The Texas Medical Center Library

DigitalCommons@TMC

Spring 3-2019

Dynamic Prediction Of Survival Data Using Single Or Multiple Longitudinal Markers

Xuehan Ren
UTHealth School of Public Health

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen

Part of the Community Psychology Commons, Health Psychology Commons, and the Public Health Commons

Recommended Citation

Ren, Xuehan, "Dynamic Prediction Of Survival Data Using Single Or Multiple Longitudinal Markers" (2019). *Dissertations & Theses (Open Access)*. 77.
https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/77

This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in Dissertations & Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digcommons@library.tmc.edu.

DYNAMIC PREDICTION OF SURVIVAL DATA USING SINGLE OR MULTIPLE LONGITUDINAL MARKERS

by

Xuehan Ren, BS

APPROVED:

Sheng Luo, PHD

Hulin Wu, PHD

Momiao Xiong, PHD

John M. Swint, PHD

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

Copyright

by

Xuehan Ren, BS, PHD

2019

DEDICATION

To my families and friends.

DYNAMIC PREDICTION OF SURVIVAL DATA USING SINGLE OR MULTIPLE LONGITUDINAL MARKERS

by

Xuehan Ren
BS, University of Science and Technology of China, 2013

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
March, 2019

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to my supervisor Dr.Sheng Luo for inspiring my research work and guiding me with a lot of effort, discussing and encouraging during the past few years. It would not be possible to write this dissertation without his guidance. He has spent his precious time to serve as my advisor and helped me with the challenges I encountered during dissertation research. I would also like to show my gratitude to my advisory committee members. My minor advisor Dr. John M. Swint and breadth advisor Dr. Momiao Xiong have kindly supported my research and made inputs to my dissertation. Thanks also goes to my dissertation chair and academic advisor Dr. Hulin Wu for his consistent support throughout my Ph. D program, along with his suggestions in the dissertation work. Last but not lease, I would like to thank my friends for supporting me and encouraging me as always. Special thanks to my parents and my husband for their unconditional love and support for so many years. All of them made the completion of this long journey an enjoyable experience.

Dynamic prediction of survival data using single or multiple longitudinal markers

Xuehan Ren, BS, PhD

The University of Texas

School of Public Health, 2019

Dissertation Supervisor: Sheng Luo, PhD

Recurrent events and time-to-event data occur frequently in longitudinal studies. In large clinical trials with survival endpoints, researchers collect a multitude of longitudinal markers. There is a growing need to utilize these rich longitudinal information to build prediction models and assess their prognostic performance. In this dissertation research, I propose a novel approach of integrating longitudinal markers in modeling the recurrent event or terminal event data, and conduct dynamic prediction of event risks. Under joint a model framework, I jointly model a longitudinal outcome and a recurrent event process with the two process correlated via shared latent function. The probability of having a new occurrence of recurrent event in a given time interval is predicted based on subject-specific longitudinal profile and disease history. When multivariate longitudinal outcomes are considered, traditional joint model method has limitation on specifying appropriate longitudinal structures and computation problem occur when using Bayesian approach. To avoid these potential issues, I employ multivariate functional principal component analysis approach which is more flexible, robust and time efficient. For terminal event data, I specify a prognostic model incorporating multivariate longitudinal information, the prediction can be updated with accumulated data over time. I also propose a recurrent event model integrating multiple longitudinal markers and conduct personalized dynamic prediction of new recurrent event risk, which helps physicians to identify patients at risk and give personalized health care.

Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Literature Review

1.1.1 Cardiovascular disease

Cardiovascular disease (CVD), a leading cause of death and disability in both men and women worldwide, causes one in every four deaths in United States and has raised a major public health concern[1]. The estimated annual costs for CVD reached 207 billion dollars in health care area in the United States [2]. Understanding the risk factors for CVD yields important insights into prevention, treatment and prediction of disease progression. To this regard, numerous clinical studies were funded by National Heart, Lung, and Blood Institute (NHLBI) aiming to identify risk factors for predicting future CVD events and further more, to develop and assess performance of different risk prediction models. For instance, NHLBI has sponsored Anti-hypertensive and Lipid-lowing Treatment to Prevent Heart Attack Trial (ALLHAT) study [3], Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium which includes 3 studies: Framingham Heart Study (FHS), Cardiovascular Health Study (CHS), and Atherosclerosis Risk in Communities (ARIC) [4]. Despite having different Cardiovascular disease as study outcome, there are commonality among the aforementioned 4 clinical trials: (1) Study participants with higher risk may experience several cardiovascular events including coronary heart disease (CHD), stroke, and heart failure during the whole study. These events occur repeatedly during their follow-up time and are correlated to each other, generating recurrent event outcome; (2) longitudinal biomarkers of participants are available in scheduled visits during their follow-up time, including systolic blood pressure, total cholesterol which were proved to be predictive in cardiovascular disease events [5, 6].

Regardless of numerous research on inference or prediction of the survival risk in cardiovascular disease area, these studies either employ simple survival models or logistic models that fail to utilize the complete longitudinal health outcome trajectories that are predictive of CVD, or fail to account for the recurrent nature of the CV event occurrence. Alternatively, leveraging frequently measured longitudinal health biomarkers and survival outcome, we propose to develop a personalized dynamic prediction model using these CVD study datasets. Our proposed model not only provides more accurate personalized predictions but also is able to dynamically update the prediction results upon the availability of updated subject-specific data in new visits.

