

RESEARCH ARTICLE

Functional principal component based landmark analysis for the effects of longitudinal cholesterol profiles on the risk of coronary heart disease

Bin Shi^{1,2}  | Peng Wei²  | Xuelin Huang² 

¹Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas

²Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas

Correspondence

Peng Wei, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
Email: pwei2@mdanderson.org

Funding information

National Science Foundation, Grant/Award Number: DMS 1612965; National Institute of General Medical Sciences, National Institutes of Health, Grant/Award Number: 5T32 GM074902; The endowed Dr. Mien-Chie Hung and Mrs. Kinglan Hung Professorship; U.S. National Institutes of Health, Grant/Award Numbers: R01CA169122, R01HL116720, U54 CA096300, U01 CA152958, 5P50 CA100632

Patients' longitudinal biomarker changing patterns are crucial factors for their disease progression. In this research, we apply functional principal component analysis techniques to extract these changing patterns and use them as predictors in landmark models for dynamic prediction. The time-varying effects of risk factors along a sequence of landmark times are smoothed by a supermodel to borrow information from neighbor time intervals. This results in more stable estimation and more clear demonstration of the time-varying effects. Compared with the traditional landmark analysis, simulation studies show our proposed approach results in lower prediction error rates and higher area under receiver operating characteristic curve (AUC) values, which indicate better ability to discriminate between subjects with different risk levels. We apply our method to data from the Framingham Heart Study, using longitudinal total cholesterol (TC) levels to predict future coronary heart disease (CHD) risk profiles. Our approach not only obtains the overall trend of biomarker-related risk profiles, but also reveals different risk patterns that are not available from the traditional landmark analyses. Our results show that high cholesterol levels during young ages are more harmful than those in old ages. This demonstrates the importance of analyzing the age-dependent effects of TC on CHD risk.

KEYWORDS

functional principal components analysis, landmark analysis, longitudinal study, survival analysis

1 | INTRODUCTION

In many clinical settings, changing disease severity and rate of disease progression are accompanied by dynamic changes in longitudinal biomarker measurements. Therefore, it is important to use the longitudinal biomarker data collected during each patient's follow-up visit to make real-time predictions of future disease events. This is dynamic prediction. Such real-time predictive information is critical for physicians and patients, so that they can actively monitor the disease and possibly plan for early treatment and prevention regimes. As a result, using longitudinal data to construct real-time prediction models has prevailed within the medical literature. The traditional statistical approach for dynamic prediction is to develop a joint model that uses some parametric models for longitudinal data. The problem is that the model may not fit well for some real data. functional principal component analysis (FPCA) is a more flexible and robust alternative

that is without distribution assumptions. In this research, we evaluate simple and easy-to-interpret-models—landmark analysis coupled with FPCA—and apply them to the dynamic prediction of cardiovascular disease (CVD).

CVD is the leading cause of morbidity and mortality among men and women in the United States,¹ and it is a complex condition that consists of a number of diseases related to atherosclerosis, which includes coronary heart disease (CHD). Some factors used to assess the risk of CHD include lipid profile, blood pressure, obesity, diabetes, and smoking. The incidence and prevalence of CHD increase exponentially with age. Therefore, selecting older people as the target population for prevention and therapy is unsurprisingly straightforward, even though the resulting reductions in CHD risk might be small. In fact, there is an ongoing debate regarding efficacy and whether targeting classical risk factors in adults older than 70 is beneficial.^{2,3} Considerable evidence has shown that higher levels of total cholesterol (TC) are associated with increased risk of CHD in adults in a younger age range (40–69 years). However, the importance of hypercholesterolemia as a risk factor for CVD in the older age group remains controversial. Thus, it is unknown whether TC or other risk factors are important targets for effective CHD prevention in elderly individuals.

Longitudinal biomarkers are often used to monitor patient health status in clinical practice. For example, the lipid profile is a good indicator of CVD risk. It is recommended that adults aged 20 and older have their serum cholesterol measured at least every 5 years. This entails a blood test called lipoprotein profiling. A TC value of less than 200 mg/dL is desirable, while a measure of 200 to 239 mg/dL is borderline high and 240 mg/dL or above is considered high. In order to use such information longitudinally in clinical practice, both landmark analysis and FPCA modeling are important techniques to handle the longitudinal effect of time-varying variables, such as TC values.

Survival models with time-dependent covariates describe the joint relationship between follow-up time T of the occurrence of an event and covariates $Z(t)$ by modeling the hazard $h(t|Z(t))$. The use of time-varying covariates typically assumes that $Z(t)$ are available for all possible timepoints. However, in practice, we can only collect a measured covariate history at discrete times and obtain a patient's vital status discretely. Estimating a patient's prognosis through these observed data would require knowledge of the future values of $Z(t)$, which involves integrating over the conditional distribution of the future covariate process given the history of $Z(t)$. Therefore, analysis can traditionally be done by making a joint model for the failure event and the history of the time-dependent covariates. We can use such an approach to derive a predictive model by conditioning on the history of time-dependent covariates and other fixed covariates to perform dynamic prediction.^{4,5} However, this approach usually requires parametric assumptions for the distribution and trajectory of longitudinal biomarkers, which may not hold in the real data analysis. Moreover, the approach of joint models involves intensive computation and complicated modeling, therefore not convenient for use in practice.

In this article, we use landmark analysis models for dynamic prediction. Compared with the joint modeling approach, the landmark model is simpler to implement and provides easier clinical interpretation. As a result, the landmark model has been used extensively in medical research. First introduced by Anderson et al,⁶ landmark analysis can dynamically adjust predictive models for survival data with follow-up data. With this approach in hand, simple proportional hazards models are able to capture the development trend of the risk over time for modeling data with time-dependent covariates.

A traditional multivariate framework is modeled with data in the form of random vectors, but under a functional data framework, it focuses on data in the form of curves, shapes and images as realizations of a stochastic process. They are intrinsically of infinite dimensions, but measured discretely. Multivariate principal component analysis is a useful tool for data dimension reduction. FPCA extends this concept by reducing random trajectories to a set of functional principal component (FPC) scores for functional data analysis. Besides dimension reduction, FPCA attempts to extract the dominant features from varied random curves around their mean trends. The principal components of FPCA, which exhibit the variability of the data through the continuous index time, are curves and can be seen naturally as the major modes of variation. Early work on FPCA was done for asymptotic properties.^{7,8} Further development of the theory and applications was achieved by Boente and Fraiman.⁹ Yao et al^{10,11} proposed to estimate the principal components via conditional expectation (PACE) for the analysis of not only regular grid data, but also highly irregular or sparse data. More recent developments involve multivariate FPCA, which differs from one-dimensional FPCA in that it aims to use a multivariate functional Karhunen-Loeve representation of the longitudinal data.^{12–14}

In the current study, we use landmark only or landmark combined with FPCA to perform the survival model estimation. We find that our two approaches yield similar conclusions for estimating the risk of CHD associated with TC levels. However, the landmark-only model provides only the overall trend, which is a marginal TC effect. By contrast, the landmark model with FPCA can decompose this overall trend into different orthogonal patterns, from which we can interpret the overall effect in detail or perform further prediction analyses in the future. We identify three different risk profiles corresponding to three distinct CHD risk levels, which may not only explain the overall trend effect, but also give the three scenarios we might encounter in clinical practice. An important finding is that a higher TC was associated

with a greater risk of CHD in the younger adult age group (<60-65 years), but in the older age (>65 years), the TC levels were no longer associated with CHD risk. This provides additional evidence for the debate as to whether it is beneficial to treat older patients in order to lower TC levels. Our results are consistent with the recent work done by two groups.^{2,15} In addition, we use FPCA-based simulation data to evaluate the two models' prediction performance. We show that the FPCA-based landmark analysis consistently has lower Brier scores (smaller prediction error) and higher values for the area under receiver operating characteristic curve (AUC), which yields better discrimination ability.

The rest of this article is organized as follows. Section 2 presents details of the statistical models that we addressed, landmark analysis and FPCA models. Section 3 presents simulation studies. Section 4 provides the results of the real data analysis. Section 5 concludes with a discussion.

2 | METHODS

2.1 | Dataset description

The Framingham Heart Study (FHS) is a family based, ongoing prospective cohort study, which investigates CVD-related risk factors for three generations of residents from the town of Framingham, Massachusetts. For our present study, we use the offspring cohort dataset that contains longitudinal phenotype measurements, including TC, low density lipoprotein cholesterol, high density lipoprotein cholesterol (HDL-C), total triglycerides, glucose, blood pressure, and body mass index, collected every 5 years for seven consecutive visits per participant. The age range represented in the data was from 17 to 85 years, which almost covers the entire life of most subjects. We downloaded the datasets from the National Center for Biotechnology Information dbGaP, study accession number phs000007.v30.p11 and phs000342.v18.p11.¹⁶

To build the survival models, we used unrelated individuals and selected the age interval of 25 to 75 years from the offspring cohort. The total sample size is 1669. The number of CHD events is 110 (6.59%), and the number of censoring is 1559 (93.41%).

2.2 | Review of landmark analysis

In medical research, we often encounter covariates that change over time. The FHS dataset includes several time-varying covariates, which are also called longitudinal biomarkers, including TC. The use of time-varying covariates allows researchers to not only explore associations, but also provides potential opportunities for making causal inference. Unfortunately, we also encounter a lot of technical difficulty, for example, choosing the covariate forms has great potential for the introduction of bias into the analysis. To this end, without making comprehensive models, we build landmark models for the time-varying covariates by constructing stacked datasets together with all the relevant information needed. Then, we can use the constructed datasets to compute the predictive probability by selecting individuals who are at risk at that moment, and then making predictions for their future risk using only their currently available information at that moment. The process of making a data selection at a moment in time is called landmarking.

