances of the latent variables $\boldsymbol{\epsilon}_i(t)$ were fixed and they were estimated using the EHR data of 2011 and 2012. To study whether the constant variance was reasonable, we estimated the changes in variances from six different time periods in windows of 2 years as well as using the whole 5-year data, and the results, as shown in supplementary Tables 1 and 2, indicate that the estimates varied little. Thus, the constant variance assumption seems to be reasonable for our application. In addition, we reestimated $\boldsymbol{\beta}_k(t)$ and $\sigma_{kl}(t)$ using the 5-year variance estimates in our approach. Supplementary Figures 2 and 3 reveal slight changes in the estimated coefficients. In fact, the absolute percentage changes between the two sets of coefficients are less than 1%, except for $\tilde{\beta}_{1\cdot}(t)$ and $\tilde{\sigma}_{1\cdot}(t)$ which have changes of up to 3%. Therefore, we could conclude that the estimation results are robust to the constant variance estimates. Moreover, to investigate the effect of the bandwidth selection as suggested by a reviewer, we report $\boldsymbol{\beta}_k(t)$ using two suboptimal bandwidths close to the optimal bandwidth in the article. Supplementary Figure 4 shows that the suboptimal estimators preserve the similar pattern to $\hat{\boldsymbol{\beta}}_k(t)$. The Canberra distance[22] between the optimal estimators and suboptimal estimators of $\beta_{kj}(t)$, $k = 1, \ldots, p, j = 0, \ldots, m$, across time, calculated as

$$d(\boldsymbol{\beta}_{kj,H_1}, \boldsymbol{\beta}_{kj,H_1'}) = \frac{1}{N}\sum_{t=t_1}^{t_N} \frac{|\boldsymbol{\beta}_{kj,H_1}(t) - \boldsymbol{\beta}_{kj,H_1'}(t)|}{|\boldsymbol{\beta}_{kj,H_1}(t)| + |\boldsymbol{\beta}_{kj,H_1'}(t)|}, \tag{14}$$

where $\boldsymbol{\beta}_{kj,H_1}$ is the vector of estimated $\{\beta_{kj}(t) : t = t_1, \ldots, t_N\}$ using the optimal bandwidth H_1 and $\boldsymbol{\beta}_{kj,H_1'}$ is the vector of these estimators using a suboptimal bandwidth H_1', as all smaller than 0.05 as given in supplementary Table 3. The conclusions could be drawn for estimating $\sigma_{kl}(t)$ (cf, supplementary Figure 5 and supplementary Table 4).

In our models, we assumed that the intensity function of the counting process only depended on the baseline covariates. This assumption can be violated if the intensity also depends on the historical marker values. Since incorporating time-dependent marker values, which are missing for most of time points, is challenging, to examine how this assumption may affect our results, we included an ad hoc marker value, defined as the mean value of HbA1c in the past 12 months, in the intensity model (1). From supplementary Table 5, the effects of the historical HbA1c level on frequencies of HBP, TC, and HbA1c are significant, while the historical HbA1c level has lower impacts on frequencies of HDL and the number of medications. Supplementary Figures 6 and 7 also reflect this phenomenon that there are slight differences between two versions of estimators for HDL and the number of medications. Although differences between two versions of estimators for HBP and TC are moderate, the new estimators still locate within or around the 95% bootstrapped confidence intervals for the original estimators. However, for HbA1c, the differences could not be ignored but the estimated curves present some unusual shapes. Therefore, further investigation is needed regarding what time-dependent marker values should be used and how missing data issues should be addressed. We will investigate it in our future work.

Since the estimation of both regression coefficients and correlations among latent processes only relies on one or two health markers, our method can be easily extended to handle a large number of health markers, where the computation can be easily parallelized to save computing time and cost. Inferences can be made based on subsampling subsets of the data. Some other extensions to the proposed method include to estimate all β's simultaneously by incorporating the covariance matrix to the estimating equations for β's, or to allow marker-specific and time-sensitive bandwidth selection during the parameter estimation (especially when smoothness of biomarker trajectories are expected to be substantially different). Another possible extension is to explicitly model the temporal dependence within the same health marker, as well as across health markers. Although with increased computational burden, an advantage of this extension is the potential to obtain a more precise assessment of the latent process given the entire history of health markers.