1.1.2 Parkinson's disease

Parkinson's disease (PD) is a chronic progressive neurodegenerative disorder manifested as tremor, rigidity, slow movement, and impaired balance in clinics, affecting about 1% of adults older than 60 years. PD is caused by the malfunction and death of neurons, which are the nerve cells in the brain. Part of the the malfunctioned and dying neurons produce dopamine, which is a chemical that helps control movement and coordination of human body.

Previous literature studied a broad range of motor and non-motor symptoms which are clinically correlated with evolution of Parkinson's disease [7–9]. However, due to substantial heterogeneity of different symptoms and subjects, it is very challenging to give accurate prognoses of disease progression. In the absence of a cure, there is a critical need to identify significant and well-validated biomarkers associated with PD progression [10]. As a complex progressive disease, a variety of endpoints have been established to evaluate PD severity and work as a criteria to categorize subjects into different disease stages. For example, Unified PD Rating Scale (UPDRS), modified Hoehn and Yahr (HY) scale, and Ambulatory capacity evaluate patients' movement ability [11, 12]. Among these measures, Hoehn and Yahr (H&Y) scale has became one of the most commonly and widely used measurement to assess overall PD dysfunction stage [13]. Schrag *et al.* [14] examined the responsiveness of different PD progression outcome measures over time and suggested the H&Y scale to be the most responsive measure. Based on motor functions, the H&Y scale has served as a good endpoint of PD progression in many published studies. The time to development of different H&Y stages from symptom onset or enrollment has been widely studied as a way to evaluate disease progression in past literature [15–18]. Although these literature discussed potential prognostic factors in predicting the H&Y stage progression, only baseline variables were included in the survival model and longitudinal information was not taken into account. Relatively few studies considered modelling longitudinal marker in prediction of PD disease outcome. He and Luo [19] proposed a joint model with multilevel item response theory sub-model for the longitudinal data and Cox propotional hazard sub-model to handle time to terminal event. They assessed the effect of tocopherol on time to initiation of symptomatic therapy in early PD patients. Iddi *et al.* [20] applied a latent time joint mixed-effects model to handle longitudinal outcomes and studied the association between different markers and PD diagnostic category. However, to our best knowledge, no previous research has been done regarding prognostication of time to the H&Y stage transition based on multiple longitudinal markers.

1.1.3 Recurrent Event Data Analysis

Many disease and clinical study outcomes may reoccur in the same participant, which makes the investigation of multiple time-to-event data for one subject important. Examples of recurrent events include recurrent strokes in elderly patients, recurrent heart failure, and recurrent heart attack [21]. An important feature of these events is that there exists intrinsic correlation between those repeated occurrences within the same subject. Ignoring this unique feature of correlation of recurrent events, the estimated event occurrence rates could be biased [21]. Hence, the most commonly used Cox proportional hazard model in biostatistics and epidemiological field will not be appropriate here, because they do not consider the correlation of the repeated occurrences.

Numerous researchers have conducted statistical research in recurrent event data analysis field. Vaida and Xu [22] proposed a general proportional hazards model with random effects for handling clustered recurrent event data. Pepe and Cai [23] have proposed semi-parametric procedures for making inferences about the mean and rate function of the counting process without the Poisson-type assumption . An approach to constructing simultaneous confidence bands for the mean function was presented by Lin *et al.* [24].

Recurrent event data has two important features. One is that the recurrent events are ordered (the second event can only occur after the first event), while another feature is that the subject can only be at risk for one event at a time (i.e. within a small time interval Δt, only one event can happen) [21]. A common approach to model the recurrent event data is to assume that the event process is poisson process so that the number of events in disjoint time intervals is assumed to be independent [25]. The key aspect of analyzing recurrent event data is to model the intensity function. Poisson process models intensity function as a function of calender time. Assuming that n events occur in time interval [0,τ], we denote the event times as $t_1, t_2, ...t_{n-1}, t_n \leq \tau$. Therefore, the total number of events up to time t is $N(t), t > 0$. Let $H(t) = \{N(s) : 0 \leq s < t\}$ denote the event process history and information on the covariate process up to time t, we can write the intensity function of the process $\lambda(t, H_i)$ as below:

$$\lambda(t|H_i(t)) = \lim_{\Delta t \to 0} \frac{Pr\{\Delta N(t) = 1 \mid H_i(t)\}}{\Delta t} = \lim_{\Delta t \to 0} \frac{Pr\{N(t + \Delta t^-) - N(t^-) = 1\}}{\Delta t}, \quad (1.1)$$

where $\Delta N(t)$ denotes the number of events over a small interval $[t, t + \Delta t)$. Denoting $\lambda(t|H_i(t)) = \lambda_i(t)$, the likelihood contribution from each subject in time period $[0, \tau]$ can be written as :

$$\left[\prod_{j=1}^{n_i} \lambda_i(t_j)\right] \times \exp\left[-\int_0^\tau \lambda_i(t)dt\right]. \quad (1.2)$$