More specifically, we construct a series of datasets for five landmark time points of ages: $L = 30, 40, 50, 60$, and 70 years as an example, which are more often denoted as time points L in the FHS datasets. At each time point of landmarking, we can fit a simple Cox model and use it to obtain a prediction of the participant's overall survival probability. For each participant who was enrolled and followed up in the study before time point L , we use the last-value-carry-forward (LVCF) method to define the biomarker value at the landmark time, denoted by covariate values $W_i^{(L)} = Z_i(\max_{t \leq L} : Z_i(t) \text{ is observed})$, for $i = 1, 2, \dots, n$.

Letting T_i and C_i denote the event time and censoring time, respectively, both are at the scale of each individual's age. We assume C_i is independent of the biomarker measurement time and event time T_i . Rather than observing T_i for all the participants, we observe only $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. For participant i , at each landmark time L , we assume a Cox proportional hazards model that specifies $\lambda_{Li}(s)$, the hazard function for $T_i - L$ at time $T_i - L = s$, as follows,

$$\lambda_{Li}(s|T_i \geq L, \{Z_i(t), 0 \leq t \leq L\}) = \lambda_{L0}(s) \exp(\vec{\alpha}^{(L)} \vec{V}_i + \beta^{(L)} W_i^{(L)}), \quad L \leq s \leq L + t_{\text{hor}}, \quad (1)$$

where $\lambda_{L0}(t)$ is a baseline hazard function, and \vec{V}_i denotes for baseline covariates, such as gender and smoking status, and $\vec{\alpha}^{(L)}$ is their coefficients. $W_i^{(L)}$ is a summary statistic of $\{Z_i(t), 0 \leq t \leq L\}$ and t_{hor} is a horizon time. It is to be specified by a data analyst. For example, it could be the current biomarker value, that is, $W_i^{(L)} = Z_i(L)$. It could also be a changing slope of the biomarker such as $W_i^{(L)} = \{Z_i(L) - Z_i(L - \Delta)\} / \Delta$. After separately fitting a regular Cox model for each landmark time point L , we stack these five datasets into one big one, and fit this unified dataset by a stratified-Cox model, with each dataset as a stratum, which we term a super model for convenience.¹⁷ We expect the coefficients $\beta^{(L)}$ of the super model to depend on L in a smooth function. For example, we can impose smoothness in a simple way, such as piece-wise constant, or we can bring more structure into the analysis by making the regression parameters $\beta^{(L)}$ as a complex function of L in a parametric way.^{17,18} In doing so, we should see several advantages to this approach. First, the super model can be fitted by using standard software for the large, stacked dataset. In addition, we can use a simple Cox model to approximate the complex survival model with several time-varying effects together. Such approximation usually works well if the time-dependent effect does not vary too much over time.¹⁸ However, fitting the Cox models for each landmark time interval separately would “ignore” the overlaps between the landmark datasets. Without loss of generality, one of the polynomial forms is adopted for the $\beta(L)$ function,^{19,20} and we specify a super model as follows:

$$\lambda_{Li}(s|T_i \geq L, \{Z_i(t), 0 \leq t \leq L\}) = \lambda_{L0}(s) \exp\{\vec{\alpha} \vec{V}_i + W_i^{(L)} \beta(L)\}, \quad L \leq s \leq L + t_{\text{hor}}, \quad (2)$$

where $\beta(L)$ is a fractional polynomial with a form such as $\beta(L) = \beta_0 + \beta_1(1/L) + \beta_2(1/\sqrt{L}) + \beta_3 \log(L) + \beta_4 \sqrt{L} + \beta_5 L + \beta_6 L^2$. It would be inappropriate to use such a super model's Wald statistics to test the statistical significance of the biomarkers' coefficients. The correct standard errors can be obtained by using the sandwich estimators to account for the “clustering” effect of the repeated data.²¹

2.3 | Proposed landmark analysis with FPCs as predictors

There are various ad hoc ways to use historical longitudinal biomarker values, such as using the current biomarker value. When this is not available, the LVCF value is used, but this is hardly a systematic approach. In many situations, the historical biomarker trajectory features (changing patterns) are more important than the current value or recent change in magnitude, in terms of predicting a future event. Since we expect the changing patterns of TC are important predictors for predicting the risk of CHD, we used FPCA methods to capture these changing patterns. First, FPCA were performed on the whole available dataset, then we integrated the FPC scores on each separate landmark time interval. Finally, we performed a Cox regression analysis in term of landmark principles by using the integrally obtained scores as our predictors in survival models.

To apply the FPCA to our longitudinal data (please see the review of FPCA method in the Appendix), in brief, we can use M eigen functions to project onto the infinite-dimensional function space of covariate random trajectories: a very good approximation can be obtained without losing too much variation. Due to the function decomposition, each $\hat{\rho}_k(t)$, where $k = 1, \dots, M$ and $t \in [0, \tau]$, may be viewed as the patterns of deviations from the population mean over time, and $\hat{\gamma}_{ik}$ describes how strongly the data from the i th subject follow the changing pattern $\rho_k(t)$ as a weight function. Therefore, we can naturally use these low-dimensional FPC scores γ_{ik} to substitute the original data as the new predictor variables in the survival analysis to model the relationship between the survival time and the patterns of the trajectories. This will improve our estimation, prediction, and interpretation in many ways.

Given the mean estimates $\hat{\mu}(t)$ and choosing certain basis function estimates $\hat{\rho}_k(t)$, the various FPC scores estimates $\hat{\gamma}_{ik}$ are obtained by integrating all available information on each landmark time interval $[0, L]$ and result in different trajectory patterns for different subjects, which approximate to the original data structure as closely as possible. So these estimates are still functions of the landmark time L . The formula is

$$\hat{\gamma}_{ik}^{(L)} = \int_0^L \{Z_i(t) - \hat{\mu}(t)\} \hat{\rho}_k(t) dt, \quad (3)$$

where the estimated $\hat{\mu}(t)$ and $\hat{\rho}_k(t)$ are obtained from the whole dataset. Since $Z_i(t)$ denotes the longitudinal biomarker trajectory for subject i and $t \geq 0$, the substituted data can be a multivariate p -dimensional vector. Here, we use one longitudinal marker as an example. Then we use $\hat{\gamma}_{ik}^{(L)}, k = 1, \dots, M$ as predictors, and the fitting survival model is as follows:

$$\lambda_{Li}(s|T_i \geq L) = \lambda_{L0}(s) \exp \left\{ \bar{\alpha}^{(L)} \bar{V}_i + \sum_{k=1}^M \beta_k^{(L)} \hat{\gamma}_{ik}^{(L)} \right\}, \quad L \leq s \leq L + t_{\text{hor}}, \quad (4)$$

where $\beta^{(L)} = (\beta_1^{(L)}, \dots, \beta_M^{(L)})'$ are the regression coefficients for the M estimated FPC score vector $\hat{\gamma}_i^{(L)} = (\hat{\gamma}_{i1}^{(L)}, \dots, \hat{\gamma}_{iM}^{(L)})'$. We choose an M value that makes the overall variation explained be larger than or equal to 95%. Please see the details in the Appendix about how to choose the M values. After fitting the above model (4) for each L separately, we fit a super model, as below, to get smooth estimates for the regression coefficient by a fractional polynomial function,

$$\lambda_{Li}(s|T_i \geq L) = \lambda_{L0}(s) \exp \left\{ \bar{\alpha} \bar{V}_i + \sum_{k=1}^M \beta_k(L) \hat{\gamma}_{ik}^{(L)} \right\}, \quad L \leq s \leq L + t_{\text{hor}}, \quad (5)$$

As before, $\beta_k(L)$ is a fractional polynomial with a form like $\beta_k(L) = \beta_{k0} + \beta_{k1}(1/L) + \beta_{k2}(1/\sqrt{L}) + \beta_{k3} \log(L) + \beta_{k4} \sqrt{L} + \beta_{k5} L + \beta_{k6} L^2$. The organization of the datasets is similar, as previously described in Section 2 for model (2). Various smoothness constraints, such as splines and fractional polynomials, can be placed on $\beta_k(L)$. Binder et al²² used simulations to compare these two approaches, and concluded that fractional polynomials better recover simpler functions, whereas splines better recover more complex functions. Generally, $\beta_k(L)$ is believed to be a simple smooth function so that fractional polynomial is better than splines in our application.

3 | SIMULATION

In order to evaluate which provides the better performance, the proposed approach of landmark (LM) coupled with FPCA or the landmark-only approach, we performed simulations to study their statistical prediction properties.

