Submitted to the Annals of Applied Statistics

A NONPARAMETRIC MIXED-EFFECTS MIXTURE MODEL FOR PATTERNS OF CLINICAL MEASUREMENTS ASSOCIATED WITH COVID-19

BY XIAORAN MA[1,a], WENSHENG GUO[2,d], MENGYANG GU[1,b], LEN USVYAT[3,e], PETER KOTANKO[4,f], AND YUEDONG WANG[1,c]

[1]*Department of Statistics and Applied Probability, University of California, Santa Barbara,* [a]*xiaoran_ma@pstat.ucsb.edu;* [b]*mengyang@pstat.ucsb.edu;* [c]*yuedong@pstat.ucsb.edu*

[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [d]*wguo@upenn.edu*

[3]*Fresenius Medical Care,* [e]*Len.Usvyat@freseniusmedicalcare.com*

[4]*Renal Research Institute, New York, NY,* [f]*Peter.Kotanko@rriny.com*

[4]*Icahn School of Medicine at Mount Sinai, New York, NY,* [f]*Peter.Kotanko@rriny.com*

Some patients with COVID-19 show changes in signs and symptoms such as temperature and oxygen saturation days before being positively tested for SARS-CoV-2, while others remain asymptomatic. It is important to identify these subgroups and to understand what biological and clinical predictors are related to these subgroups. This information will provide insights into how the immune system may respond differently to infection and can further be used to identify infected individuals. We propose a flexible nonparametric mixed-effects mixture model that identifies risk factors and classifies patients with biological changes. We model the latent probability of biological changes using a logistic regression model and trajectories in the latent groups using smoothing splines. We developed an EM algorithm to maximize the penalized likelihood for estimating all parameters and mean functions. We evaluate our methods by simulations and apply the proposed model to investigate changes in temperature in a cohort of COVID-19-infected hemodialysis patients.

1. Introduction. The Coronavirus Disease 2019 (COVID-19) pandemic has had a profound impact on humanity, and the emergence of new variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) challenges healthcare systems worldwide. Much research has examined changes in biological variables for diagnosis and prognosis during and after an infection caused by the SARS-CoV-2 (Pimentel et al., 2020; Malik et al., 2021; Chaudhuri et al., 2022). Early detection of SARS-CoV-2 infection based on changes in readily available measurements such as temperature and arterial oxygen saturation during the incubation period is crucial for isolating and treating contagious individuals (de Moraes Batista et al., 2020; Wu et al., 2020; Kukar et al., 2021; Monaghan et al., 2021). Indicators of disease severity and prognosis are essential to the clinical management of COVID-19 patients (Jiang et al., 2020; Malik et al., 2021; Gallo Marin et al., 2021). Identifying potential changes in an individual is challenging because the clinical presentation of COVID-19 varies greatly from asymptomatic infection to critical illness (Harahwa et al., 2020; Souza et al., 2020; da Rosa Mesquita et al., 2021). It is difficult to predict how the disease will manifest itself in an individual. Identifying biological variables, their longitudinal patterns, variations between individuals, and associations with demographic and clinical characteristics would aid the development of a risk-stratified approach to patient care.

Keywords and phrases: clustering, Coronavirus Disease 2019, COVID-19, EM algorithm, mixed-effects model, mixture model, SARS-CoV-2, severe acute respiratory syndrome coronavirus 2, spline.

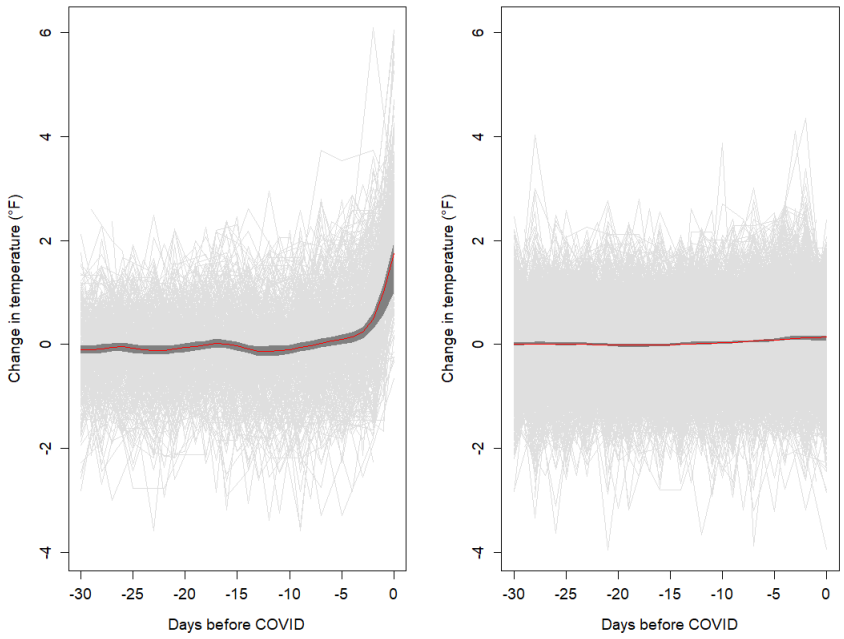

FIG 1. *The temperature profiles before confirmation time in 3,293 COVID-19 HD patients. Symptomatic (left) and asymptomatic (right) groups were classified by the method in this paper. Red lines are the estimates of mean functions, and shaded areas are 95% bootstrap confidence intervals.*

Nearly 786,000 people in the United States have end-stage renal disease (ESRD). About 488,000 ESRD patients travel to clinics to receive life-sustaining hemodialysis (HD) treatments and cannot shelter in place (USRDS, 2020; NIDDK, 2021). HD patients suffer from a host of comorbidities, such as diabetes and cardiac disease, putting them at an increased risk for complications from COVID-19. In addition, HD patients have reduced responses to SARS-CoV-2 vaccines (Simon et al., 2021). Thus, there is a pressing need to identify potential coronavirus carriers and develop procedures to curb the spread among HD patients. Numerous studies (Bivona, Agnello and Ciaccio, 2021; Malik et al., 2021; Gallo Marin et al., 2021) focus on the general population but only a few center on HD patients (Monaghan et al., 2021). The thrice-weekly in-center HD treatments provide results of a large number of clinical and treatment variables that are stored in patients' electronic health records (EHRs) and thus readily available for analysis. Utilizing EHRs, Chaudhuri et al. (2022) observed significant changes in many biological variables due to COVID-19 infection. However, these results estimated at the population level are not directly applicable to individual detection and prediction. For each patient, we compute temperature change as the difference between measured temperatures minus the average temperatures during a period free of COVID-19 infection. Figure 1 presents temperature change profiles before confirmation time in a cohort of COVID-19 HD patients. Since the body temperature of some patients raised a couple of days before being tested positive for COVID-19, these patterns are indicative of COVID-19 infection. Nevertheless, the temperatures of other patients remain relatively unchanged. We observed similar patterns for other biological variables, including pulse rate, systolic blood pressure, interdialytic weight gain, serum levels of albumin and ferritin, and counts of neutrophils and lymphocytes (Chaudhuri et al., 2022).

This paper focuses on estimating changes in biological and clinical indicators in in-center HD patients with COVID-19. We have two objectives: (a) clustering patients into two groups:

symptomatic and asymptomatic; and (b) associating the group probability with comorbidities and demographic/clinical characteristics. Existing methods for clustering longitudinal and functional data (see Bouveyron and Brunet (2014) and Jacques and Preda (2014) for reviews) do not apply since they were proposed only for objective (a). We proposed a novel Nonparametric Mixed-Effects Mixture (NMEM) model to fulfill both objectives (a) and (b). We model the probability of a latent group label using a logistic regression model and the response variable when the latent group label is given using a nonparametric mixed-effects model. Joo et al. (2009) considered a similar model with P-spline fixed effects only for the characteristics of urban groundwater recharge. Lu and Song (2012) studied a Bayesian model with B-spline fixed effects and random intercepts for the treatment effect of heroin use. The main contributions of this paper are as follows: (1) modeling both the group mean and subject deviation trajectories using smoothing splines, which makes the proposed NMEM model more flexible than those in Joo et al. (2009) and Lu and Song (2012). Our method extends that in Ma and Zhong (2008) by allowing the probability to depend on covariates; (2) introducing an L_1 regularization method for variable selection, which Joo et al. (2009) and Lu and Song (2012) did not consider; and (3) investigating changes in temperature in a cohort of COVID-19-infected hemodialysis patients using the proposed method. We note that the proposed method is general, which is not limited to the application illustrated in this paper.