1.1.4 Overview of Dynamic Prediction via Joint Models

Joint models (JM) of longitudinal measurements and survival data have been widely studied in the past two decades (first proposed by Faucett & Thomas [26] and Wulfsohn & Tsiatis [27]), there is an increasing number of literature investigated the application of joint models to large clinical studies, e.g., Henderson *et al.* [28], Han *et al.* [29] and Crowther *et al.* [30]. Usually, a joint model involves two sub-models, one is a mixed effects model for longitudinal outcomes while the other is a Cox model for survival events. To link the two sub-models together, shared random effects are usually utilized. Previous literature have compared and investigated the association structure for random effects [31]. For inference purpose, joint models can be used to estimate the parameters associated with the longitudinal models as well as the survival models. Although there exists many literature discussing joint model of longitudinal markers and time-to-event data, few studies has investigated the joint model of longitudinal markers and recurrent event data. Henderson [28] proposed a joint model of longitudinal data and recurrent event data, with association between two process captured by correlated latent trajectory. Liu and Huang [32] established a JM approach where a more complex setting was considered, i.e. a repeated measures process and a recurrent events process were correlated, both subject to a terminal event. In addition to model inference, a novel usage of JM is to provide dynamic prediction of the risk of target event and the trajectories of biomarkers. Rizopoulos [33] has proposed a Monte Carlo approach to conduct dynamic prediction on time-to-event data, which predicts subject-specific survival probability using the joint model framework. The key feature of the dynamic prediction framework is that the prediction can always be dynamically updated given the additional information of longitudinal trajectories and longer history of survival process. There are some literature of the dynamic prediction problem under joint model of longitudinal and survival data framework [34, 35]. Regarding the prediction for recurrent event data, Krol *et al* [36] extended the usage of joint model to include recurrent event process, and estimated the probability of having a terminal event in specific time interval, given historical longitudinal data and recurrent event times. An alternative approach was proposed by Musoro *et al* [37] by employing landmark method to handle the longitudinal data as time-fixed covariate at different landmark time points, the dynamic prediction by landmarking was then extended to recurrent event data in this way. However, to our best knowledge, providing personalized dynamic prediction of recurrent events by joint model method remains an open question. To propose a prediction model that incorporates clinical information and the recurrent disease history is of importance here. In this dissertation, we develop a Bayesian joint model which consist of two sub-models, one is the linear mixed effect model to model longitudinal trajectory, another is recurrent event model with intensity function from Poisson process to handle the recurrent event process. The dynamic prediction of recurrent events is derived and implemented in application datasets.

1.1.5 Multivariate Principal Component Analysis

In most literature, dynamic predictions via joint models have been restricted to only include single longitudinal outcome. However, emerging evidence has suggested that many diseases are correlated with multiple longitudinal clinical outcomes. Moreover, in large clinical studies, rich clinical information including observations of multiple longitudinal biomarkers is collected. Incorporating these clinical information in prediction of event of interest becomes an urgent need. Extending the JM approach to incorporate multivariate longitudinal outcomes has a major limitation that we need to specify appropriate parameter distributions and correlation structure between longitudinal processes. In addition, the computation intensity is another concern when utilizing joint model under Bayesian framework, especially with large sample size. As the number of candidate longitudinal markers increase, the computation cost increases exponentially. To handle the multivariate longitudinal outcomes and avoid involving in the aforementioned issues, we consider an alternative approach which is more flexible, robust and computationally efficient. Yao [38] first proposed a nonparametric approach to perform functional principal components analysis (FPCA) on sparse longitudinal data. Since then, researchers has extended the usage of FPCA to joint analysis of repeated measurements and survival data [39]. In the two-step approach proposed by Holte *et al* [40], the feature of longitudinal trajectory was extracted and represented by the functional principal component (FPC) scores estimated from separate model in step one, and the estimated FPC scores were included in the survival model as new risk factors to build association between longitudinal trajectory and the survival outcome. In regards of prediction, Yan *et al* [41] extended the FPCA framework further to dynamic prediction area. Their work updates the estimated FPCA scores as new longitudinal information come into available. However, these studies only incorporated a single longitudinal marker in the survival model. Moreover, the Cox model they used to model the terminal event is not suitable when the event of interest has recurrent feature (e.g. CVD events). To our best knowledge, there is no existing literature has investigated the prediction of recurrent event utilizing information from multiple longitudinal markers. In this dissertation work, we develop an novel approach to fit this critical need.

1.2 Public Health Significance

Even with the contemporary medical techniques nowadays, cardiovascular disease still acts as a leading cause of death and disability in both men and women in United States and worldwide. It causes one in every four deaths and has became a major public health concern. Investigation in risk factors is a key step to cardiovascular disease treatment and prevention. There is strong evidence suggesting that persons with healthier lifestyle could significantly reduce the risk of incidence in cardiovascular events (e.g. quit smoking, avoiding obesity, exercise routinely, consume green food, and keep a healthy diet) [42]. Therefore, identifying and quantifying risk factors is of great interest in modern public health area.

While discovering significant and well-validated risk factors, predicting risk of the future event is another important mission. Accurate prediction of the risk of a subject's future cardiovascular event in a specific given time period enables physicians to make personalized diagnostic and treatment. Based on subject-specific longitudinal profiles and updated event history, the proposed work can help physicians target people with higher risk of developing a CVD events in a near future. This task helps address the critical need for model-based personalized dynamic predictions of future longitudinal health outcome trajectories and CVD events.