3.1 | Simulation setting

We performed a simulation study by mimicking the FHS as a real data example. (Please see the detailed analysis of FHS data in the following section of this article.) Specifically, we conducted strategies similar to those of previous work by Wei's group.^{16,23} First, we simulated a time-varying biomarker of subject i by using the first three TC FPCA components ($\gamma_{i1}, \gamma_{i2}, \gamma_{i3}$) in the following model: $Y_i(s) = \mu(s) + \gamma_{i1}\rho_1(s) + \gamma_{i2}\rho_2(s) + \gamma_{i3}\rho_3(s) + \epsilon$, where $\rho_1(s), \rho_2(s)$, and $\rho_3(s)$ are obtained from three eigen functions, $i = 1, 2, \dots, n$, and $s \in [25, \tau_i]$, $\tau_i \in [35, 75]$, they are recorded exam ages, which follow independent normal distributions with means of seven visit ages [34.5, 42.7, 47.1, 50.3, 54.0, 58.1, 60.8] and variances of those visit ages [9.0², 9.0², 9.1², 9.0², 9.0², 8.9², 8.9²]. $\epsilon \sim N(0, 0.1)$. Three FPC scores are independently simulated from normal distributions: $\gamma_{i1} \sim N(0, 32.3^2)$, $\gamma_{i2} \sim N(0, 26.5^2)$, and $\gamma_{i3} \sim N(0, 15.3^2)$. Then the time to CHD disease is generated according to the formula: $\lambda_i(s) = \lambda_0(s) \exp(\beta_1(s)\gamma_{i1} + \beta_2(s)\gamma_{i2} + \beta_3(s)\gamma_{i3})$. We adopted the coefficient values based on estimating the effects of TC on the CHD risk by using the FPCs as covariates in the Cox analysis for the FHS data.

We consider a common baseline hazard functional form: $\lambda_0(s) = \lambda s^{\lambda-1} \exp(\eta)$, which follows a Weibull distribution with $\lambda = 1.5$, $\eta = -7$. We adjusted such parameter values to make its distribution of time to CHD increasingly similar to that of the real data in FHS. We also modified three estimated effects of TC FPC scores based on the original estimates, and let $(\beta_1, \beta_2, \beta_3) = (-0.27, 0.30, -0.48)$, sample size $n = 1669$, replication times $Q = 500$. Without loss of generality, random censoring times are generated independently from a uniform distribution, and we set approximately 25% of the event times to be right-censored.

LM coupled with FPCA compared with landmark-only models may have good estimation, or better prediction performance on functional data. In order to assess them, we conduct the above simulation and focus on the predictive performance.²⁴ The Brier score is used to assess the calibration capacity by calculating the mean square difference between the true vs predicted survival probabilities of the two models. To incorporate censoring in survival data, the weighted version of the Brier scores is calculated by using inverse probability as censoring weights according to the following formula:²⁵

$$\hat{BS}(t, L) = \frac{1}{n_L} \sum_{i \in R(L)} \left\{ \frac{(1 - \hat{S}_{L_i}(t|Z_i(L))^2 I(T_{iL} > t))}{\hat{S}_c(t)} + \frac{(0 - \hat{S}_{L_i}(t|Z_i(L))^2 I(T_{iL} \leq t, \delta_i = 1))}{\hat{S}_c(T_{iL})} \right\},$$

where n_L denotes the size of risk set $R(L)$ at the landmark time L ; $T_{iL} = T_i - L$; $\hat{S}_c(t) = P(C_i > t)$ is the survival function of censoring time as the inverse weights, which can be estimated via the Kaplan-Meier method by letting $C_{iL} = C_i - L$; and $\hat{S}_{L_i}(t|Z_i(L))$ is the predicted survival probability for individual i with biomarker variable $Z_i(L)$ at the prediction time t , which can be estimated by Equation (1) in the landmark-only model or Equation (4) in the LM coupled with FPCA model. As an example of LM coupled with FPCA, $\hat{S}_{L_i}(t) = \{\hat{S}_{L_0}(t)\}^{\exp\{\sum_{k=1}^M (\hat{\gamma}_{ik}^{(L)} \hat{\rho}_k^{(L)})\}}$, where $\hat{S}_{L_0}(t) = \exp\{-\sum_{s \leq t} \hat{\lambda}_{L_0}(s)\}$, and $\hat{\lambda}_{L_0}(s)$ is a baseline hazard function. We used the R package “pec” to compute it.²⁵

To assess the discrimination ability of the two models, we compute the AUC values according to the following formula:^{5,26}

$$\widehat{AUC}(t, L) = Pr[\hat{S}_i(t|Z_i(L)) < \hat{S}_j(t|Z_j(L)) | \tilde{T}_i \in (L, t] \cap \tilde{T}_j > t],$$

where $\hat{S}_i(t|Z_i(L))$ is the predicted survival probability for individual i with biomarker variable $Z_i(L)$ at the prediction time t . At each evaluated time point, AUC summarizes all the sensitivity and specificity along the curve and gives the discriminating power to distinguish between subjects with different risk levels, as defined by the above formula. In a similar way, the weighted method is applied to incorporate censoring information to calculate the concordance index for survival models.²⁷ For simplicity, a nonparametric approach is used to compute the AUC values, and it is similar to compute the Wilcoxon statistics for survival data.²⁸ Here, we used R package “tdROC” to perform the calculations.²⁹

3.2 | Simulation results

Compared with the landmark-only approach, the major difference of our proposed LM coupled with FPCA is that we apply a more comprehensive and statistically principled method to extract predictive features from the longitudinal biomarker data. The landmark-only option implements the method of LVCF. However, LM coupled with FPCA first uses the whole dataset to compute the FPC scores, then uses all available accumulative information up to the landmark time point for prediction. In order to assess such differences, we implemented the FPCA-based method to generate functional data, and then studied the empirical performance of the two approaches. Table 1 summarizes the predictive performance results.

Prediction at landmark time was set at $L = 35, 50, 65, 75$ years in exam ages, and three prediction horizon times were investigated at $t_{hor} = t - L = 1, 3, 5$ years. Table 1 shows that, compared with the landmark-only model, the LM coupled with FPCA model produces a much smaller prediction error, in average tenfold lower in Brier scores, for all settings. At the same time, the AUC values of the proposed model are about 30% higher than landmark-only models in all the settings. This indicates our proposed functional approach performs much better than the LVCF-based method in terms of calibration and discrimination, respectively. One can infer from our simulation studies that, when we deal with functional data, especially sparse and irregular grids, or frequently missing-caused irregularity, LVCF-based methods mis-specify the models and could cause biased coefficient estimates and inaccurate prediction. Functional approach based models might account for such irregularity, minimize the adverse effects, pursue models flexibility, and obtain better performance.

4 | APPLICATION TO THE FHS

4.1 | Lowess curves analysis for TC

It has been reported that the following major risk factors affect lipid profiles, such as high blood pressure, cigarette smoking, family history of early heart disease, old age, and obesity (BMI 30 or higher). These are also the risk factors for CHD events, so they are highly correlated together during disease developments. In general, cholesterol is a fat-like substance in the blood, which is necessary for the build-up of artery walls. However, high blood cholesterol is one of the major risk factors for heart disease. Too much cholesterol in the blood can cause the deposit or formation of plaque in vessel walls,

TABLE 1 Evaluation of predictive performance: Brier scores and AUC values

Time		Brier scores		AUC values	
<i>L</i> (age: years)	<i>t</i> _{hor}	LM-only (1e-2)	LM-FPCA (1e-2)	LM-only (1e-1)	LM-FPCA (1e-1)
35	1	5.565	0.353	5.028	6.718
35	3	5.846	0.424	5.033	6.778
35	5	6.177	0.491	5.037	6.865
50	1	5.565	0.353	5.030	6.725
50	3	5.846	0.424	5.035	6.782
50	5	6.177	0.490	5.039	6.868
65	1	5.565	0.353	5.024	6.720
65	3	5.846	0.424	5.029	6.778
65	5	6.177	0.491	5.033	6.865
75	1	5.565	0.353	5.034	6.715
75	3	5.846	0.424	5.039	6.777
75	5	6.177	0.490	5.043	6.863

Abbreviations: AUC, area under receiver operating characteristic curve; FPCA, functional principal component analysis; LM, landmark.

hardening of arteries, and narrowing or total blockage of blood flow to the heart, which eventually can lead to a heart attack.