The rest of the paper is organized as follows. We introduce the NMEM model and estimation procedure in Section 2. The real data analysis is reported in Section 3. The simulation and comparisons with other methods are presented in Section 4.

2. Nonparametric Mixed-Effects Mixture Model.

2.1. *Model Specification.* Denote y_{ij} as the observation from subject i at time t_{ij} where $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ and $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})^T$ be vectors of observations and time points from subject i. Let u_{ik} be a latent variable such that $u_{ik} = 1$ if subject i belongs to group k and $u_{ik} = 0$ otherwise. In this paper, for simplicity, we consider two latent groups (e.g., symptomatic and asymptomatic). The extension to more than two groups is straightforward. Denote $\boldsymbol{x}_i$ as a vector of covariates from subject i. We assume the following NMEM model:

$$p_{i1} := P(u_{i1} = 1) = \frac{\exp(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1)}{1 + \exp(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1)}, \tag{1}$$

$$\boldsymbol{y}_i = f_k(\boldsymbol{t}_i) + \mathbf{Z}_{i(k)} \mathbf{b}_{i(k)} + \boldsymbol{\epsilon}_i \text{ if } u_{ik} = 1,\ k = 1, 2, \tag{2}$$

where p_{i1} is the probability of subject i belonging to group 1; $f_k(\boldsymbol{t}_i) = (f_k(t_{i1}), f_k(t_{i2}), ..., f_k(t_{in_i}))^T$ is the mean function in group k, f_k, evaluated at time points $\boldsymbol{t}_i$; $\mathbf{Z}_{i(k)}$ is the design matrix for random effects; $\mathbf{b}_{i(k)} \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{G}_{i(k)})$ are the random effects associated with subject i nested within group k; $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ are the random errors; and $\mathbf{I}_{n_i}$ is an identity matrix with dimension n_i. We assume that u_{ik}, $\mathbf{b}_{i(k)}$, and $\boldsymbol{\epsilon}_i$ from different subjects are independent.

The logistic regression model in equation (1) models the probability of subject i belonging to group 1 as a function of covariates, and the nonparametric mixed-effect model in equation (2) models longitudinal trajectories of subject i given the group label. Given a latent group, since the trajectory of the mean function is usually unknown and could be nonlinear, we model the shape of the mean function nonparametrically using a cubic spline. Specifically, we scale time points into the interval $[0, 1]$, denote the first and second derivative of a function f as f' and f'', and assume that $f_k \in W_2^2[0, 1]$ where

$$W_2^2[0, 1] = \{f : \ f \text{ and } f' \text{ are absolutely continuous}, \int_0^1 (f''(t))^2 dt < \infty\}$$

is a Sobolev space. The space can be decomposed into two subspaces $W_2^2[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{f:\ f'' = 0\}$ and $\mathcal{H}_1 = \left\{f:\ f(0) = 0,\ f'(0) = 0,\ \int_0^1 (f''(t))^2 < \infty\right\}$ are reproducing kernel Hilbert spaces (RKHS) with reproducing kernels (RK) $R_0(s,t) = 1 + k_1(s)k_1(t)$, $R_1(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$, $k_1(t) = t - 0.5$, $k_2(t) = \frac{1}{2}\left(k_1^2(t) - \frac{1}{12}\right)$, and $k_4(t) = \frac{1}{24}\left(k_1^4(t) - \frac{k_1^2(t)}{2} + \frac{7}{240}\right)$. See Wang (2011) for more details.

The random effects $\mathbf{b}_{i(k)}$ model deviation of subject i's trajectory from the group mean. Different models may be considered for different applications. For example, one may include random intercepts and slopes. Since different subjects may have different nonlinear shapes in our application, we will consider smooth random effects associated with cubic splines (Wang, 1998a). Specifically, in addition to random intercepts and slopes, we will consider a zero mean Gaussian process with a covariance function proportional to the RK $R_1(s,t)$. Details are given in Section 3.

2.2. *Model Estimation.* Denote parameters in the covariance matrix of random effects $\mathbf{G}_{i(k)}$ as $\boldsymbol{\zeta}_k$. We need to estimate β_0, $\boldsymbol{\beta}_1$, σ^2, f_1, f_2, $\boldsymbol{\zeta}_1$, and $\boldsymbol{\zeta}_2$. Denote $N = \sum_{i=1}^m n_i$ as the total number of observations from all subjects. Let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_m^T)^T$, $\boldsymbol{u}_1 = (u_{11}, \ldots, u_{m1})^T$, and $\mathbf{V}_{ik} = \mathbf{Z}_{i(k)}\mathbf{G}_{i(k)}\mathbf{Z}_{i(k)}^T + \sigma^2\mathbf{I}_{n_i}$. The complete data likelihood of $(\boldsymbol{y}, \boldsymbol{u}_1)$ is calculated as $p(\boldsymbol{y}, \boldsymbol{u}_1) = p(\boldsymbol{y}|\boldsymbol{u}_1)p(\boldsymbol{u}_1) = \prod_{i=1}^m p(\boldsymbol{y}_i|u_{i1})p(u_{i1})$. We estimate all parameters and nonparametric functions using the following penalized likelihood:

$$
\begin{aligned}
l_c = &\left(\sum_{i=1}^m \left[u_{i1}(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}_1) - \log\left(1 + \exp\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}_1\right)\right)\right] + \lambda_0\|\boldsymbol{\beta}_1\|\right) \\
&- \left(\sum_{i=1}^m u_{i1}\left[\frac{n_i}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}_{i1}| + \frac{1}{2}\left(\boldsymbol{y}_i - f_1(\boldsymbol{t}_i)\right)^T\mathbf{V}_{i1}^{-1}\left(\boldsymbol{y}_i - f_1(\boldsymbol{t}_i)\right)\right] + N\lambda_1\int_0^1 (f_1'')^2 dt\right) \\
&- \left(\sum_{i=1}^m u_{i2}\left[\frac{n_i}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}_{i2}| + \frac{1}{2}\left(\boldsymbol{y}_i - f_2(\boldsymbol{t}_i)\right)^T\mathbf{V}_{i2}^{-1}\left(\boldsymbol{y}_i - f_2(\boldsymbol{t}_i)\right)\right] + N\lambda_2\int_0^1 (f_2'')^2 dt\right),
\end{aligned}
\tag{3}
$$

where the first part of the first line corresponds to $\log p(\boldsymbol{u}_1)$, the L_1 penalty with tuning parameter λ_0 in the second part of the first line is used to select covariates in $\boldsymbol{x}$, the first parts on the second and third lines correspond to $\log p(\boldsymbol{y}|\boldsymbol{u}_1)$, and the second parts in the second and third lines with smoothing parameters λ_1 and λ_2 control the smoothness of the mean functions in two groups.

Since the latent variables $\boldsymbol{u}_1$ are not observed, we use the EM algorithm to estimate the parameters. The E-step of the EM algorithm involves taking the conditional expectation of the likelihood conditional on the data and previously updated parameter values. Denote $\boldsymbol{\theta} = (f_1, f_2, \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \sigma^2, \lambda_0, \lambda_1, \lambda_2)$ as all the parameters to be estimated in both groups. Note that u_{ik} is Bernoulli. The conditional expectation of the latent variable u_{ik}

$$
w_{ik} := P(u_{ik} = 1|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}) = \frac{P(u_{ik} = 1|\hat{\boldsymbol{\theta}})P(\boldsymbol{y}_i|u_{ik} = 1, \hat{\boldsymbol{\theta}})}{\sum_{k'=1}^2 P(u_{ik'} = 1|\hat{\boldsymbol{\theta}})P(\boldsymbol{y}_i|u_{ik'} = 1, \hat{\boldsymbol{\theta}})},
\tag{4}
$$

where $k \in \{1, 2\}$, $\hat{\boldsymbol{\theta}}$ represents the estimated parameters from the last iteration, and $P(\boldsymbol{y}_i|u_{ik} = 1, \hat{\boldsymbol{\theta}})$ is a multivariate Gaussian density function of $\boldsymbol{y}_i$ with mean $f_k(\boldsymbol{t}_i)$ and variance $\mathbf{V}_{ik}$.