Chapter 2

Article 1: Dynamic prediction using joint models of longitudinal and recurrent event data: A Bayesian perspective

Dynamic prediction using joint models of longitudinal and recurrent event data: A Bayesian perspective

2.1 Introduction

As a leading cause of death and disability in both men and women worldwide, cardiovascular disease (CVD) causes one in every four deaths in United States and has became a major public health concern [1]. The estimated annual costs for CVD reached 207 billion dollars in health care area in the United States [2]. It is of importance to understand the risk factors for CVD which yields important insights into prevention, treatment and prediction of disease progression. National Heart, Lung, and Blood Institute (NHLBI) has funded many clinical studies aiming to identify risk factors for predicting future CVD events. Anti-hypertensive and Lipid-lowing Treatment to Prevent Heart Attack Trial (ALLHAT) study [3] is one of the largest CVD studies that sponsored by NHLBI. Previous literatures have identified hypertension, high cholesterol level, diabetes etc. are highly related to risk of developing cardiovascular disease [43, 44].

An important feature of cardiovascular disease is that the primary event outcomes are often recurrent, i.e. patients can experience multiple CVD events during the follow-up period. Examples of these recurrent events include strokes, reoccurred heart failure, and recurrent heart attack [21]. An important feature of these events is that there exists intrinsic correlation between those repeated occurrences within the same subject. Ignoring this unique feature of correlation of recurrent events, the estimated risk could be biased [21]. Hence, the commonly used Cox proportional hazard model will not be appropriate here, due to the fact that correlation of the repeated occurrences is not considered. As the recurrent event process tend to be related to longitudinal markers, analyzing the two processes separately may lead to biased estimation. Thus, to propose an statistical approach that jointly models longitudinal markers and recurrent events together is of importance here.

First proposed by Faucett & Thomas [26] and Wulfsohn & Tsiatis [27], joint models (JM) of longitudinal measurements and survival data (including terminal event and recurrent event data) have been widely studied in the past decades and a increasing number of studies applied joint models to large clinical studies, e.g., Henderson *et al.* [28], Han *et al.* [29] and Crowther *et al.* [30]. A common joint model often involves two sub-models, one is a mixed effects model for longitudinal outcomes while the other is a Cox model for survival events. In such setting, shared random effects are usually utilized to link the two sub-models together. Previous literature compared and investigated the association structure for random effects [45]. Some extensions of the joint models have been developed, e.g., relaxing some of the normality assumptions of random effects in order to make a more general assumption [31].

Another novel usage of JM is to provide dynamic prediction of the risk of target event and the trajectories of biomarkers. Rizopoulos [33] has developed methodology to make personalized predictions using the joint model framework. The prediction is dynamic in that the prediction can always be dynamically updated given the additional information of longitudinal trajectories and longer history of survival process. Although there are some literatures of the dynamic prediction problem under joint model of longitudinal and survival data framework [34, 35], to our best knowledge, providing personalized dynamic prediction in recurrent events in large CVD studies remains an open question. Numerous prediction models estimating the risk of developing CVD events are either simple survival model or logistic regression models [46]. To propose a prediction model that incorporates clinical information and the recurrent disease history is of importance here. In this article, we develop a Bayesian joint model approach to fit this critical need. The proposed model consists of two sub-models, one is the linear mixed effect model to model longitudinal trajectory, another is recurrent event model with intensity function from Poisson process to handle the recurrent event process. The two sub-models are linked via shared latent trajectory, while a parameter is assigned to access the strength of association between these two process.

The rest of the article is organized as follows. In Section 2, we introduce the motivating cardiovascular disease study that motivates this article. The recurrent nature of the CVD events is visually displayed. The joint model framework and dynamic prediction method are illustrated in Section 3, upon which we propose our approach to model recurrent event process with longitudinal trajectory. We conduct simulations in Section 4 to validate inference accuracy of our proposed model, the prediction performance is also accessed by utilizing time-dependent Area Under the Curve (AUC) and Brier score (BS) as indexes of prediction accuracy. In Section 5, we apply our proposed model on the motivating dataset, the Parallel MCMC method is employed to address our computation issue caused by large sample size. For prediction purpose, we also specify a traditional recurrent event model (referred as simple recurrent event model later), which only take into account baseline covariates. The prediction performance of our model outperform simple recurrent event model in regards of higher time-dependent AUC, indicating incorporation of longitudinal clinical information improves subject-specific prediction accuracy for risk of new CVD events. In the last section, we summarize the findings in our study, and discuss limitation of our work and some possible future directions.

2.2 Motivating dataset

The methodological research is motivated by Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), a randomized, double-blinded, active-controlled clinical trial conducted from February 1994 through March 2002 [3]. As the primary outcome was composite fatal coronary heart disease (CHD) or non-fatal myocardial infarction (MI), the aim of ALLHAT study is to determine whether newer antihypertensive agents, including α-blocker doxazosin, differ from the reference treatment chlorthalidone in regards of coronary heart disease events and other CVD events. During the study follow-up period, 15,255 subjects were randomized to chlorthalidone arm and 9,061 subjects were randomized to doxazosin arm, out analysis is conducted on subjects from these two arms. The maximum follow-up time in analytical dataset was 5.57 years. As death is treated as independent censoring in our study, subjects who were alive at the end of the trial, or lost to follow-up during the trial are considered as independent termination (right censored). We are interested in the composite outcome of three types of CVD recurrent events (CHD, stroke, heart failure) before death or censoring. To visualize the data structure, we present Figure 2.1 to show the time plot of four selected subjects who experienced composite CVD events before the end of their follow-up. For instance, subject 1 experienced one CVD event in 59 month and was censored on 92 month, while subject 4 experienced 5 recurrent CVD events and died in 67 month. Death process is treated as non-informative censoring as our main interest lies in predicting next recurrent event.