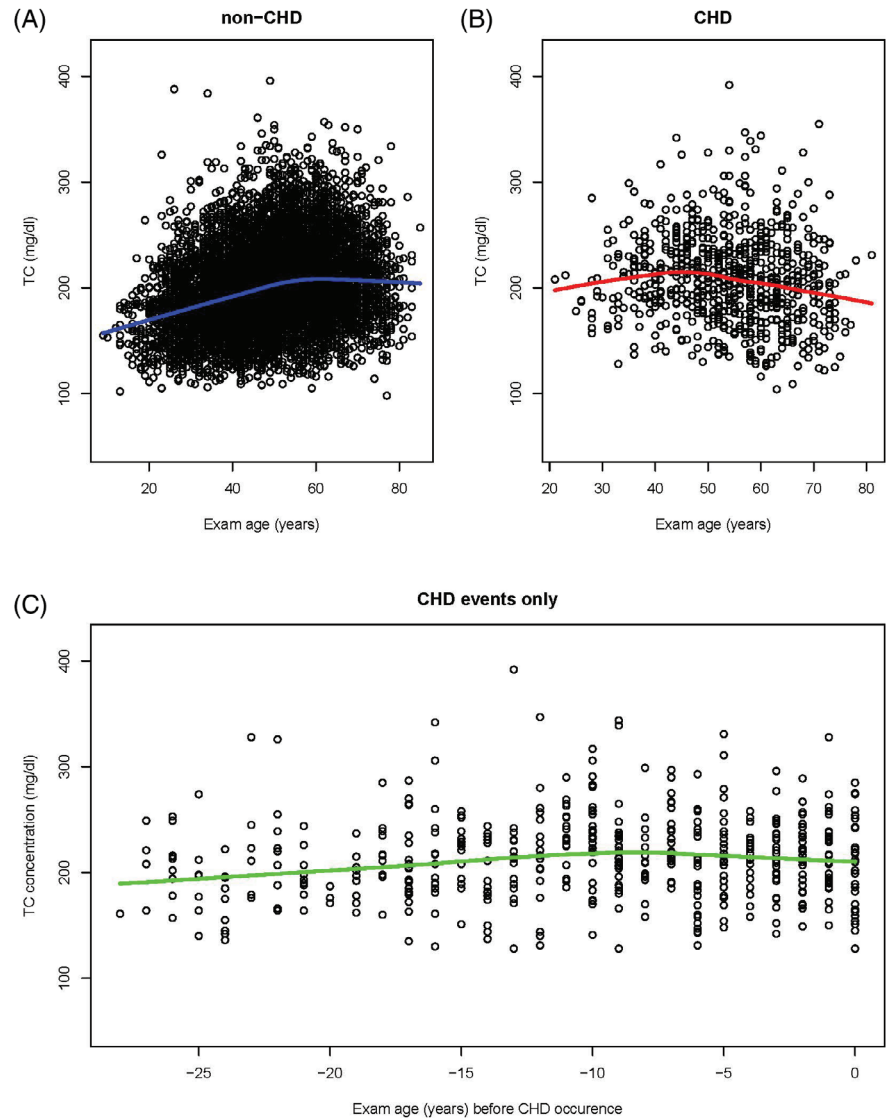
In this study, we collected TC levels and the ages of participants at the time of the exam for seven visits from the FHS dataset, and separated the corresponding participants into two groups: participants who experienced CHD (CHD event occurred) and participants who did not experience CHD (non-CHD event). Then we assessed the time-varying information by constructing the three lowess smoothing plots (Figure 1). For non-CHD events, the lowess plot shows that the TC smoothing curve starts from levels lower than 200 mg/dL; with increasing exam age, TC slowly increases until ages greater than 55, at which time TC decreases a bit and gradually flattens out (Figure 1A). For the group “CHD event occurred,” the lowess curve shows that the TC level starts with 200 mg/dL, then increases gradually until reaching peak level at age 50, slowly decreasing after, and is lower than 200 mg/dL after age 60 (Figure 1B). We also plotted the lowess smoothing curves for TC levels against the time-varying windows closest to when the CHD events occurred (length of ages; Figure 1C). This curve starts from a lower TC level, which is a bit closer to 200 mg/dL at a time that is 25 years before the CHD event occurs, then slowly increases to the TC level peak (higher than 220 mg/dL) at 9 years before the CHD event occurs. After that, it gradually decreases to 200 mg/dL at the time point closest to when the CHD event occurs.

In general, from these three plots, we can see in the CHD event occurred group, that the TC level is higher than in the non-CHD group during earlier adult ages. But during later ages, the TC levels become similar among the two groups. We also see that the overall trend of the TC level increases with age until about 10 years before the CHD event, at which time the TC level is maintained at a high level.

4.2 | Landmark analysis

We performed the landmark analysis on the longitudinal biomarkers TC. First, separated datasets were constructed with stratification on landmark time points. Then, we fitted the landmark Cox model in Equation (1), based on the different landmark datasets separately. So the effect of TC is fixed at each landmark time interval. We plotted the coefficients against the landmark time points with their 95% confidence interval (CI) values in Figure 2A. This plot shows that at the beginning ages, the coefficient is higher than zero and increases over time until age 40, reaching to a peak. This indicates that high TC is associated with greater risk of CHD during earlier adult ages. After age 40, the curves decrease with age until 55, at which point the coefficient CIs include zero. This means that TC was not associated with CHD event occurrence. After age 55, the plot shows a second increasing trend for unknown reasons until age 60, and then decreases to below zero (95% CIs include zero). The overall trend implies that for adults of younger age (below 40), higher TC levels correlate with

FIGURE 1 Plots of TC values against the participants' ages at the exam time, showing the changing trend of the time-varying covariate TC with exam ages for participants who experienced a CHD event and those who did not (non-CHD). A, TC lowess curve for participants experiencing non-CHD; B, TC lowess curve for CHD participants; C, Lowess curve of TC on the time windows close to and before CHD occurred for participants who experienced CHD only. CHD, coronary heart disease; TC, total cholesterol [Colour figure can be viewed at wileyonlinelibrary.com]



greater risk of CHD. With increasing age, the CHD risk associated with TC level decreases and then vanishes at older ages (after around age 55).

At the same time, we conducted the super model analysis described by Equation (2) in our methods, Section 2.2. By stacking the individual landmark datasets together, we can estimate the coefficients dependent on the landmark time L . The regression coefficients in such a model can be treated as weighted average effects of the time-varying covariates over the landmark time. We brought a smooth structure into the analysis by modeling the regression parameters as a fractional polynomial specified in Equation (2). It turns out that among those terms in $\beta(L)$, only the $\log(L)$ term is statistically significant. Moreover, since the youngest age of subjects we selected is 25 years, we actually used $\frac{L-25}{100}$ instead of L throughout our analyses. Consequently, the smooth function form is $\beta(L) = \beta_0 + \beta_1 * \log\left(\frac{L-25}{100}\right)$. This smooth curve is plotted in Figure 2B. We see similar patterns as the unsmoothed plot (Figure 2A): at younger adult ages, higher TC is associated with greater risk for the CHD; with increasing age, this risk becomes smaller and vanishes at older ages.

4.3 | Results of landmark analysis with FPCs as predictors

Since FPCA can be applied for capturing the underlying TC trajectory patterns via the PACE approach, we checked the density of data time points by the design plot to see whether the longitudinal TC values in the FHS data are suitable for PACE. The details of the design plot (Figure 6) of TC can be found in the Appendix.

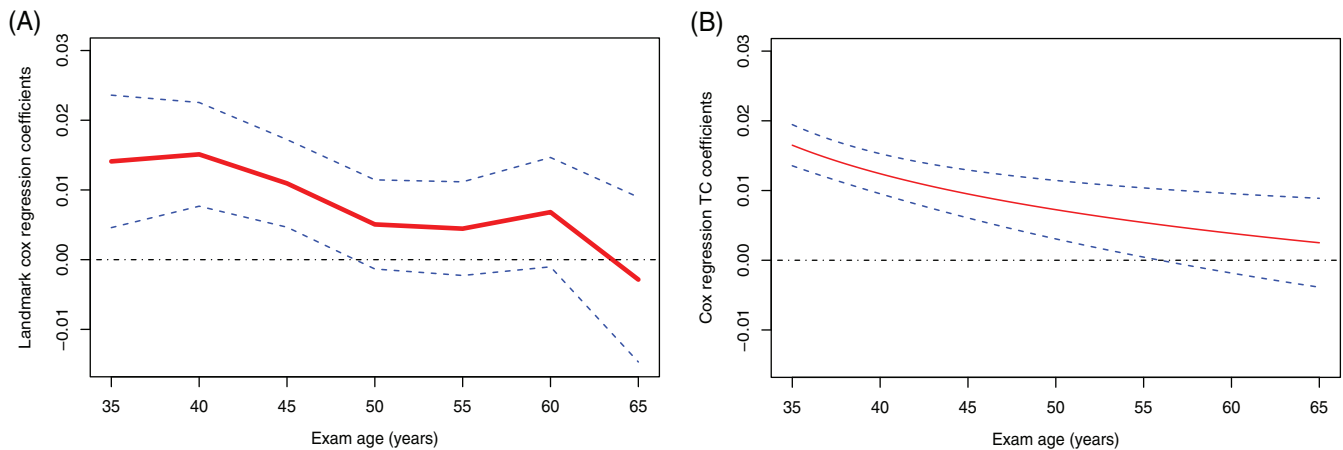


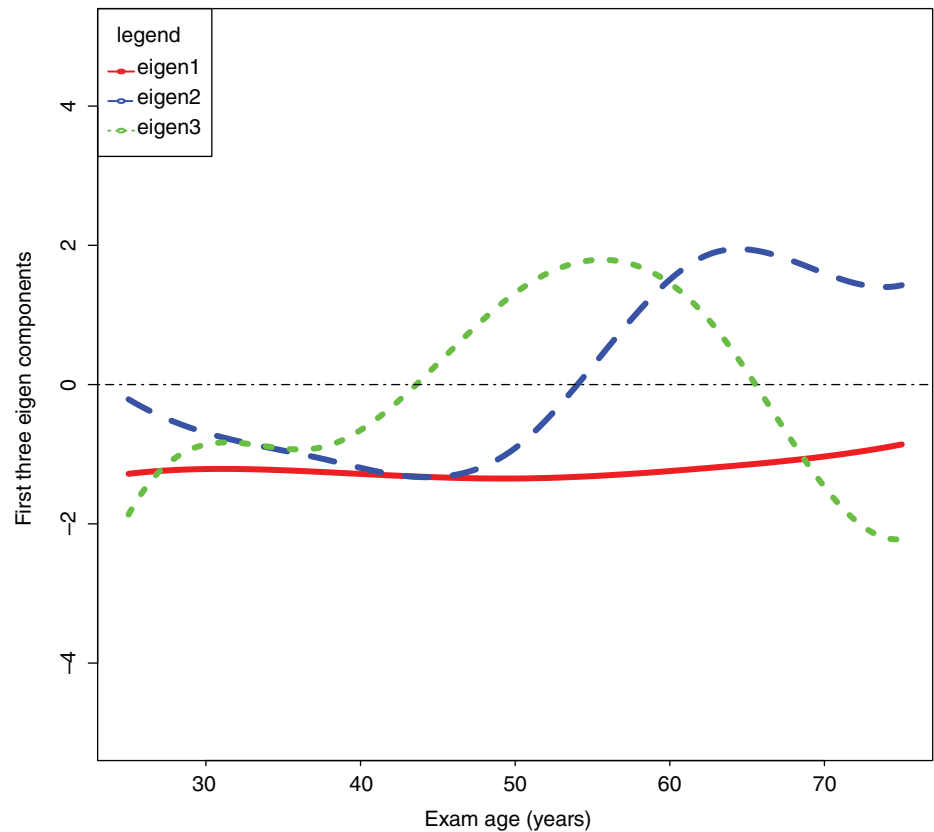
FIGURE 2 Plots of TC coefficients against participants' ages at the time of the exam after adjusting baseline covariates of gender and smoking status, showing the changing trend of the time-varying covariate TC effect with exam ages for landmark models. A, Sliding landmark effects with pointwise 95% confidence intervals for survival analysis of CHD against TC in the model; B, smooth curves on the coefficient of the landmark model against time using the formula $\beta(L) = \beta_0 + \beta_1 * \log(\frac{L-25}{100})$. CHD, coronary heart disease; TC, total cholesterol [Colour figure can be viewed at wileyonlinelibrary.com]