As stated in Section 1, the latent processes can be also viewed as projections of the health markers onto a lower dimensional space. Therefore, our method can be used for identifying latent clusters among patients as illustrated in our application, and at the same time can also play a role in learning personalized disease prognosis and personalized disease management. For example, the summary of latent processes can be used to improve the understanding of treatment propensity scores in EHRs when learning individualized treatment rules. Lastly, the latent processes can be included in disease outcome models as prognostic or predictive health markers.

ACKNOWLEDGEMENT

This research is supported by US NIH grants GM124104, NS073671, and MH117458. The codes "EHR Latent" to implement these methods are available from https://blogs.cuit.columbia.edu/yw2016/software/.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY

The codes "EHR Latent" to implement these methods are available from https://blogs.cuit.columbia.edu/yw2016/software/.

ORCID

Jitong Lou https://orcid.org/0000-0002-1049-0416
Yuanjia Wang https://orcid.org/0000-0002-1510-3315

REFERENCES

1. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J Med Internet Res*. 2005;7(1):e3.
2. Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic health records and quality of diabetes care. *N Engl J Med*. 2011;365(9): 825-833.
3. Herrin J, Graca B, Nicewander D, et al. The effectiveness of implementing an electronic health record on diabetes care and outcomes. *Health Serv Res*. 2012;47(4):1522-1540.
4. Verbeke G, Fieuws S, Molenberghs G, Davidian M. The analysis of multivariate longitudinal data: a review. *Stat Methods Med Res*. 2014;23(1):42-59.
5. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data: an overview and update. *J Agric Biol Envir Stat*. 2003;8(4):387-419.
6. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York, NY: Springer; 2000.
7. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. New York, NY: Springer; 2005.
8. Lambert P, Vandenhende F. A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Stat Med*. 2002;21(21):3197-3217.
9. Gueorguieva RV, Sanacora G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat Med*. 2006;25(8):1307-1322.
10. Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*. 2002;89(1):111-128.
11. Fan J, Zhang W. Statistical methods with varying coefficient models. *Stat Interface*. 2008;1(1):179-195.
12. Henao R, Lu JT, Lucas JE, Ferranti J, Carin L. Electronic health record analysis via deep Poisson factor models. *J Mach Learn Res*. 2016;17(1):6422-6453.
13. Ho JC, Ghosh J, Steinhubl SR, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. *J Biomed Inform*. 2014;52:199-211.
14. Miscouridou X, Perotte A, Elhadad N, Ranganath R. Deep survival analysis: non-parametrics and missingness. Paper presented at: Proceedings of the 3rd Machine Learning for Healthcare Conference. Palo Alto, California: PMLR; 2018;85:244-256.
15. DeMaesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. *Chemom Intell Lab Syst*. 2000;50:1-18.
16. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982;10(4):1100-1120.

17. Abramowitz M. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, DC: US Government printing office; 1964.
18. Cao H, Zeng D, Fine JP. Regression analysis of sparse asynchronous longitudinal data. *J R Stat Soc Ser B Stat Methodol*. 2015;77(4):755-776.
19. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American college of cardiology/American heart association task force on practice guidelines. *J Am Coll Cardiol*. 2014;63(25 Pt B):2889-2934.
20. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American college of cardiology/American heart association task force on clinical practice guidelines. *J Am Coll Cardiol*. 2018;71(19):e127-e248.
21. American Diabetes Association. 8 Pharmacologic approaches to glycemic treatment: standards of medical care in diabetes-2018. *Diabetes Care*. 2018;41(Suppl 1):S73-S85.
22. Lance GN, Williams WT. Computer programs for hierarchical polythetic classification ("similarity analyses"). *Comput J*. 1966;9(1):60-64.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lou J, Wang Y, Li L, Zeng D. Learning latent heterogeneity for type 2 diabetes patients using longitudinal health markers in electronic health records. *Statistics in Medicine*. 2021;40:1930–1946. https://doi.org/10.1002/sim.8880