During the study, participants' clinical information was collected and recorded during each scheduled visits. Systolic blood pressure (SBP) is the most frequently measured clinical outcome in ALLHAT study. Utilizing this rich clinical information in predicting the probability of having next recurrent event is of our interest. It is well established that high blood pressure, i.e. hypertension is one of the main risk factors for CVD events, previous literature has studied the mechanisms of how high blood pressure can cause CHD, stroke, heart failure [47]. In terms of prediction of risk of CVD events, hypertension, as a known significant risk factor, is usually included in the statistical model. Staessen

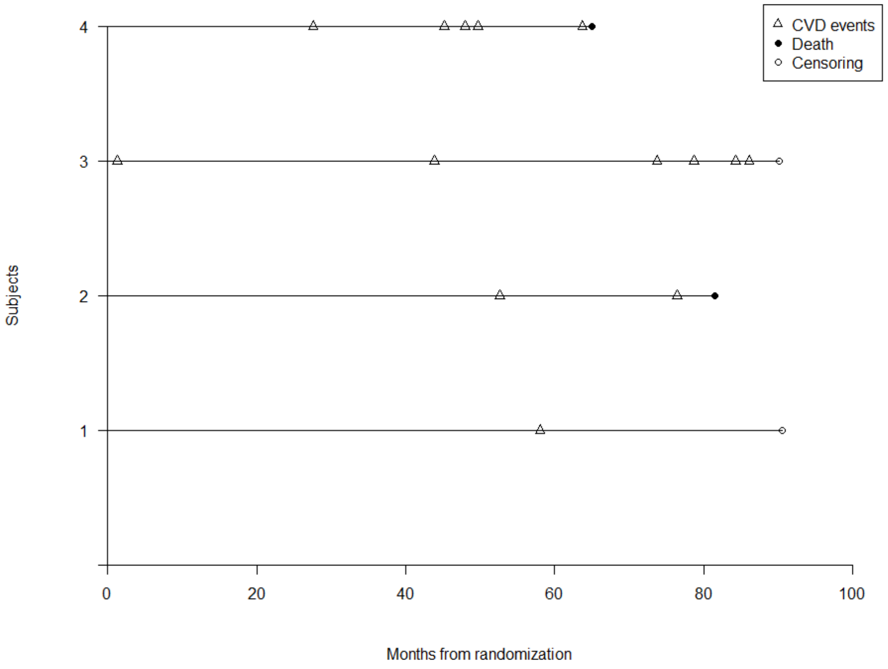


Figure 2.1: Data display of CVD events in ALLHAT study

et al [48] found that in untreated patients with isolated systolic hypertension, the systolic blood pressure predicted cardiovascular risk. More recent publications pointed out there was significant positive relationship between higher SBP and CVD events [34, 46]. On the other hand, subjects with previous CVD have high risks of CVD recurrence [49]. Thus, incorporating the SBP trajectory and recurrent disease history is essential to personalized-prediction of next occurrence of a CVD event. Of the existing literature about joint modeling of longitudinal measurements and recurrent events, Henderson [28] proposed a joint model of longitudinal data and recurrent event data, with association between two process captured by correlated latent trajectory. Moreover, Liu and Huang [32] considered a more complex setting where a repeated measures process and a recurrent events process were correlated, both subject to a terminal event. While these literature mainly focused on model inference, Krol *et al* [36] extended the usage of joint model of longitudinal data, recurrent events, and a terminal event to dynamic prediction area. They derived the estimated probability of having a terminal event in specific time interval, given all previous history and no event occurred before prediction starting time. Regarding prediction of recurrent events, Musoro *et al* [37] extended dynamic prediction by landmarking to recurrent event data. Using landmark method, they handled the longitudinal data as time-fixed covariate at different landmark time point. However, to our best knowledge, no research has been done in providing personalized dynamic prediction in the recurrent events using joint model approach. In this article, we propose a joint model of longitudinal data and recurrent event data and extend to dynamic prediction of next occurrence of recurrent event within given time interval. Instead of using landmark method, the proposed joint model approach simultaneously models time updated longitudinal biomarker and recurrent event process.

2.3 Methods

2.3.1 Joint model specification

Some clinical outcomes may occur in a recurrent fashion, making it important to investigate recurrent event data. In many studies, clinical bio-markers of targeted patients are often collected longitudinally, utilizing this information helps us to identify risk factors of recurrent events and predict the risk of future occurrence. There are different ways to incorporate longitudinal information in recurrent event model building, to model the longitudinal process and recurrent event process at same time, we extend the typical joint model method to handle recurrent event data.

Let $\boldsymbol{y}_i(t) = \{y_i(t_{ij})\}$ be the vector of longitudinal observation for subject i at time t_{ij}, where $i = 1, \ldots, N$, and $j = 1, \ldots, m_i$. Let T_{ik} be the recurrent event times from study onset for subject i, $k = 0, \ldots, n_i$, where n_i denotes the number of recurrent cardiovascular events (including CHD, stroke, and HF). We use a linear mixed effects sub-model to model the longitudinal health outcome and a poisson intensity recurrent event sub-model for modeling the re-occurrence of CVD events. These two sub-models are correlated by random effects and expected values of longitudinal outcomes.