The conditional expectation of l_c becomes

$$
\mathbb{E}(l_c \,|\, \boldsymbol{y}, \hat{\boldsymbol{\theta}}) =
$$

$$(5)\quad \left(\sum_{i=1}^{m}\left[w_{i1}(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}_1)-\log\left(1+\exp\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}_1\right)\right)\right]+\lambda_0\|\boldsymbol{\beta}_1\|\right)$$

$$-\left(\sum_{i=1}^{m}w_{i1}\left[\frac{n_i}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{V}_{i1}|+\frac{1}{2}\left(\boldsymbol{y}_i-f_1(\boldsymbol{t}_i)\right)^T\mathbf{V}_{i1}^{-1}\left(\boldsymbol{y}_i-f_1(\boldsymbol{t}_i)\right)\right]+N\lambda_1\int_0^1(f_1'')^2dt\right)$$

$$(6)$$

$$-\left(\sum_{i=1}^{m}w_{i2}\left[\frac{n_i}{2}\log(2\pi)+\frac{1}{2}\log|\mathbf{V}_{i2}|+\frac{1}{2}\left(\boldsymbol{y}_i-f_2(\boldsymbol{t}_i)\right)^T\mathbf{V}_{i2}^{-1}\left(\boldsymbol{y}_i-f_2(\boldsymbol{t}_i)\right)\right]+N\lambda_2\int_0^1(f_2'')^2dt\right)$$

where $w_{i2}=1-w_{i1}$ and $\hat{\boldsymbol{\theta}}$ represents the estimated parameters from the last iteration.

The M-step involves maximizing $E(l_c\mid\boldsymbol{y},\hat{\boldsymbol{\theta}})$. We maximize equations (5) and (6) separately since they involve two disjoint set of parameters $\boldsymbol{\theta}_1=(\beta_0,\boldsymbol{\beta}_1,\lambda_0)$ and $\boldsymbol{\theta}_2=(f_1,f_2,\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2,\sigma^2,\lambda_1,\lambda_2)$ respectively.

Equation (5) is the penalized likelihood of a logistic regression model with an L_1 penalty. We apply the existing method and software package in R (Friedman, Hastie and Tibshirani, 2010) to update the estimate of $\boldsymbol{\theta}_1$. We use 10-fold cross-validation to select the tuning parameter λ_0.

For equation (6), the mean functions f_1 and f_2 are modeled using cubic splines. Based on the decomposition of the Sobolev space $W_2^2[0,1]=\mathcal{H}_0\oplus\mathcal{H}_1$, the estimated mean functions can be expressed as (Wang, 2011)

$$(7)\qquad f_k(t_{ij})=\sum_{\nu=1}^{2}d_{k\nu}\phi_\nu(t_{ij})+\sum_{l=1}^{e}c_{kl}\xi_l(t_{ij}),\quad k=1,2,$$

where $\phi_1(t)=1$ and $\phi_2(t)=t$ are basis functions of $\mathcal{H}_0$, $\xi_l(t)=R_1(z_l,t)$, and $\{z_1,\cdots,z_e\}$ are e distinct points in the set $\{t_{ij},\ i=1,\ldots,m,\ j=1,\ldots,n_i\}$. Let $\boldsymbol{d}_k=(d_{k1},d_{k2})^T$, $\boldsymbol{c}_k=(c_{k1},\ldots,c_{ke})^T$, and $(\tau_1,\ldots,\tau_N)^T$ be the stacked vectors of all time points $(t_{11},\ldots,t_{1n_1},\ldots,t_{mn_m})^T$. Denote $\mathbf{S}$ as an $N\times 2$ matrix with the (ι,ν)th entry $\phi_\nu(\tau_\iota)$, $\mathbf{R}$ as an $N\times e$ matrix with (ι,l)th entry $R_1(\tau_\iota,z_l)$, and $\mathbf{Q}$ as an $e\times e$ matrix with the (l,k)th entry $R_1(z_l,z_k)$. Let $\mathbf{W}_k$ be a block diagonal matrix of size $N\times N$ where the i-th block is $w_{ik}\mathbf{V}_{ik}^{-1}$. It is easy to verify that the target equation (6) is proportional to

$$(8)\qquad \sum_{i=1}^{m}\left(w_{i1}\log|\mathbf{V}_{i1}|\right)+\left(\boldsymbol{y}-\mathbf{S}\boldsymbol{d}_1-\mathbf{R}\boldsymbol{c}_1\right)^T\mathbf{W}_1\left(\boldsymbol{y}-\mathbf{S}\boldsymbol{d}_1-\mathbf{R}\boldsymbol{c}_1\right)+N\lambda_1\boldsymbol{c}_1^T\mathbf{Q}\boldsymbol{c}_1$$

$$(9)\qquad +\sum_{i=1}^{m}\left(w_{i2}\log|\mathbf{V}_{i2}|\right)+\left(\boldsymbol{y}-\mathbf{S}\boldsymbol{d}_2-\mathbf{R}\boldsymbol{c}_2\right)^T\mathbf{W}_2\left(\boldsymbol{y}-\mathbf{S}\boldsymbol{d}_2-\mathbf{R}\boldsymbol{c}_2\right)+N\lambda_2\boldsymbol{c}_2^T\mathbf{Q}\boldsymbol{c}_2.$$

We estimate the variance components $(\sigma^2,\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)$ and components corresponding to the mean functions $(\boldsymbol{c}_1,\boldsymbol{d}_1,\lambda_1,\boldsymbol{c}_2,\boldsymbol{d}_2,\lambda_2)$ alternatively.

When fixing the variance components $(\sigma^2,\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)$, we estimate the two mean functions and their smoothing parameters in equation (8) and (9) separately. Each one is the penalized least square for smoothing spline regression with correlated data. The minimizer of each satisfies the following equations (Gu, 2013):

$$(10)\qquad \begin{pmatrix}\mathbf{S}^T\mathbf{W}_k\mathbf{S} & \mathbf{S}^T\mathbf{W}_k\mathbf{R}\\ \mathbf{R}^T\mathbf{W}_k\mathbf{S} & \mathbf{R}^T\mathbf{W}_k\mathbf{R}+N\lambda_k\mathbf{Q}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}_k\\ \boldsymbol{c}_k\end{pmatrix}=\begin{pmatrix}\mathbf{S}^T\mathbf{W}_k\boldsymbol{y}\\ \mathbf{R}^T\mathbf{W}_k\boldsymbol{y}\end{pmatrix},\quad k=1,2.$$

Note that $\mathbf{W}_k$ is fixed at this step. Let $w_{ik}\mathbf{V}_{ik}^{-1}=\mathbf{P}_{ik}^T\mathbf{P}_{ik}$ be the Cholesky decomposition of $w_{ik}\mathbf{V}_{1k}^{-1}$ and $\mathbf{P}_k=\text{diag}(\mathbf{P}_{1k},\cdots,\mathbf{P}_{mk})$. Then $\mathbf{W}_k=\mathbf{P}_k^T\mathbf{P}_k$ is the Cholesky decomposition of $\mathbf{W}_k=\text{diag}(w_{1k}\mathbf{V}_{1k}^{-1},\cdots,w_{mk}\mathbf{V}_{mk}^{-1})$. With the transformations $(\boldsymbol{y}_k,\mathbf{S}_k,\mathbf{R}_k)=$

$\mathbf{P}_k^T(\boldsymbol{y}, \mathbf{S}, \mathbf{R})$, the weight matrices $\mathbf{W}_k$ in equation (10) are absorbed into other vectors/matrices and the equations reduce to those under the independent cases. Therefore, we can apply computational methods in Gu (2013) for independent data to update $\boldsymbol{c}_k$ and $\boldsymbol{d}_k$ with smoothing parameters λ_k selected by the generalized maximum likelihood method. More recent advances in computational methods can be found in Ma, Huang and Zhang (2015) and Sun, Zhong and Ma (2021).

When fixing the mean functions, we estimate the variance components by minimizing (8) and (9) together using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) algorithm (Byrd et al., 1995; Zhu et al., 1997). Since both equations contain the common variance of random error σ^2, we profiled it out and minimized the profiled likelihood. We provide the calculation of profiled likelihood in the Supplementary Material (Ma et al., 2023).