We performed FPCA for TC trajectory analysis by using all the available FHS data, and obtained all the eigen functions. These eigen functions described the patterns of deviations from the population mean over time. We selected the first three eigen functions, and their corresponding FPC scores explained 72.0%, 17.5%, and 10.0% of the variation, respectively, with a total 99% of the variation. As demonstrated in Figure 3, the first eigen function showed an almost constant reduction from the mean values at the time interval 25 to 60 years and started increasing a bit in the age interval 60 to 75 years. The second eigen function showed a slowly decreasing trend over time from zero to a negative value at age interval 25 to 45 years, then started to show an increase to a positive peak value at close to age 65 years. The third one had a bell shape, which increased slowly from 25 to 55 years of age, and then started to decrease.

Since FPCA can capture the individual changing patterns of TC from the longitudinal data by using mutually orthogonal eigen functions, we obtained the FPC scores by integrating on the seven time windows: ages 25-40, 25-45, 25-50, 25-55, 25-60, 25-65, and 25-70 separately resulting in the accumulative scores, which represent the cumulative biomarker historical values on each time interval. The averages of the standard deviation for the scores at the seven age intervals are 0.4488, 0.13022, and 0.1185 for FPC I, II, and III, respectively. These calculated FPC scores act as our new covariates and are used to conduct the landmark Cox survival analysis, as specified in Equation (4). Then we separately plotted the regression coefficient against the landmark time points with their 95% CI values in Figure 4A,C,E. In order to give a general picture of the FPCs' patterns and their corresponding effects briefly in Table 2. The details of interpretation are as follows. As shown in Figure 4A, for the first FPC score, the regression coefficients start with a low level (negative value) at the beginning, age 40, increase to around zero at age 55, continuously increase to a peak at age 65, and then start decreasing to near zero until age 70. That is to say, subjects with positive values for the first FPC score during early ages 40 to 55 would have lower CHD risks than subjects with negative scores during these ages. Since the first FPC score is the inner product of the centered TC and the first eigen function, a positive value for the first FPC score is corresponding to a below-average TC value. Therefore, the negative regression coefficients for ages 40 to 55 in Figure 4A indicate that a below-average level of TC has protective effects during the younger adult ages. On the other hand, the positive regression coefficients after age 55 in Figure 4A indicate that, with increasing age, the protective effects of below-average TC values vanish. Because the first eigen function and its FPC scores explain 72.0% of total variation, this should be the main conclusion of our analysis.

As shown in Figure 4C, the regression coefficient for the second FPC score starts with a significantly positive value at age 40, and decreases to and stays at nonsignificantly negative after around age 55. Figure 3 shows that the second eigen function has an initial decreasing and later increasing pattern. Following a similar argument as above, from these two plots, we conclude that patients whose TC values follow such a pattern might have higher CHD risks at early ages, but not old age. Figure 4E shows that the regression coefficient for the third FPC score remains around zero throughout all the ages. This implies that a TC profile as indicated by the third eigen function in Figure 3 does not have much effect

FIGURE 3 Plots of three TC eigen functions against the participants' ages for FPCA analysis without mean functions. The first three eigen components plotted using FPCA explain 72.0%, 17%, and 10% of the relevant variation. FPCA, functional principal component analysis; TC, total cholesterol [Colour figure can be viewed at wileyonlinelibrary.com]



on CHD risks. We also imposed a smooth function on the three regression coefficients against the landmark time points with the formula $\beta(L) = \beta_0 + \beta_1 * \log(\frac{L-25}{100})$, as specified in Equation (5), and then plotted three smooth curves, as seen in Figure 4B,D,F. We obtained similar patterns with smoothed curves.

From the above analysis of FPC scores, we were not only able to well capture the mean longitudinal profiles of biomarkers, but also could characterize the variability of individual temporal patterns, leading to a parsimonious representation of longitudinal profile variability. Figure 5 shows the observed longitudinal TC curves of individuals with the top 5% and bottom 5% of each of the three FPC scores. Although the trajectories of different subjects varied among the three FPC scores, we could see from I to III that variation increased a bit sequentially. We could also see the cases present in the top 5% of FPC I scores was lower than the ones present in the bottom 5% of FPC I scores. A consistent pattern across all three FPCs was that the TC levels, starting at a higher level, were associated with more CHD events or higher event percentages; however, higher TC levels occurring in older ages were not associated with a higher CHD risk (eg, as seen from Figure 5A vs B).

In conclusion, the three time-varying coefficient plots exhibit three different risk profiles (Figure 4). The overall trend was consistent with the above landmark analysis using the original data (Figure 2) and our conclusion was also consistent with the results of previous studies, which we discuss in the following section.

At the same time, we compared the prediction performance for the two models in Sections 4.2 and 4.3, the AUC values were calculated at the landmark time points $L = 35, 50, 65, 75$ years in the exam ages, and horizon time is 5 years. The FPCA approach results in higher AUCs (0.6908, 0.6907, 0.6904, 0.6900) than the LVCF method (0.5905, 0.5884, 0.5891, 0.5917). This means the FPCA-based landmark model has a better ability to distinguish between high-risk and low-risk subjects. These results were corroborated by simulation studies in Section 3.

5 | DISCUSSION AND CONCLUSION

Traditional Cox survival analysis fits the proportional hazards model by using baseline biomarkers as covariates, which are measured at a single time point for each individual. This study used FPCA as a robust way to capture the trajectory patterns of longitudinal biomarker data. Then, this summary information can be used as predictors in the traditional Cox survival