The joint model for recurrent event and longitudinal processes is:

$$\begin{gathered} y_i(t) = \boldsymbol{X}_i^Y(t)\boldsymbol{\alpha} + \boldsymbol{Z}(t)_i\boldsymbol{u}_i + \boldsymbol{V}_R(t)\boldsymbol{\zeta} + e_i(t) = f_i(t) + e_i(t) \\ r_i(t) = r_0(t)\exp\{\boldsymbol{Z}_i^R\boldsymbol{\beta} + \nu f_i(t) + v_i\}, \end{gathered} \tag{2.1}$$

To allow flexibility and variation of baseline risk intensity, random effect v_i is added in the recurrent event submodel to explain the variation between subjects in baseline hazard. We assume independence between $\boldsymbol{u}_i$ and v_i. The two sub-models are linked via an association parameter ν, which quantifies the strength of correlation between the expected longitudinal outcome and the hazard of recurrent CVD events. Parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ denote the estimated coefficients of longitudinal process related covariates and recurrent event process related covariates, respectively. One key assumption in the linear mixed effects model is that all measurements from each patients are independent conditioning on the random effects vector $u_{i,1}$ and $u_{i,2}$. Thus, the likelihood for the longitudinal process of subject i is:

$$l_i^Y = \prod_{j=1}^{m_i} \frac{1}{\sqrt{2\pi}\sigma_e} \exp\Bigg[- \frac{(y_i(t_{ij}) - \boldsymbol{X}_i^Y(t_{ij})\boldsymbol{\alpha} - \boldsymbol{Z}_i(t_{ij})\boldsymbol{u}_i - \boldsymbol{V}_R(t)\boldsymbol{\zeta})^2}{2\sigma_e^2}\Bigg]. \tag{2.2}$$

Under the assumption that the number of recurrent events in non-overlapping time intervals is a poisson process, we are able to model the event process via intensity function. As in Eq (2.1), the covariate vector $\boldsymbol{Z}_i^R$ can be identical or different from covariate vector $\boldsymbol{X}_i^Y$ in the longitudinal sub-model. To increase the robustness of our model fitting, we utilize piece-wise constant baseline hazard model to obtain estimators for both fixed effects and random effects. We divide the total follow-up time interval $[0, \tau]$ by using time knots $\tau_t = (0, \tau_1, ..., \tau_R)$ by quantile of event times, and denote the baseline hazard vector as $g = (g_0, g_1, ..., g_{R-1})$. Then we can define the piecewise constant hazard function as $h_0(t) = \sum_{r=0}^{R-1} g_r I_r(t)$, where indicator function $I_r(t) = 1$, if $\tau_r \leq t < \tau_{r+1}$ and 0 if otherwise. Therefore, the likelihood of the recurrent events process for subject i is:

$$\begin{aligned} l_i^R &= \prod_{k=0}^{n_i} h_i(t_{ik})^{\sigma_{ik}} S_i(x_i) \\ &= \prod_{k=0}^{n_i} \Big[r_0(t_{ik}) \exp\big\{ v_i + \boldsymbol{Z}_i^R\boldsymbol{\beta} + \nu f_i(t_{ik}) \big\} \Big]^{\delta_{ik}} \cdot \exp\Big[- \int_0^{x_i} r_0(t) \exp\big\{ v_i + \boldsymbol{Z}_i^R\boldsymbol{\beta} + \nu f_i(t) \big\} dt \Big], \end{aligned}$$

where δ_{ik} is the indicator of a recurrent event at time t_{ik} and x_i is the observed follow-up time. Here, the longitudinal outcome $\boldsymbol{y}_i$ is assumed to be independent of time $\boldsymbol{t}$, conditioning on the random effects $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$. Thus, the full likelihood for subject i is

$$l_i = l_i^Y \cdot l_i^R \cdot f(\boldsymbol{u}_i) \cdot f(\boldsymbol{v}_i), \tag{2.3}$$

where $f(\boldsymbol{u}_i)$ and $f(\boldsymbol{v}_i)$ is the density function of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ respectively. Our unknown parameter vector is $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \nu, \sigma_1, \sigma_2, \rho, \sigma_v, \sigma_\zeta, \sigma_e\}$.

2.3.2 Dynamic prediction

We randomly partition the dataset into two part, one is the training datasest which is used to build the model, another is the validation dataset to access the prediction performance of proposed model. After obtaining the posterior samples of parameter vectors from inference of training dataset, we illustrate the derivation and procedure to conduct dynamic prediction of each subject in validation dataset. Suppose a new subject i had n_i (e.g., $n_i = 0, 1, 2, \cdots$) recurrent events up to time t, with longitudinal profile $\mathcal{y}_i(t) = \{y_i(t_{ij}); 0 \leq t_{ij} \leq t\}$, we would like to predict his/her probability of having the $n_i + 1$ recurrent event before time $t' = t + \Delta t$ (e.g., 1 year), denoted by $\pi_i(t'|t) = P(T_{i,n_i+1} \leq t' | T_{i,n_i+1} > t, \mathcal{y}_i(t), \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector in model (2.1).