The initial estimates of variance components and the mean function are calculated using the linear mixed effects (LME) form of smoothing splines. These estimates are then used for the first inner iteration. The existing package nlme can be used for fitting these LME models. Details can be found in Wang (1998b) and Xu and Wang (2021).

We summarize the entire EM algorithm in Algorithm 1. The stopping criteria are set as follows. Let $\tilde{\boldsymbol{\theta}}$ represent all the estimated values except the penalty parameter λ_0 and the smoothing parameters λ_1 and λ_2. Let $d_{EM}^{[\mathcal{O}]} = \frac{\|\tilde{\boldsymbol{\theta}}^{[\mathcal{O}]} - \tilde{\boldsymbol{\theta}}^{[\mathcal{O}-1]}\|^2}{\|\tilde{\boldsymbol{\theta}}^{[\mathcal{O}-1]}\|^2 + \kappa_1}$ be the relative change of the estimates at the $\mathcal{O}$th EM iteration and $d_{inner}^{[\mathcal{O},\mathcal{I}]} = \frac{\|\tilde{\boldsymbol{\theta}}^{[\mathcal{O},\mathcal{I}]} - \tilde{\boldsymbol{\theta}}^{[\mathcal{O},\mathcal{I}-1]}\|^2}{\|\tilde{\boldsymbol{\theta}}^{[\mathcal{O},\mathcal{I}-1]}\|^2 + \kappa_2}$ be the relative change of the estimates at the $\mathcal{I}$th inner iteration inside the $\mathcal{O}$th EM iteration. The inner iteration stops if $d_{inner}^{[\mathcal{O},\mathcal{I}]} \leq D_{inner}$ and the EM iteration stops if $d_{EM}^{[\mathcal{O}]} \leq D_{EM}$. The maximum number of iterations are denoted as $\mathcal{O}_{max}$ and $\mathcal{I}_{max}$, respectively.

Algorithm 1: EM algorithm for fitting the NMEM model

Input : longitudinal data

Output: estimates of group probabilities, variance components, parameters of covariates and mean functions: $\hat{p}_{ik}$, $\hat{\boldsymbol{\zeta}}_k$, $\hat{\sigma}^2$, $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_1$, $\hat{f}_k$

randomly assign subjects with probability 0.5 into two groups where group 1 contains n_1 subjects and group 2 contains n_2 subjects;

set $\hat{p}_{ik} = n_k/N$ for all i;

get the initial estimates of $\hat{\boldsymbol{\zeta}}_k, \hat{\sigma}^2$ and $\hat{f}_k$ using the method introduced in Wang (1998b) and Xu and Wang (2021);

while $(\{d_{EM}^{[\mathcal{O}]} > D_{EM}\} \wedge \{\mathcal{O} <= \mathcal{O}_{max}\})$ **do**

  E-step:

  compute w_{ik} as in equation (4) using the current estimates $\hat{p}_{ik}$, $\hat{\boldsymbol{\zeta}}_k$, $\hat{\sigma}^2$ and $\hat{f}_k$;

  M-step:

  update $\hat{\boldsymbol{\beta}}$ using logistic regression as in equation (5) and calculate $\hat{p}_{ik}$ as in equation (1);

  while $(\{d_{inner}^{[\mathcal{O},\mathcal{I}]} > D_{inner}\} \wedge \{\mathcal{I} <= \mathcal{I}_{max}\})$ **do**

    solve equation (8) and (9) separately to update f_k with fixed $\hat{\boldsymbol{\zeta}}_k$ and $\hat{\sigma}^2$;

    minimize the sum of equation (8) and (9) to update $\hat{\boldsymbol{\zeta}}_k$ and $\hat{\sigma}^2$ with fixed f_k using L-BFGS-B;

  end

end

return $\hat{p}_{ik}$, $\hat{\boldsymbol{\zeta}}_k$, $\hat{\sigma}^2$, $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_1$, $\hat{f}_k$

We regard the L_1 penalty in the logistic regression as a variable selection process. After variable selection, we rerun Algorithm 1 without the L_1 penalty and get the point estimate of

all parameters. We construct bootstrap confidence intervals for the estimated mean functions and all parameters. We note that even though we are only interested in two clusters for our specific problem, the number of clusters in other applications is usually unknown. Model selection methods, such as BIC described in Ma and Zhong (2008), can be used to select the number of clusters.

3. Temperature Profiles in COVID-19-Infected HD Patients. This section applies our methods to investigate temperature change profiles before COVID-19 confirmation in HD patients. We consider 3,293 ESRD patients who received in-center HD treatment from Fresenius Medical Care and had positive PCR tests during 2020-01-01 and 2021-8-31. We align patients at their first positive PCR test date and analyze their temperature measurements 30 days before the PCR test. The observation window is defined as -30 to 0, where 0 is the PCR test date. The average number of observations for each patient is around 14. To focus on the changing pattern, we subtract each patient's temperature from the average temperature between -60 to -31 days, a period free of COVID-19 infection. Figure 1 shows temperature change profiles for all patients. Our goals are to identify two latent groups (symptomatic and asymptomatic), estimate the probability of belonging to each group for each subject, and associate the group probability with demographic and clinical characteristics.

Based on preliminary analyses, we consider an NMEM model with equation (1) and the following model in place of equation (2):

$$\boldsymbol{y}_i = f_k(\boldsymbol{t}_i) + b_{i1(k)}\mathbf{1} + b_{i2(k)}\boldsymbol{t}_i + s_{i(k)}(\boldsymbol{t}_i) + \boldsymbol{\epsilon}_i, \text{ if } u_{ik} = 1, k = 1, 2, \tag{11}$$

where groups 1 and 2 correspond to patients with and without temperature changes, f_1 and f_2 are the mean functions of these two groups, $b_{i1(k)}$, $b_{i2(k)}$, and $s_{i(k)}(\boldsymbol{t}_i) = (s_{i(k)}(t_{i,1}), \cdots, s_{i(k)}(t_{i,n_i}))^T$ are random intercept, slope, and a vector of smooth random effects associated with subject i nested within group k, and $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_i})$ are random errors independent of the random effects. Model (11) is a special case of the proposed model (2) with $\mathbf{Z}_{i(k)} = (\mathbf{1}, \boldsymbol{t}_i, \mathbf{I}_{n_i})$ and $\boldsymbol{b}_{i(k)} = (b_{i1(k)}, b_{i2(k)}, s_{i(k)}(t_{i,1}), \cdots, s_{i(k)}(t_{i,n_i}))^T$. We assume unstructured covariance for the random intercept and slope. In addition to the random intercept and slope, the smooth random effect $s_{i(k)}(t)$ is a function of t that allows a more flexible deviation of subject i from the population mean (Wang, 2011). We assume that $s_{i(k)}(t)$ follows a zero mean Gaussian process with a covariance function equal to the reproducing kernel of $\mathcal{H}_1$ (Wang, 1998a). Specifically, for the combined random effects, we assume that

$$\boldsymbol{b}_{i(k)} = \begin{pmatrix} b_{i1(k)} \\ b_{i2(k)} \\ \hline s_{i(k)}(t_{i,1}) \\ \vdots \\ s_{i(k)}(t_{i,n_i}) \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \left[\begin{array}{cc|c} \sigma^2_{inter,k} & \sigma_{is,k} & \mathbf{0} \\ \sigma_{is,k} & \sigma^2_{slope,k} & \\ \hline \mathbf{0} & & \sigma^2_{non,k} R_1(\boldsymbol{t}_i, \boldsymbol{t}_i) \end{array}\right]\right), \tag{12}$$

for $k \in \{1, 2\}$, where $\sigma^2_{inter,k}$, $\sigma^2_{slope,k}$, and $\sigma_{is,k}$ are variances and covariance of the random intercept and slope, $\sigma^2_{non,k}$ is the variance of smooth random effect, and $R_1(\boldsymbol{t}_i, \boldsymbol{t}_i)$ is an $n_i \times n_i$ matrix with the jkth entry as $R_1(t_{ij}, t_{ik})$. For model (11), we have $\boldsymbol{\zeta}_k = (\sigma^2_{inter,k}, \sigma_{is,k}, \sigma^2_{slope,k}, \sigma^2_{non,k})$, $k \in \{1, 2\}$.