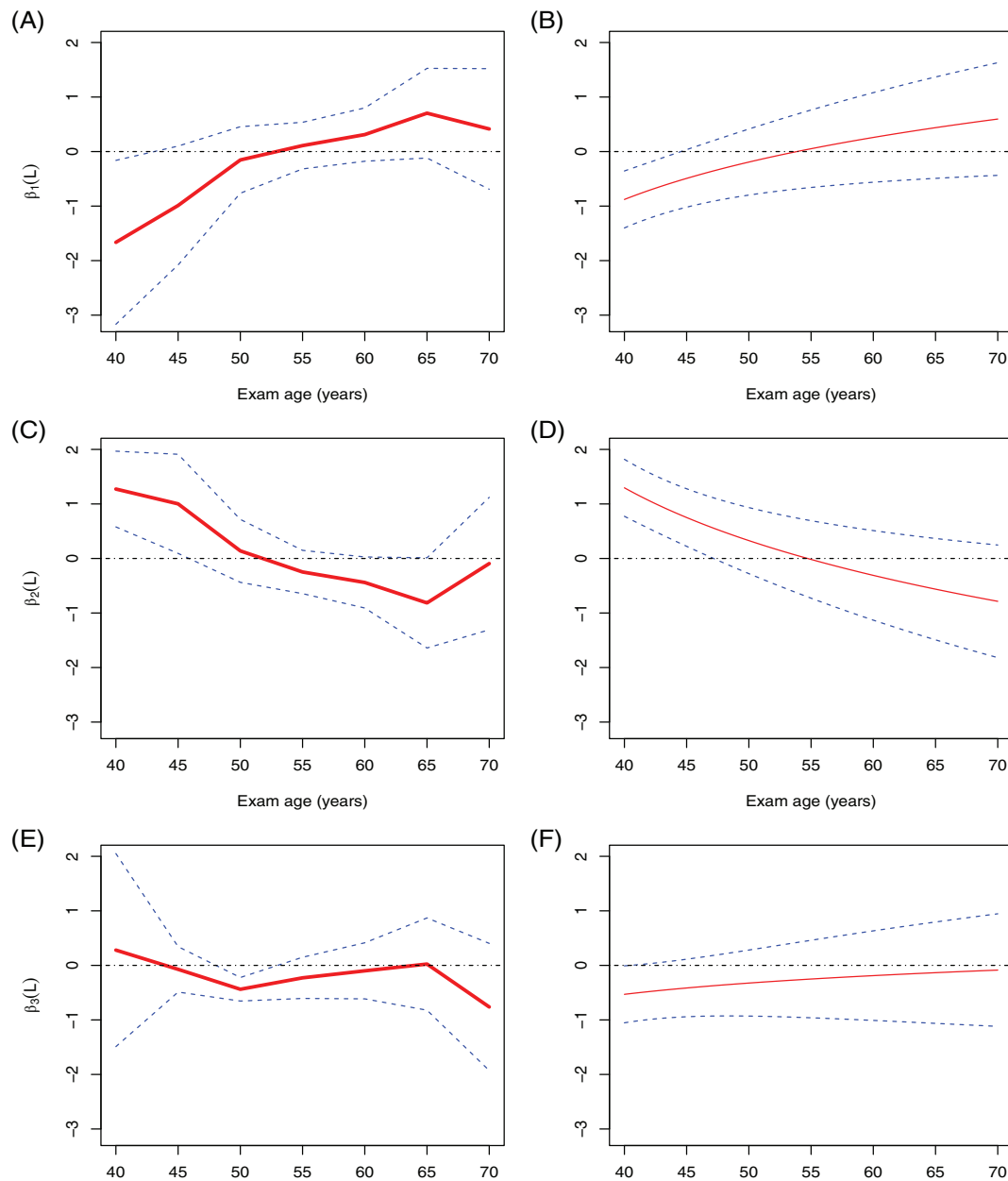


FIGURE 4 A, C, E, Sliding landmark effects with point-wise 95% CIs for coefficients of TC FPCA components I ($\beta_1(L)$), II ($\beta_2(L)$), and III ($\beta_3(L)$), respectively, against exam ages after adjusting baseline covariates gender and smoking status in the Cox model for CHD risks. Each data point and 95% CI multiply the standard deviation of the corresponding functional principal component scores I, II, and III, respectively. B, D, F, Smooth curves on the coefficient of landmark models for FPCA components I, II, and III, respectively, against time using the formula: $\beta_k(L) = \beta_{k0} + \beta_{k1} * \log(\frac{L-25}{100})$, $k = 1, 2, 3$. CHD, coronary heart disease; CI, confidence interval; FPCA, functional principal component analysis; TC, total cholesterol [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Summary for patterns and effects of three FPCs

FPCs	Patterns	Effects
I	Constantly below average	Early age protective and late age no more protective
II	Early decreasing and late increasing	Early age harmful and late age risk-neutral
III	Increasing at 40 and decreasing at 60	No much effect and equal to average level

Abbreviation: FPC, functional principal component.

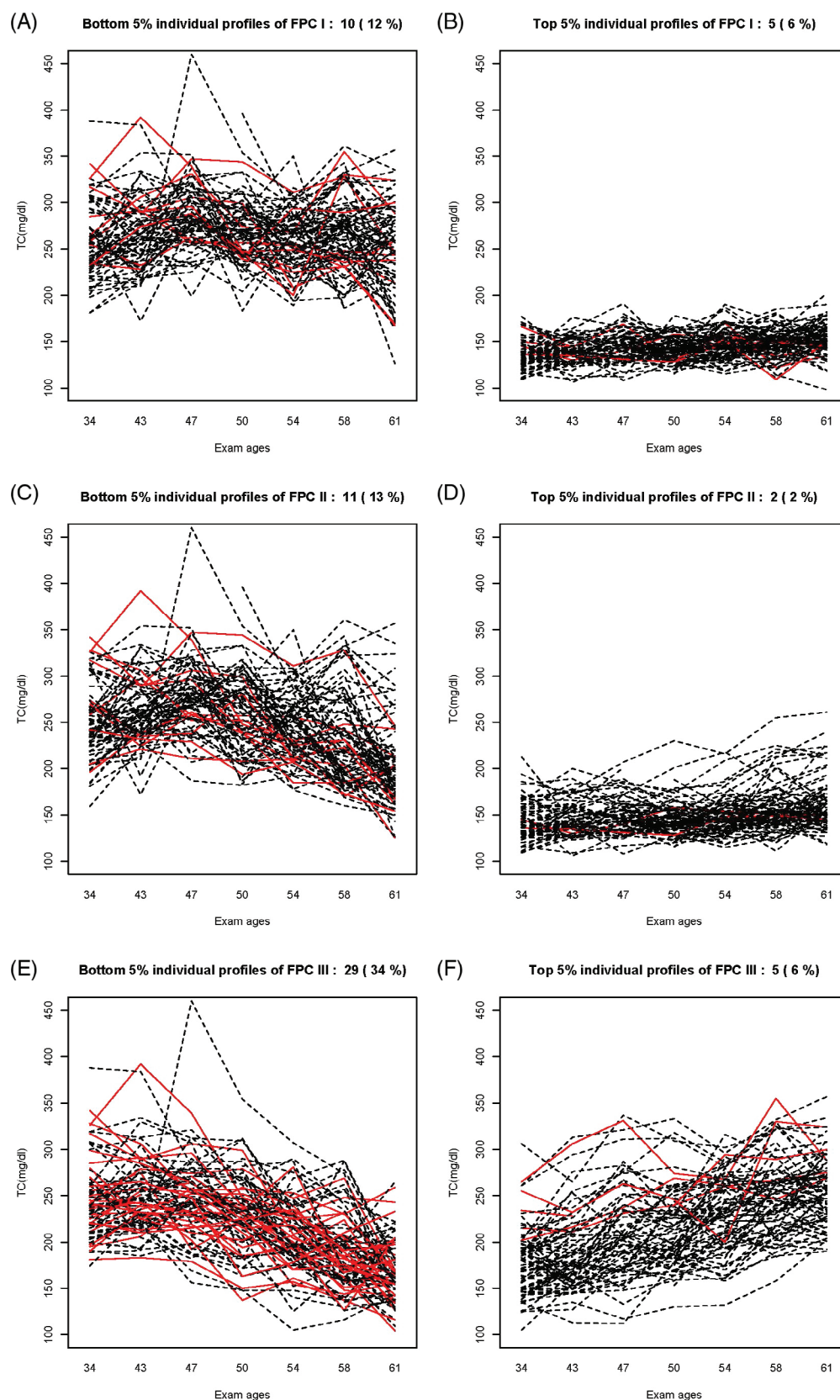


FIGURE 5 Individual TC longitudinal profiles by selecting from top 5% and bottom 5% of FPC scores. A, B, shows individual TC longitudinal profiles selected from top 5%, (B) and bottom 5% (A) of FPC scores I. Correspondingly, there exists a number (percentage) of CHD events (shown in red lines): 10(12%) in the top 5% subjects, and 5(6%) in the bottom 5% subjects. C, D, shows individual TC longitudinal profiles selected from top 5% (D) and bottom 5% (C) of FPC scores II. Correspondingly, there exists a number (percentage) of CHD events (shown in red lines): 11(13%) in the top 5% subjects, and 2(2%) in the bottom 5% subjects. E, F, shows individual TC longitudinal profiles selected from top 5% (F) and bottom 5% (E) of FPC scores III. Correspondingly, there exists a number (percentage) of CHD events (shown in red lines): 29(34%) in the top 5% subjects, and 5(6%) in the bottom 5% subjects. CHD, coronary heart disease; FPC, functional principal component; TC, total cholesterol [Colour figure can be viewed at wileyonlinelibrary.com]

analysis to conduct real-time dynamic prediction of the time to a specific event of interest. We have adapted the landmark analysis by using two different strategies. In one strategy, we first partitioned the datasets based on landmark methods, then we conducted the LVCF to “impute” the current biomarker values. In the second strategy, we first performed FPCA on the whole dataset, then used the obtained FPCA components to project FPC scores onto each separate time interval, which we used as our predictors in survival models. Comparing these two methods, we found the second to be more stable and accurate, though final conclusions are similar. These results make sense as we borrowed information across subjects to estimate the FPCA scores at the landmark time intervals. With more data, prediction accuracy increases and is more stable, which is consistent with the dynamic prediction process. When we have prior knowledge of the whole process of biomarker development, we can obtain more stable and precise predictions. So we focus on the second method in this study.

Numerous observational studies and clinical trials have shown mixed findings regarding the association between TC and CHD in older populations. Our findings contrast with those of some previously published reports. For example, Aronow et al³⁰ performed a prospective study that recruited 708 elderly patients to follow-up with for 41 months to assess the risk factors of CHDs, including serum TC level. The results suggested that serum TC was significantly associated with CHD events in univariate or multivariate analysis. Rubin and colleagues showed similar results in a very large cohort of 2746 white men aged 60 to 79 years.³¹ A recent study showed an increased mortality risk for men with very high HDL-C, but an inverse relationship between HDL-C and CHD events for women.³² On the other hand, our findings were corroborated by a large body of studies.^{2,33,34} Han et al¹⁵ recently reported their work on statin use and CHD risk among older people (>65 or >75 years old). They found that the statin group had significantly lower cholesterol levels in the 6-year follow-up, but no reduced risk of CHD, which is consistent with our analysis of the FHS data. The debates surrounding this issue involve public health problems, but also health economics. As the proportion of the population that is elderly increases world-wide, with particular growth within the United States, the cost of providing daily statin therapy to older adults must be justified by clear benefit. Without a clear benefit, such therapy should be avoided. Our findings demonstrate the importance of analyzing age-dependent effects of TC on CHD risks.

In this article, we applied FPCA for longitudinal continuous covariates. However, FPCA can also be used for longitudinal binary or count data. There exist two approaches to address such issues: covariance decomposition based and probabilistic based FPCA. The covariance decomposition-based approach is the standard and popular way for identifying modes of variation in functional continuous data. Hall et al³⁵ applied the FPCA for binary functional data by positing a smooth latent Gaussian process and then decomposing this process on the variance-covariance matrix of demeaned functional observations. However, when this approach is extended to the exponential family data, as demonstrated in Gertheiss et al, it has an inherent bias due to reliance on a marginal mean estimate rather than conditional mean.³⁶ Probabilistic FPCA is a more appealing and generalized alternative to the covariance smoothing approach for longitudinal categorical variables. This FPCA framework is approached as a likelihood-based model from a Bayesian perspective which easily accounts for sparse or irregular data. At the same time, the probabilistic framework avoids the bias inherent in the covariance decomposition approach. The reasons behind this advantage are that the likelihood approach is often used a link function to relate the expected value of observed data to a smooth latent process, and moreover, all parameters in this approach are estimated simultaneously rather than sequentially. Goldsmith et al extended the probabilistic FPCA via a fully Bayesian approach,³⁷ while Van der Linde applied the generalized FPCA to binary and count data through a variational Bayesian framework which approximates the Taylor's expansion on a log likelihood function.³⁸ In principle, all these techniques can be adapted to our landmark-FPCA framework in the future.