We can derive $\pi_i(t'|t)$ as follows:

$$\begin{aligned}
&\pi_i(t'|t) \\
&= \int\int P(T_{i,n_i+1} \leq t'|T_{i,n_i+1} > t, \boldsymbol{u}_i, v_i, \boldsymbol{y}_i(t), \boldsymbol{\theta}) P(\boldsymbol{u}_i|T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i|T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&\approx \frac{1}{M}\sum_{m=1}^{M} 1 - \exp\left[- \int_t^{t'} r_i^{(m)}(s|\boldsymbol{u}_i^{(m)}, v_i^{(m)}, \boldsymbol{\theta}^{(m)}) ds \right].
\end{aligned} \tag{2.4}$$

Here $\boldsymbol{\theta}^{(m)}$ is the mth sample ($m = 1, \ldots, M$, where M is the number of post burn-in samples) of parameter vector $\boldsymbol{\theta}$. For random effects, $\boldsymbol{u}_i^{(m)}$ and $\boldsymbol{v}_i^{(m)}$ is the mth sample of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ respectively. The term $r_i^{(m)}$ denotes the intensity function from poission process conditioning on mth copy of $\boldsymbol{\theta}$ and corresponding random effects $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$. Detailed derivation can be found in Appendix.

The key steps to approximate the event probability $\pi_i(t^{'}|t)$ are obtaining samples for random effect $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$. The posterior samples of $\boldsymbol{u}_i$ come from the posterior distribution $P(\boldsymbol{u}_i|T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i)$. Specifically, conditional on the mth posterior sample $\boldsymbol{\theta}^{(m)}$, we draw the mth sample of $\boldsymbol{u}_i$ from its posterior distribution

$$\begin{aligned}
P(\boldsymbol{u}_i|T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}^{(m)}, v_i) &= \frac{P(\boldsymbol{y}_i(t), T_{i,n_i+1} > t, \boldsymbol{u}_i|\boldsymbol{\theta}^{(m)}, v_i)}{P(\boldsymbol{y}_i(t), T_{i,n_i+1} > t|\boldsymbol{\theta}^{(m)}, v_i)} \propto P(\boldsymbol{y}_i(t), T_{i,n_i+1} > t, \boldsymbol{u}_i|\boldsymbol{\theta}^{(m)}, v_i) \\
&= P(\boldsymbol{y}_i(t)|\boldsymbol{u}_i, \boldsymbol{\theta}^{(m)}) P(T_{i,n_i+1} > t|\boldsymbol{u}_i, \boldsymbol{\theta}^{(m)}, v_i) P(\boldsymbol{u}_i|\boldsymbol{\theta}^{(m)}),
\end{aligned}$$

where $P(T_{i,n_i+1} > t|\boldsymbol{u}_i, \boldsymbol{\theta}^{(m)}, v_i) = \prod_{k=0}^{n_i} \left[r_0(t_{ik}) \exp\{\boldsymbol{Z}_i^R \boldsymbol{\beta} + \nu f_i(t_{ik}) + v_i\} \right]^{\sigma_{ik}} \cdot \exp\Bigg[-$

$\int_0^t r_0(s) \exp\left\{\boldsymbol{Z}_i^R \boldsymbol{\beta} + \nu f_i(s) + v_i\right\} ds\Bigg]$.

We then draw the posterior samples of v_i from distribution $P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta})$ as follows:

$$\begin{aligned} P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}^{(m)}, \boldsymbol{u}_i^{(m)}) &= \frac{P(T_{i,n_i+1} > t, v_i | \boldsymbol{\theta}^{(m)}, \boldsymbol{u}_i^{(m)})}{P(T_{i,n_i+1} > t | \boldsymbol{\theta}^{(m)}, \boldsymbol{u}_i^{(m)})} \propto P(T_{i,n_i+1} > t, v_i | \boldsymbol{\theta}^{(m)}, \boldsymbol{u}_i^{(m)}) \\ &= P(T_{i,n_i+1} > t | v_i, \boldsymbol{\theta}^{(m)}, \boldsymbol{u}_i^{(m)}) P(v_i | \boldsymbol{\theta}^{(m)}), \end{aligned}$$

For each of $\boldsymbol{\theta}^{(m)}$, $m = 1, \ldots, M$, we use adaptive rejection metropolis sampling (ARMS) [50] in `R HI` package to draw 50 samples of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ and retain the final sample. This process is repeated for the M saved values of $\boldsymbol{\theta}$. Once the posterior distributions of random effects are simulated, all calculations become straightforward and produce the entire distribution of the future trajectory (health outcomes and risk) of a new subject. For example, the outcome trajectory at time t' is $y_i(t') | \boldsymbol{u}_i \sim N(\boldsymbol{Z}_i(t')^Y \boldsymbol{\alpha}^{(m)} + \boldsymbol{u}_i^{(m)}, \sigma_e^{2(m)})$. Suppose that patient i does not have a recurrent event by time t', then the outcome history is updated to $\boldsymbol{y}_i(t')$. We can dynamically update the posterior distribution to $p(\boldsymbol{u}_i | T_i > t', \boldsymbol{y}_i(t'), \boldsymbol{\theta}^{(m)})$ and $p(v_i | T_i > t', \boldsymbol{\theta}^{(m)})$, draw new samples of $\boldsymbol{u}_i$ and v_i, and obtain the updated predictions.