Based on exploratory analysis and literature, we consider the following covariates for the logistic regression model (1): gender, race, ethnicity, vintage, diabetes, hypertension, BMI (calculated using height and the average weight after each dialysis treatment during the observation window), age on the treatment date, and vascular access type with three options:

arteriovenous fistula (AVF), arteriovenous grafts (AVG), and central venous catheter (CVCATH).

For the stopping criteria, we set $\kappa_1 = \kappa_2 = 10^{-5}$, $D_{EM} = 10^{-5}$, $D_{inner} = 10^{-5}$, $\mathcal{O}_{max} = 100$, and $\mathcal{I}_{max} = 5$. Different hyperparameters parameters lead to similar fits. After variable selection, we refit the model with selected variables and without the L_1 penalty. To avoid over-fitting, we set a lower bound for λ_k as $\log_{10}(N\lambda_k) = 0$ where $N = 46837$. The variance components in equation (12) are estimated using the covariance matrix of bivariate normal where $\sigma_{is,1} = \rho\sigma_{inter,1}\sigma_{slope,1}$. All variances are estimated using natural log transformation, and a tangent transformation is used for the correlation parameter ρ.

We perform the following diagnostics to evaluate the model assumptions. First, to check the independence assumption between random errors in $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})^T$ and model potential temporal correlation within patient i, we consider a continuous AR(1) random error structure with covariance between ϵ_{ik} at time t_{ik} and ϵ_{il} at time t_{il} equals $\sigma^2\phi^{|t_{ik}-t_{il}|}$. We extend our estimation and computational methods to fit models with correlated random errors. The estimate of ϕ equals 2.85×10^{-297} with 95% bootstrap confidence interval [$6.02 * 10^{-307}$,$3.86 * 10^{-38}$], indicating the autocorrelation is ignorable. Second, to check the normality assumptions for random errors and random effects, we compute estimates of random effects using the following formula in Wang (1998a):

$$\mathbf{Z}_{i(k)}\hat{\mathbf{b}}_{i(k)} = \mathbf{Z}_{i(k)}\hat{\mathbf{G}}_{i(k)}\mathbf{Z}'_{i(k)}\hat{\mathbf{V}}_{ik}^{-1}(\boldsymbol{y} - \mathbf{S}\hat{\boldsymbol{d}}_k - \mathbf{R}\hat{\boldsymbol{c}}_k),$$

where $\hat{\mathbf{G}}_{i(k)}$,$\hat{\mathbf{V}}_{ik}$, $\hat{\boldsymbol{d}}_k$, and $\hat{c}_k$ are the estimates, and a threshold of 0.5 was used to cluster patients into two groups. The estimated random effects and residuals' QQ plots (not shown) show no evidence of substantial violations of the normality assumptions.

We summarize estimates of coefficients associated with covariates, variance components, and their 95% bootstrap confidence intervals based on 1000 samples from the fitted model in Table 1.

We conclude that white and elderly patients were less likely to have a change in temperature. Age can reflect the strength of the patient's immune system. Elderly patients usually have a weaker immune system and are thus less responsive to the infection.

With the probability threshold for clustering set as 0.5, our method identified 468 (14.21%) out of 3,293 COVID-19 patients who experienced an increase in temperature before the positive PCR test. Figure 2 shows the estimated mean change functions and their 95% confidence intervals in the two groups. We note that confidence intervals for the mean functions are narrow due to a large number of total observations. As we can see, in group 1, the increase in temperature started about nine days and accelerated about four days before the PCR test date.

Table 2 shows several interesting characteristics that set patients with fever apart from those who did not have fever. Two observations are striking: first, age is higher in patients who did not have fever, and second, the rate of whites without fever was higher compared to the ones with fever. The observation regarding age agrees with clinical experience. Fever is less prevalent in the elderly (> 65 years) than in younger adults. The body temperature in the elderly is physiologically lower compared to middle-aged adults. In frail, older adults, fever is absent in 30 to 50 percent, even in serious infections such as pneumonia or endocarditis (Musgrave and Verghese, 1990; Henschke, 1993). The blunted febrile response in older adults is mostly due to impairment in multiple systems responsible for thermoregulation (e.g., shivering, vasoconstriction, hypothalamic regulation, and thermogenesis by brown adipose tissue) (Mackowiak, 1997). The impact of race on fever is less clear. Interestingly, in a study, fever was less prevalent in blacks compared to whites with temporal measurement and more prevalent with oral measurement (Bhavani et al., 2022). In that study of patients with suspected (non-COVID) infection, in 265 black patients with paired measurements within the first hour of presentation, fever prevalence was 23.4% with temporal and 35.8% with oral measurement. In 281 white patients, the prevalence was 27.8% with temporal and 26.0% with oral measurement.

TABLE 1

Result of data analysis. Upper: estimates of variance components. Lower: estimates of coefficients in the logistic regression model. We only report the variables selected by Lasso. We report the point estimate and 95% bootstrap confidence intervals. The subscripts for variance components are defined in equation (12). The subscripts for covariates with a "Y" indicate the patient has that comorbidity.

Parameters	Estimate	95% CI
$\hat{\sigma}^2$	0.4136	(0.4027,0.4239)
$\hat{\sigma}^2_{inter,1}$	0.0955	(0.0571,0.1854)
$\hat{\sigma}^2_{inter,2}$	0.0620	(0.0000,0.0752)
$\hat{\sigma}_{is,1}$	-0.0809	(-0.1435,-0.0291)
$\hat{\sigma}_{is,2}$	-0.0135	(-0.0268,0.0053)
$\hat{\sigma}_{slope,1}$	0.5130	(0.3793,0.6539)
$\hat{\sigma}_{slope,2}$	0.0956	(0.0467,0.1251)
$\hat{\sigma}^2_{non,1}$	14.8065	(9.0019,20.3593)
$\hat{\sigma}^2_{non,2}$	0.2195	(0.0000,1.0791)
$\hat{\beta}_0$	-1.0028	(-1.6899,0.2948)
$\hat{\beta}_{White}$	-0.3734	(-0.625,-0.0432)
$\hat{\beta}_{DiabetesY}$	-0.1085	(-0.3421,0.1256)
$\hat{\beta}_{BMI}$	0.0139	(-0.0047,0.0268)
$\hat{\beta}_{age}$	-0.0107	(-0.0203,-0.0023)
$\hat{\beta}_{AVG}$	0.2057	(-0.1251,0.5811)
$\hat{\beta}_{CVCATH}$	-0.1675	(-0.5195,0.1863)

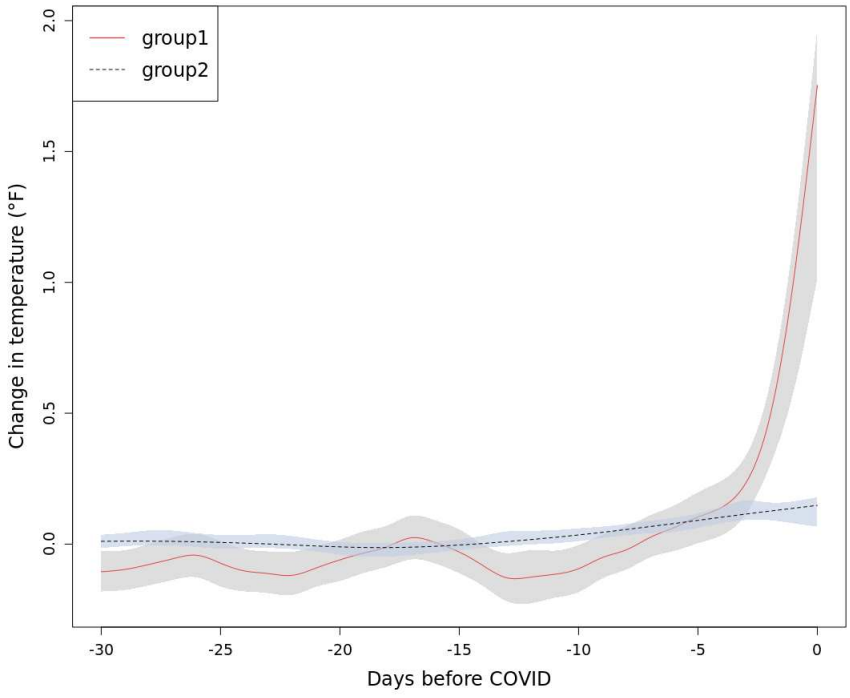

FIG 2. *Estimated mean functions of two groups and their 95% bootstrap confidence intervals.*

TABLE 2
Statistics of patients in different groups. The abbreviation "std" stands for standard deviation.