ACKNOWLEDGEMENTS

This research of Bin Shi was supported by the National Institute of General Medical Sciences, National Institutes of Health through grant 5T32 GM074902, and the National Science Foundation through grant DMS 1612965. This research of Peng Wei was partially supported by the U.S. National Institutes of Health grants R01HL116720 and R01CA169122. This research of Xuelin Huang was supported by the USA National Institutes of Health grants U54 CA096300, U01 CA152958, and 5P50 CA100632, and the Dr. Mien-Chie Hung and Mrs. Kinglan Hung Endowed Professorship. The authors declare that there are no conflicts of interest. The authors thank the reviewers for helpful and constructive comments. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This article was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.


DATA AVAILABILITY STATEMENT

We have provided the R codes for performing the proposed FPCA-based landmark analysis and simulation study as online supplemental materials. Due to dbGaP data use policy, we are not allowed to provide the Framingham Heart Study real data; instead, we have provided the real data-based simulated datasets in the example R codes.

ORCID

Bin Shi  <https://orcid.org/0000-0001-5744-4014>

Peng Wei  <https://orcid.org/0000-0001-7758-6116>

Xuelin Huang  <https://orcid.org/0000-0003-1192-9336>

REFERENCES

1. Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. *Ann Trans Med.* 2016;4(13):1-12.
2. Redberg RF, Katz MH. Statins for primary prevention: the debate is intense, but the data are weak. *Jama.* 2016;316(19):1979-1981.
3. Petersen LK, Christensen K, Kragstrup J. Lipid-lowering treatment to the end? a review of observational studies and RCTs on cholesterol and mortality in 80+-year olds. *Age Ageing.* 2010;39(6):674-680.
4. Portela A, Esteller M. Epigenetic modifications and human disease. *Nature Biotechnol.* 2010;28(10):1057-1068.
5. Yan F, Lin X, Huang X. Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *Ann Appl Stat.* 2017;11(3):1649-1670.
6. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol.* 1983;1(11):710-719.
7. Dauxois J, Pousse A, Romain Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J Multivariate Anal.* 1982;12(1):136-154.
8. Besse P, Ramsay JO. Principal components analysis of sampled functions. *Psychometrika.* 1986;51(2):285-311.
9. Boente G, Fraiman R. Kernel-based functional principal components. *Stat Probab Lett.* 2000;48(4):335-345.
10. Yao F, Müller HG, Clifford AJ, et al. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics.* 2003;59(3):676-685.
11. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc.* 2005;100(470):577-590.
12. Jacques J, Preda C. Model-based clustering for multivariate functional data. *Comput Stat Data Anal.* 2014;71:92-106.
13. Chiou JM, Chen YT, Yang YF. Multivariate functional principal component analysis: a normalization approach. *Statistica Sinica.* 2014;24(4):1571-1596.
14. Happ C, Greven S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J Am Stat Assoc.* 2018;113(522):649-659.
15. Han BH, Sutin D, Williamson JD, et al. Effect of statin treatment vs usual care on primary cardiovascular prevention among older adults: the ALLHAT-LLT randomized clinical trial. *JAMA Internal Med.* 2017;177(7):955-965.
16. Cao Y, Rajan SS, Wei P. Mendelian randomization analysis of a time-varying exposure for binary disease outcomes using functional data analysis methods. *Genetic Epidemiol.* 2016;40(8):744-755.
17. Van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scand J Stat.* 2007;34(1):70-85.
18. Zheng Y, Heagerty PJ. Partly conditional survival models for longitudinal data. *Biometrics.* 2005;61(2):379-391.
19. Duong H, Volding D. Modelling continuous risk variables: introduction to fractional polynomial regression. *Vietnam J Sci.* 2014;11:1-5.
20. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS STATA and R programs. *Comput Stat Data Anal.* 2006;50(12):3464-3485.
21. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc.* 1989;84(408):1074-1078.
22. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Stat Med.* 2013;32(13):2262-2277.
23. Wei P, Tang H, Li D. Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genet Epidemiol.* 2014;38(7):638-651.
24. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics.* 2005;61(1):92-105.
25. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw.* 2012;50(11):1-23.
26. Yan F, Lin X, Li R, Huang X. Functional principal components analysis on moving time windows of longitudinal data: dynamic prediction of times to event. *J Royal Stat Soc Ser C (Appl Stat).* 2018;67(4):961-978.
27. Stare J, Perme MP, Henderson R. A measure of explained variation for event history data. *Biometrics.* 2011;67(3):750-759.
28. Tang LL, Balakrishnan N. A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data. *J Stat Plann Infer.* 2011;141(1):335-344.
29. Li L, Wu C. *tdROC: Nonparametric Estimation of Time-Dependent ROC Curve from Right Censored Survival Data. R Package Version 1.0.* New York, NY: John Wiley & Sons, Inc.; 2016. <https://CRAN.R-project.org/package=tdROC>.
30. Aronow WS, Herzig AH, Etienne F, D'Alba P, Ronquillo J. 41-month follow-up of risk factors correlated with new coronary events in 708 elderly patients. *J Am Geriatr Soc.* 1989;37(6):501-506.

31. Rubin SM, Sidney S, Black DM, Browner WS, Hulley SB, Cummings SR. High blood cholesterol in elderly men and the excess risk for coronary heart disease. *Ann Internal Medic.* 1990;113(12):916-920.
32. Wilkins JT, Ning H, Stone NJ, et al. Coronary heart disease risks associated with high levels of HDL cholesterol. *J Am Heart Assoc.* 2014;3(2):1-7.
33. De Ruijter W, Westendorp RG, Assendelft WJ, et al. Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study. *BMJ.* 2009;338:1-8.
34. Krumholz HM, Seeman TE, Merrill SS, et al. Lack of association between cholesterol and coronary heart disease mortality and morbidity and all-cause mortality in persons older than 70 years. *Jama.* 1994;272(17):1335-1340.
35. Hall P, Müller HG, Yao F. Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J Royal Stat Soc Ser B (Stat Methodol).* 2008;70(4):703-707.
36. Gertheiss J, Goldsmith J, Staicu AM. A note on modeling sparse exponential-family functional response curves. *Comput Stat Data Anal.* 2017;105:46-52.
37. Goldsmith J, Zipunnikov V, Schrack J. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics.* 2015;71(2):344-353.
38. Van Der Linde A. Variational Bayesian functional PCA. *Comput Stat Data Anal.* 2008;53(2):517-533.
39. Bosq D. *Linear Processes in Function Spaces.* New York, NY: Springer; 2000.
40. Mas A. Weak convergence for the covariance operators of a Hilbertian linear process. *Stoch Process Their Appl.* 2002;99(1):117-135.
41. Li H, Staudenmayer J, Carroll RJ. Hierarchical functional data with mixed continuous and binary measurements. *Biometrics.* 2014;70:802-811.
42. Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *J Royal Stat Soc Ser B (Methodol).* 1991;53(1):233-243.

How to cite this article: Shi B, Wei P, Huang X. Functional principal component based landmark analysis for the effects of longitudinal cholesterol profiles on the risk of coronary heart disease. *Statistics in Medicine.* 2021;40:650–667. <https://doi.org/10.1002/sim.8794>

APPENDIX A. FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

For performing FPCA on the longitudinal biomarkers from FHS datasets, we assume that all the biomarkers change smoothly and continuously over time. Then we model individuals' biomarker trajectories as independent realizations from a stochastic process $Z(t)$. We treat the underlying $Z(t)$ as a real-valued, square integrable function, which belongs to the separable Hilbert space $L^2[0, \tau]$, where $t \in (0, \tau)$, and τ is the maximum follow-up time, which is a compacted time interval in $[0, \infty)$. For the observed data, let Y_{ij} be the observed biomarker of the j th time point observations from the i th subject across random time interval $t_{ij} \in [0, \tau]$. t_{ij} is for $j = 1, 2, \dots, n_i$, where n_i is the maximum observed time point for i th subject, and subject $i = 1, 2, \dots, n$. Thus, $Z_i(t)$ is the underlying biomarker trajectory of subject i , and often not directly observable due to measurement errors, but must be reconstructed from noisy observations Y_{ij} . Therefore, we model each longitudinal biomarker as follows:

$$Y_{ij} = Z_i(t_{ij}) + \varepsilon_{ij}, \quad (\text{A1})$$

where ε_{ij} are independent measurement error terms with $E(\varepsilon_{ij}) = 0$ and $\text{Var}(\varepsilon_{ij}) = \sigma^2$.