Variable	All	Temperature Increased	Temperature Unchanged
Female (%)	44	42	44
White (%)	64	55	66
Hispanic (%)	21	20	21
Diabetes (%)	52	49	52
Hypertension (%)	79	80	79
AVF (%)	67	68	67
AVG (%)	13	15	12
BMI (std)	30.84 (8.20)	31.78 (9.35)	30.68 (7.99)
Age (std)	61.69 (14.19)	59.58 (14.03)	62.04 (14.19)
Vintage (std)	4.64 (4.09)	4.69 (4.40)	4.64 (4.04)

4. Simulation Study. We conduct simulations to evaluate the performance of the proposed method and compare it with the previous work in Lu and Song (2012). Lu and Song (2012) considered the following model:

$$P(u_{ik}=1)=p_{ik}=\frac{\exp(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}_k)}{\sum_{k'=1}^{2}\exp(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}_{k'})}, \quad i=1,\ldots,m,$$

$$\boldsymbol{y}_i=\tilde{\mathbf{X}}_i^{[1]}f_k(\mathbf{t}_i)+\tilde{\mathbf{X}}_i^{[2]}\boldsymbol{\alpha}_k+\mathbf{Z}_{i(k)}\mathbf{b}_{i(k)}+\boldsymbol{\epsilon}_{ik} \quad \text{if } u_{ik}=1, \tag{13}$$

where $\tilde{\mathbf{X}}_i^{[1]}$ and $\tilde{\mathbf{X}}_i^{[2]}$ are the design matrices of fixed effects $f_k(\mathbf{t}_i)$ and $\boldsymbol{\alpha}_k$ respectively, $\mathbf{Z}_{i(k)}$ is the design matrix for the random effects $\mathbf{b}_{i(k)}$, and the variances of random errors $\boldsymbol{\epsilon}_{ik}$ are assumed to be different for different groups.

Our method differs from Lu and Song's methods in both the model structure and estimation approach. Lu and Song (2012) modeled the nonparametric functions f_k's using P-splines while we use smoothing splines. Our proposed model includes a smooth random effect for flexibility and allows different random effects in different groups. Lu and Song (2012) estimate parameters in a Bayesian framework while we estimate parameters using penalized likelihood. In addition, our estimation procedure includes an L_1 penalty for variable selection.

We generate data using model (1) and (11), which was used in real data analysis in Section 3. We set $m=3293$, n_i's are randomly selected from integers in the interval $[16,31]$, and parameters as their estimates when possible.

We generate latent variables u_{i1} for $i=1,\cdots,m$ using equation (1). We include all nine covariates (before variable selection) for the group probability in equation (1) and generated them according to their estimated marginal distributions. Specifically, six categorical variables are generated according to their empirical ratios in each category. Three continuous variables are generated from an exponential distribution with a rate parameter of 0.22 (vintage), a Gamma distribution with a shape parameter of 15.58 and a rate parameter of 0.51 (BMI), and a normal distribution with a mean 61.69 and standard deviation 14.20 (age). Parameters in three continuous distributions are set to be the maximum likelihood estimates. For the $\boldsymbol{\beta}$ coefficients in the logistic model (1), we set the intercept as the estimate in Table 1, coefficients for gender, ethnicity, vintage, diabetes, hypertension, BMI, and vascular access types as zero since they are either not selected in the variable selection process or the confidence interval contains zero in the final estimation. The coefficients for race and age are set to be the estimates in Table 1. The latent variables u_{i1} for $i=1,\cdots,m$ are generated according to a Bernoulli distribution with probability given in equation (1).

We use equation (11) to generate responses $\boldsymbol{y}_i$. Since Lu and Song (2012)'s model does not have a smooth random effect, we consider two simulation settings: with smooth random

TABLE 3

Simulation results: average accuracy in the two settings. NMEM stands for our method, LS stands for LU and Song's method. We also report the average mean squared error (MSE) of estimated mean functions.

	Without $\sigma^2_{non,k}$		**With** $\sigma^2_{non,k}$	
	NMEM	LS	NMEM	LS
Accuracy	0.9341	**0.9344**	**0.8239**	0.6351
MSE f_1	**0.0017**	0.0018	**0.0240**	0.0494
MSE f_2	**0.0000**	0.0001	**0.0007**	0.0158

effect where $\sigma^2_{non,1} = 14.8$ (estimate from the real data), $\sigma^2_{non,2} = 14$ and without smooth random effect where $\sigma^2_{non,1} = \sigma^2_{non,2} = 0$. In both settings, the values of all other parameters apart from the variance of smooth random effect are set to be the estimates in the real data analysis in Table 1. We first randomly generate the number of observations n_i from a discrete uniform distribution on $[16, 31]$ for each subject i, and then randomly select n_i days between -30 to 0 as $\boldsymbol{t}_i$. The values of mean functions f_k at each time point are set to be the estimates in the real data analysis. Independent and identically distributed random errors are generated from a normal distribution with mean zero and variance equals the estimate from the real data analysis.

To fit model (13), we set $\tilde{\mathbf{X}}_i^{[1]}$ as an identity matrix and $\tilde{\mathbf{X}}_i^{[2]} = 0$. Lu and Song (2012) used the random permutation sampler approach in Frühwirth-Schnatter (2001) to deal with the label-switching problem caused by the symmetric prior of the parameters in different components. We use the variance of random slope for the permutation sampler in our simulations.

Each simulation is replicated 100 times. We use the same stopping criteria as in the real data analysis. The comparison results are presented in Table 3, Figure 3 and Figure 4.

Both methods performed well when no smooth random effect existed. The clustering accuracy and MSEs of function estimates are almost identical. The existence of a smooth random effect allows nonlinear individual departure from the population mean function, thus making clustering and estimation more difficult. This is reflected in the smaller accuracy and larger MSEs in Table 3. As expected, our method achieves better accuracy and lower MSEs than Lu and Song's method when smooth random effect existed. In general, our method has smaller biases in the estimates of variance components (Figures 3 and 4).

We also conducted a simulation study to test the performance of variable selection. The values of all parameters, including the variance of the smooth random effects, are set to be the estimates in the real data analysis in table 1. For variable selection in the model (1), on average, 0.3 of the two non-zero parameters (not including the intercept) are mistakenly excluded from the model, and 2.95 out of the eight zero parameters are mistakenly selected. The over-selection behavior agrees with the previous literature (Tibshirani, 1996; Chetverikov, Liao and Chernozhukov, 2021).

To evaluate the performance of the proposed method under sparse longitudinal data situation, in addition to generating n_i's from $[16, 31]$, we consider two alternative settings where n_i's are randomly generated from the sets of integers in the intervals $[4, 8]$ and $[6, 23]$ respectively while keeping other parameters unchanged. Table 4 summarizes clustering accuracies and mean squared errors under the three different settings for n_i's. The proposed method performed reasonably well under sparse situations, and the performance improves as observations become denser.

We included the codes in the supplementary materials and posted them to GitHub (`https://github.com/HubDaniel/NMEM`). A subset of data is available at the NIH RADx Data Hub (`https://radx-hub.nih.gov/`).

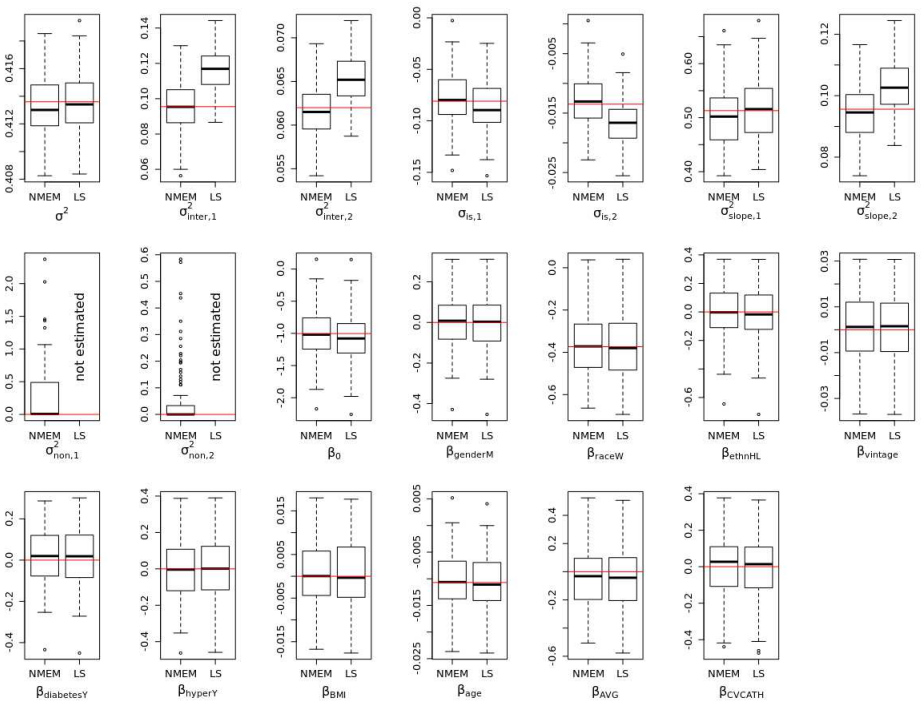

FIG 3. *Simulation results without smooth random effects: box plots of estimated parameters from NMEM and Lu and Song's method (LS). Red lines are true values of parameters.*

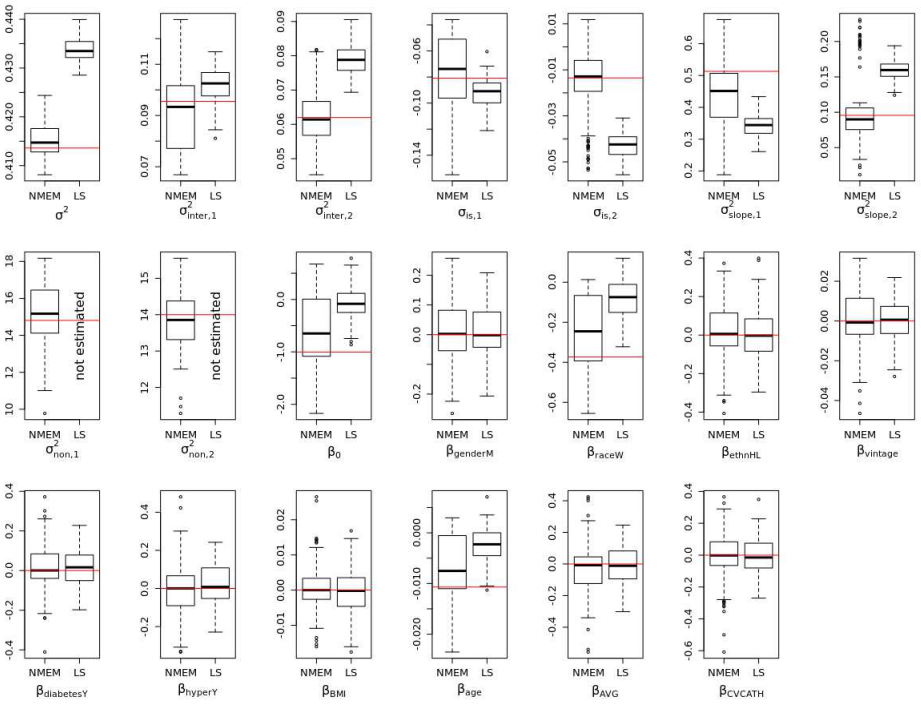

FIG 4. *Simulation results with smooth random effects: box plots of estimated parameters from NMEM and Lu and Song's method (LS). Red lines are true values of parameters.*

TABLE 4
Simulation results: average accuracy and the average mean squared error (MSE) of estimated mean functions.

	With $\sigma^2_{non,k}$		
Intervals for n_i	[4,8]	[6,23]	[16,31]
Accuracy	0.6446	0.7354	0.8239
MSE f_1	0.0524	0.0394	0.0240
MSE f_2	0.0026	0.0014	0.0007

5. Conclusion and Future Work. This article proposes a unified method for clustering longitudinal trajectories and relating the subgroups to other biological and clinical predictors. A flexible nonparametric mixed-effects mixture model is proposed to identify risk factors and classify patients with a change in body temperature before the diagnosis of COVID-19. We model the change in temperature using smoothing splines. We use penalized likelihood and the EM algorithm to estimate the mean functions, variance components, and covariates associated with the clustering probability. A simulation study shows that our method performs well and, under certain scenarios, outperforms existing methods. The results of data analysis suggest that different demographic characteristics influence the immune system response and provide an improved understanding of patient groups that may or may not experience a more severe course following infection with SARS-CoV-2.

The identification of two clusters with respect to pre-diagnosis dynamics of body temperature is novel and significant. It indicates the existence of biological phenotypes that may react to infection differently during the COVID-19 incubation period. While the underlying biological reasons are unclear, it will be of interest in future analysis to explore the clinical course and patient outcomes after diagnosis of COVID-19 differ. One could hypothesize that patients with a pre-diagnostic rise in temperature may have a more severe infection (e.g., higher virus load) or augmented inflammatory response (e.g., higher levels of interleukin-6) that may translate into worse COVID-19 outcomes.

The identifiability of mixture models has received a great deal of attention, and some general conditions were provided in the existing literature (Teicher, 1963; Holzmann, Munk and Gneiting, 2006; Wang, Yao and Huang, 2014; Aragam et al., 2020; Wong, Zeng and Lin, 2022). However, these general conditions are complicated and difficult to verify. Specific conditions for the identifiability of the proposed model warrant further research.

Acknowledgments. We would like to thank the editor and two anonymous reviewers for their insightful comments that significantly improved the manuscript.

Funding. This research is partially supported by NIH grants R01-DK130067, R01-HL161303, R01-DK117208, and NSF grant DMS-2053423.

SUPPLEMENTARY MATERIAL

Supplement A: Profiled likelihood
(doi: ; .pdf). Profiled likelihood used for estimation.

Supplement B: Code-Independent random error
(doi: ; .zip). Code for the implementation with independent random error.

Supplement C: Code-CAR1 random error
(doi: ; .zip). Code for the implementation with CAR1 random error.

REFERENCES

ARAGAM, B., DAN, C., XING, E. P. and RAVIKUMAR, P. (2020). Identifiability of nonparametric mixture models and Bayes optimal clustering. *The Annals of Statistics* **48**. https://doi.org/10.1214/19-AOS1887

BHAVANI, S. V., WILEY, Z., VERHOEF, P. A., COOPERSMITH, C. M. and OFOTOKUN, I. (2022). Racial differences in detection of fever using temporal vs oral temperature measurements in hospitalized patients. *JAMA* **328** 885. https://doi.org/10.1001/jama.2022.12290

BIVONA, G., AGNELLO, L. and CIACCIO, M. (2021). Biomarkers for prognosis and treatment response in COVID-19 patients. *Annals of Laboratory Medicine* **41** 540–548. https://doi.org/10.3343/alm.2021.41.6.540

BOUVEYRON, C. and BRUNET, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* **71** 52–78. https://doi.org/10.1016/j.csda.2012.12.008

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16** 1190–1208. https://doi.org/10.1137/0916069

CHAUDHURI, S., LASKY, R., JIAO, Y., LARKIN, J., MONAGHAN, C., WINTER, A., NERI, L., KOTANKO, P., HYMES, J., LEE, S., WANG, Y., KOOMAN, J. P., MADDUX, F. and USVYAT, L. (2022). Trajectories of clinical and laboratory characteristics associated with COVID-19 in hemodialysis patients by survival. *Hemodialysis International* **26** 94–107. https://doi.org/10.1111/hdi.12977

CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2021). On cross-validated Lasso in high dimensions. *The Annals of Statistics* **49**. https://doi.org/10.1214/20-AOS2000