Given $Z_i(t)$ defined as above, we suppose there exists the mean function $E[Z_i(t)] = \mu(t)$ and covariance function $E[(Z_i(t) - \mu(t)) * (Z_i(v) - \mu(v))] = G(t, v)$, for t, v in $[0, \tau]$. As noted, FPCA relies on an expansion of the data in terms of the eigen functions of the covariance function $G(t, v)$. Mercer's theorem³⁹ guarantees that there exists an orthonormal basis for decomposition of a symmetric and nonnegative definite covariance function, such that: $G(t, v) = \sum_{k=1}^{\infty} \lambda_k \rho_k(t) \rho_k(v)$, where the eigen functions: $\rho_k(t)$, $0 \leq t \leq \tau$, and the eigen values: $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$, for $k = 1, 2, \dots, \infty$. The eigen functions $\{\rho_k(t)\}_{k=1}^{\infty}$ are satisfied as follows:

$$\int (\rho_k(t))^2 dt = 1, \text{ and } \int \rho_k(t) \rho_j(t) dt = 0, \text{ for any } j \neq k, j, k = 1, 2, \dots, \infty. \quad (\text{A2})$$

Correspondingly, the individual biomarker temporal profile $Z_i(t)$ can be expanded by the Karhunen-Loeve representation⁴⁰ in term of this orthonormal basis, as follows:

$$Z_i(t) = \mu(t) + \sum_{k=1}^{\infty} \gamma_{ik} \rho_k(t), \quad i = 1, 2, \dots, n, \quad \text{and } t \in [0, \tau], \quad (\text{A3})$$

where $\gamma_{ik} = \int_0^{\tau} \{Z_i(t) - \mu(t)\} \rho_k(t) dt$ is the k th FPCA score, which are uncorrelated random variables with a mean of 0 and a variance λ_k .

To achieve dimensional reduction, each sample function can be sufficiently well approximated, without losing much variation, by the first M components, which changes the notation from $\sum_{k=1}^{\infty} \gamma_{ik} \rho_k(t)$ to $\sum_{k=1}^m \gamma_{ik} \rho_k(t)$, where $m = 1, 2, \dots, M$. Then the model we consider is

$$Z_i(t) = \mu(t) + \sum_{k=1}^M \gamma_{ik} \rho_k(t). \quad (\text{A4})$$

To choose M , which is the number of eigen functions needed to provide a reasonable approximation for the infinite-dimensional process, we may base it on the fraction of variance, as explained as at least 95%, as suggested by Reference 41. Alternately, M can be selected using cross-validation methods based on the minimized prediction error⁴² or some other methods, such as Akaike or Bayesian information criterion. In practice, the choice of the basis is decided by the shape of the curve. Usually, the basis should be chosen according to the following criterion: the expansion by eigen functions approximates the curve itself as closely as possible. The common choices for the function basis include polynomial, B-splines, wavelets, and the Fourier basis.

The above FPCA framework for functional data is a flexible method for capturing the trajectories of longitudinal biomarkers. The underlying random function curves can be reconstructed as a linear combination of the orthonormal basis, defined by the eigen functions of its covariance matrix with random coefficients γ_{ik} considered as the weight function for each basis. This is the so-called data projection or transformation from the original correlated scale $Z_i(t)$ to the new orthogonally coordinated system. However, in order to estimate such a process, the time grids of longitudinal data need to be dense and regular. In reality, functional data are usually a collection of noise-corrupted observations. Data measurements can be sparse, error-prone, and taken at irregular time points. To deal with this incongruence, Yao et al¹¹ proposed to use PACE to obtain smooth estimators. In this study, we use the same strategy to estimate the FPCA model components. The mean, covariance, and eigen functions are assumed to be smooth. For estimating the mean function $\hat{\mu}(t)$, we use one-dimensional, local linear kernel smoothers to estimate the function surface, which is obtained by weighted least squares. The estimated covariance $\hat{G}(t, v)$, given $t, v \in [0, \tau]$, is obtained by fitting a two-dimensional kernel smoother with all pairwise products $\{Y_{ij} - \mu(t_{ij})\} \{Y_{ih} - \mu(t_{ih})\}$ for $j \neq h$ in the i th subject. The estimates $\hat{\rho}_k(t)$ and $\hat{\lambda}_k$ of eigen functions and eigen values that correspond to the solutions are obtained by decomposing the smoothed covariance through the following eigen equation:

$$\int_0^{\tau} \hat{G}(t, v) \hat{\rho}_k(t) dt = \hat{\lambda}_k \hat{\rho}_k(t). \quad (\text{A5})$$

Traditionally, FPC scores γ_{ik} have been estimated by numerical integration. It works well when the grid density of the measurements for each subject is sufficiently large. However, when data are sparsely and irregularly measured or contaminated with measurement errors, PACE provides the best way to estimate them. Briefly, PACE carries out FPCA under the normal assumption. This approach is necessary if the number of measurements per subject is too small, such as one or two observations per subject. In order to obtain a good approximation for the following integral,

$$\hat{\gamma}_{ik} = \int_0^{\tau} \{Z_i(t) - \hat{\mu}(t)\} \hat{\rho}_k(t) dt. \quad (\text{A6})$$

The approach is as follows: the quantity $\hat{\mu}(t)$ and $\hat{\rho}_k(t)$ are estimated using the entire set of observed data $\{Y_{ij}, i = 1, \dots, n, j = 1, \dots, n_i\}$ and borrowing strength from the data on all subjects. As suggested by Yao, assuming γ_{ik} and ε_{ij} to be jointly Gaussian and predicting the random effects γ_{ik} based on its conditional expectation,¹¹ we have $\hat{\gamma}_{ik} = \hat{E}(\gamma_{ik}|Y_i)$. Specifically, let $Y_i = (Y_{i1}, \dots, Y_{in_i})'$, $\hat{\mu}(\vec{t}_i) = (\hat{\mu}(t_{i1}), \dots, \hat{\mu}(t_{in_i}))'$, and $\hat{\rho}_k(\vec{t}_i) = (\hat{\rho}_k(t_{i1}), \dots, \hat{\rho}_k(t_{in_i}))'$, the best linear prediction of FPCA scores is

$$\hat{\gamma}_{ik} = \hat{E}(\gamma_{ik}|Y_i) = \hat{\lambda}_k \hat{\rho}_k(\vec{t}_i) \hat{\Sigma}_{Y_i}^{-1} (Y_i - \hat{\mu}(\vec{t}_i)), \quad (\text{A7})$$

where $\hat{\Sigma}_{Y_i} = \text{cov}(Y_i, Y_i) = \text{cov}(Z_i, Z_i) + \sigma^2 I_{N_i}$, $Z_i = (Z_{i1}, \dots, Z_{in_i})'$. That is to say, the (j, h) entry of the $N_i \times N_i$ matrix $\hat{\Sigma}_{Y_i}$ is $(\hat{\Sigma}_{Y_i})_{j,h} = \hat{G}(t_{ij}, t_{ih}) + \sigma^2 \delta_{jh}$, with $\delta_{jh} = 1$ if $j = h$, and 0 otherwise. Under a Gaussian assumption, this method provides the best linear unbiased prediction estimator. However, it is a robust estimation, irrespective of whether the Gaussian assumption holds.

APPENDIX B. DESIGN PLOT

From the design plot (Figure 6) of TC, we can see the total number of data points are sufficiently dense to cover the surface of variance-covariance matrix. Because the maximum follow-up time was 30 years, we have $\max|T_{ij} - T_{ik}| = 30$. Therefore, the design plot is blank in the area specified by $|y - x| > 30$.

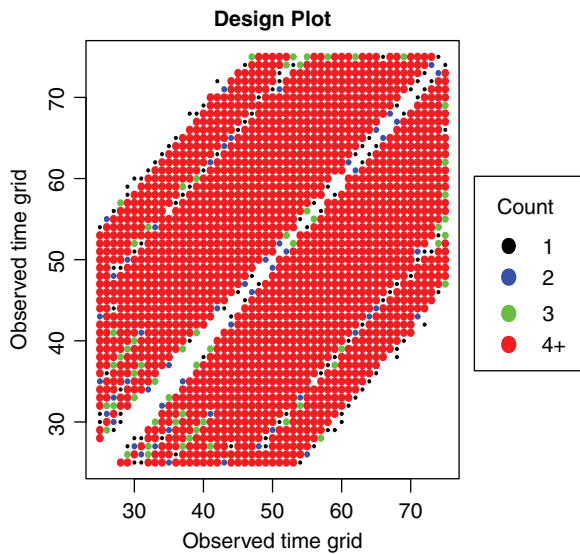


FIGURE 6 Scatter plot for all the pairs $(T_{ij}; T_{ik})$ of time grids of TC values in FHS data, where $i = 1, \dots, n$, and $j, k = 1, \dots, n_i$. FHS, Framingham Heart Study; TC, total cholesterol [Colour figure can be viewed at wileyonlinelibrary.com]

APPENDIX C. SELECTION OF M IN FPCA

As shown in Table 3, when we let M vary from 1 to 6, the explained variance increased from 68.6% to 99.9% correspondingly. We noted that using the percentage of variance explained at least 95% selected the true number $M = 3$ across all 500 simulation replications. We found that the Brier scores were stable when $M \leq 3$, then increased dramatically when $M > 3$. However, the AUC values were quite similar for all selected M values. We can see that the Brier scores were more sensitive than the AUC values with respect to the choice of M . This is because the Brier score measures the mean squared difference between the true and the predicted survival probabilities (so-called calibration error), while the AUC value, a rank-based measure, is less sensitive to perturbations.

TABLE C1 Selection of M in FPCA and its effect on predictive performance in terms of Brier score and AUC

M	Selection of M	LM-FPCA	
	Variation explained (%)	Brier scores (1e-2)	AUC values (1e-1)
1	68.60	0.479	6.379
2	90.40	0.491	6.704
3	96.00	0.489	6.856
4	99.80	1.057	6.769
5	99.85	1.781	6.536
6	99.90	5.981	6.538

Abbreviations: AUC, area under receiver operating characteristic curve; FPCA, functional principal component analysis; LM, landmark.