DA ROSA MESQUITA, R., FRANCELINO SILVA JUNIOR, L. C., SANTOS SANTANA, F. M., FARIAS DE OLIVEIRA, T., CAMPOS ALCÂNTARA, R., MONTEIRO ARNOZO, G., RODRIGUES DA SILVA FILHO, E., GALDINO DOS SANTOS, A. G., OLIVEIRA DA CUNHA, E. J., SALGUEIRO DE AQUINO, S. H. and FREIRE DE SOUZA, C. D. (2021). Clinical manifestations of COVID-19 in the general population: systematic review. *Wiener klinische Wochenschrift* **133** 377–382. https://doi.org/10.1007/s00508-020-01760-4

DE MORAES BATISTA, A. F., MIRAGLIA, J. L., RIZZI DONATO, T. H. and PORTO CHIAVEGATTO FILHO, A. D. (2020). COVID-19 diagnosis prediction in emergency care patients: A machine learning approach preprint, Epidemiology. https://doi.org/10.1101/2020.04.04.20052092

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**. https://doi.org/10.18637/jss.v033.i01

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96** 194–209. https://doi.org/10.1198/016214501750333063

GALLO MARIN, B., AGHAGOLI, G., LAVINE, K., YANG, L., SIFF, E. J., CHIANG, S. S., SALAZAR-MATHER, T. P., DUMENCO, L., SAVARIA, M. C., AUNG, S. N., FLANIGAN, T. and MICHELOW, I. C. (2021). Predictors of COVID-19 severity: A literature review. *Reviews in Medical Virology* **31** 1–10. https://doi.org/10.1002/rmv.2146

GU, C. (2013). *Smoothing spline ANOVA models*, 2nd ed ed. *Springer series in statistics*. Springer, New York, NY.

HARAHWA, T. A., LAI YAU, T. H., LIM-COOKE, M.-S., AL-HADDI, S., ZEINAH, M. and HARKY, A. (2020). The optimal diagnostic methods for COVID-19. *Diagnosis* **7** 349–356. https://doi.org/10.1515/dx-2020-0058

HENSCHKE, P. J. (1993). Infections in the elderly. *Medical Journal of Australia* **158** 830–834. https://doi.org/10.5694/j.1326-5377.1993.tb137672.x

HOLZMANN, H., MUNK, A. and GNEITING, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics* **33** 753–763. https://doi.org/10.1111/j.1467-9469.2006.00505.x

JACQUES, J. and PREDA, C. (2014). Functional data clustering: A survey. *Advances in Data Analysis and Classification* **8** 231–255. https://doi.org/10.1007/s11634-013-0158-y

JIANG, X., COFFEE, M., BARI, A., WANG, J., JIANG, X., HUANG, J., SHI, J., DAI, J., CAI, J., ZHANG, T., WU, Z., HE, G. and HUANG, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua* **62** 537–551. https://doi.org/10.32604/cmc.2020.010691

JOO, Y., BRUMBACK, B., LEE, K., YUN, S.-T., KIM, K.-H. and JOO, C. (2009). Clustering of temporal profiles using a Bayesian logistic mixture model: Analyzing groundwater level data to understand the characteristics of urban groundwater recharge. *Journal of Agricultural, Biological, and Environmental Statistics* **14** 356–373. https://doi.org/10.1198/jabes.2009.07100

KUKAR, M., GUNČAR, G., VOVKO, T., PODNAR, S., ČERNELČ, P., BRVAR, M., ZALAZNIK, M., NOTAR, M., MOŠKON, S. and NOTAR, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Scientific Reports* **11** 10738. https://doi.org/10.1038/s41598-021-90265-9

LU, Z. and SONG, X. (2012). Finite mixture varying coefficient models for analyzing longitudinal heterogenous data. *Statistics in Medicine* **31** 544–560. https://doi.org/10.1002/sim.4420

MA, P., HUANG, J. Z. and ZHANG, N. (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika* **102** 631–645.

MA, P. and ZHONG, W. (2008). Penalized clustering of large-scale functional data with multiple covariates. *Journal of the American Statistical Association* **103** 625–636. https://doi.org/10.1198/016214508000000247

MA, X., GUO, W., GU, M., USVYAT, L., KOTANKO, P. and WANG, Y. (2023). Supplement to "A Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with COVID-19".

MACKOWIAK, P. A. (1997). *Fever: Basic Mechanisms and Management*, 2nd ed ed. Lippincott-Raven.

MALIK, P., PATEL, U., MEHTA, D., PATEL, N., KELKAR, R., AKRMAH, M., GABRILOVE, J. L. and SACKS, H. (2021). Biomarkers and outcomes of COVID-19 hospitalisations: systematic review and meta-analysis. *BMJ Evidence-Based Medicine* **26** 107–108. https://doi.org/10.1136/bmjebm-2020-111536

MONAGHAN, C. K., LARKIN, J. W., CHAUDHURI, S., HAN, H., JIAO, Y., BERMUDEZ, K. M., WEINHANDL, E. D., DAHNE-STEUBER, I. A., BELMONTE, K., NERI, L., KOTANKO, P., KOOMAN, J. P., HYMES, J. L., KOSSMANN, R. J., USVYAT, L. A. and MADDUX, F. W. (2021). Machine learning for prediction of patients on hemodialysis with an undetected SARS-CoV-2 infection. *Kidney360* **2** 456–468. https://doi.org/10.34067/KID.0003802020

MUSGRAVE, T. and VERGHESE, A. (1990). Clinical features of pneumonia in the elderly. *Seminars in Respiratory Infections* **5** 269–75.

NIDDK (2021). Kidney disease statistics for the united states NIDDK.

PIMENTEL, M. A. F., REDFERN, O. C., HATCH, R., YOUNG, J. D., TARASSENKO, L. and WATKINSON, P. J. (2020). Trajectories of vital signs in patients with COVID-19. *Resuscitation* **156** 99-106. https://doi.org/10.1016/j.resuscitation.2020.09.002

SIMON, B., RUBEY, H., TREIPL, A., GROMANN, M., HEMEDI, B., ZEHETMAYER, S. and KIRSCH, B. (2021). Haemodialysis patients show a highly diminished antibody response after COVID-19 mRNA vaccination compared with healthy controls. *Nephrology Dialysis Transplantation* **36** 1709–1716. https://doi.org/10.1093/ndt/gfab179

SOUZA, T. H., NADAL, J. A., NOGUEIRA, R. J. N., PEREIRA, R. M. and BRANDÃO, M. B. (2020). Clinical manifestations of children with COVID-19: A systematic review. *Pediatric Pulmonology* **55** 1892–1899. https://doi.org/10.1002/ppul.24885

SUN, X., ZHONG, W. and MA, P. (2021). An asymptotic and empirical smoothing parameters selection method for smoothing spline ANOVA models in large samples. *Biometrika* **108** 149–166. https://doi.org/10.1093/biomet/asaa047

TEICHER, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics* **34** 1265–1269. https://doi.org/10.1214/aoms/1177703862

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

USRDS (2020). Unites states renal data system annual data report.

WANG, Y. (1998a). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 159–174. https://doi.org/10.1111/1467-9868.00115

WANG, Y. (1998b). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* **93** 341–348. https://doi.org/10.1080/01621459.1998.10474115

WANG, Y. (2011). *Smoothing splines: methods and applications*. *Monographs on statistics and applied probability*. CRC Press, Boca Raton, FL.

WANG, S., YAO, W. and HUANG, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statistics & Probability Letters* **93** 41–45. https://doi.org/10.1016/j.spl.2014.06.010

WONG, K. Y., ZENG, D. and LIN, D. Y. (2022). Semiparametric latent-class models for multivariate longitudinal and survival data. *The Annals of Statistics* **50**. https://doi.org/10.1214/21-AOS2117

WU, J., ZHANG, P., ZHANG, L., MENG, W., LI, J., TONG, C., LI, Y., CAI, J., YANG, Z., ZHU, J., ZHAO, M., HUANG, H., XIE, X. and LI, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results preprint, Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.04.02.20051136

XU, D. and WANG, Y. (2021). Low-rank approximation for smoothing spline via eigensystem truncation. *Stat* **10**. https://doi.org/10.1002/sta4.355

ZHU, C., BYRD, R. H., LU, P. and NOCEDAL, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software* **23** 550–560. https://doi.org/10.1145/279232.279236