2.3.3 Bayesian inference

We use Bayesian inference based on Markov chain Monte Carlo (MCMC) posterior simulations to infer unknown parameters. We use non-informative priors on all parameters in vector $\boldsymbol{\theta}$. To be more specific, the prior distribution of all elements in the coefficient vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and the the association parameter ν are normal distribution with mean equal to 0 and standard deviation being 10. To ensure the positivity of variance parameter, we use inverse gamma distribution with $\alpha = 0.01, \beta = 0.01$ as the prior distribution of $\sigma_1, \sigma_2, \sigma_v, \sigma_\zeta, \sigma_e$, so the corresponding variance is 100. For the correlation parameter ρ between random intercept and random slope, we use the uniform prior distribution $\rho \sim U(-1, 1)$.

There are variety of MCMC sampler that can be used to conduct Bayesian inference, including `WinBUGS`, `OpenBUGS`, `JAGS`, `Stan` and etc. We employ `Stan` to fit our proposed model, because unlike `WinBUGS`, which uses Gibbs sampler, `Stan` uses Hamiltonian Monte Carlo (HMC) sampler. Relying on deterministic mechanism inspired by Hamiltonian dynamics, HMC method generates coherent exploration of target distribution. When compared to standard Gibbs method, such approach not only improves posterior samples convergence speed, but also gives the resulting estimator with stronger validity [51]. Our proposed model is fitted in `RStan` while specifying aforementioned full likelihood function and prior distributions of unknown parameters. To diagnose the convergence of posterior samples, we use trace plot to decide the samples are converged if no apparent trends can be viewed from the plot. Another criteria to ensure convergence is Gelman and Rubin [52] potential scale reduction statistic. we ensure the scale reduction statistics $\hat{R}$ are smaller than 1.1 for all parameters as an additional way to ensure convergence. In simulation, 2,000 after burn-in samples are draw for each parameter in $\boldsymbol{\theta}$, the convergence of these posterior samples are ensured by meeting aforementioned criteria.

2.3.4 Parallelizing MCMC with Random Partition Trees

As large clinical studies being more and more prevalent, the scale of clinical datasets has brought challenges to conventional MCMC sampling techniques. For these large datasets, conducting Bayesian inference can be extremely time consuming. Moreover, processing such large datasets in a single computer likely encounters the problem of non-sufficient memory. Although there are various attempts to address these issues [53, 54], they either rely on asymptotic normality of posterior distributions, or have the drawback of insufficient of resampling. To the end, Wang *et al.* [55] proposed a embarrassingly parallel MCMC (EP-MCMC) approach to address the large sample size problem without assuming any distribution of posterior samples. The theoretical property of this EP-MCMC approach grants its good performance when being applied to different models. The algorithm consists of two step, first it partitions the data into multiple subsets and independently run MCMC sampler on each of them, then it applies random partition trees to combine the posterior draws from subset.

Simulation is conducted to ensure EP-MCMC method is valid under our proposed model setting. The simulation setting is illustrated in Section 2.4. To conduct EP-MCMC, we randomly divide 2000 samples into 10 subsets, each subset contains 200 samples and sufficient recurrent events. We then conduct Bayesian data analysis employing our proposed model in each subset, which results in 2,000 after burn-in MCMC posterior samples for every parameter of interest. For example, for the 12 parameters, we will have 10 matrices of size $2,000$ by 12, each coming from one subset. After obtaining the posterior samples of parameters in multiple subsets, we implement `PART` algorithm [55] in `Matlab` to combine the posterior samples from 10 subsets. This algorithm aggregates sub-chain posterior MCMC samples and draws a certain number of (e.g. 10,000) samples for each parameter from the combined posterior with k-dimensional tree (k-d tree) partition rules. Following aforementioned example, after we conduct the aggregation step, a matrix of size $10,000 \times 12$ is obtained as the posterior samples for the whole dataset. The posterior mean of the samples suggest that the parameters estimated by this EP-MCMC approach have minor bias from the true values. Also, the coverage probability does not show evidence of biased estimation or overly conservative standard error estimates, which ensures the validity of this parallel MCMC approach under our proposed model.

Besides the ability of dealing with large sample problem in application, EP-MCMC method has additional benefit in regards of efficiency in estimating survival probabilities of all subjects. To avoid over-fitting, we separate the whole data into 10 subsets and estimate parameters in training set (consists of 9 of the subsets), dynamic prediction is then conducted to estimate event-free probability in validation set (the unused 1 subset). Instead of having to repeat the inference MCMC sampling process 10 times to get the estimated survival probabilities for all subjects (each time, survival probabilities of subjects from 1 validation set are estimated), EP-MCMC method enables us to aggregate posterior samples from selected 9 subsets to obtain the estimation of parameters in the training set. In this way, we no longer need to run MCMC sampler for 10 times in each pooled training set, but just use the posterior samples of different subsets to obtain inference results. As the Bayesian MCMC sampling technique is usually time-consuming, this approach save a lot of computation time.

2.3.5 Assessing predictive performance

Accurate identification of patients with higher risk of having a new future event is one of the most interested features in statistical models. It is important to access how well our proposed risk prediction model performs and how accurately it predicts a future event within specific given time period. Here, we access the prediction performance in three aspect, global discrimination ability (the ability that the model correctly classifies event and non-event.), validation performance (how well the model predicts the data), comparison with other model. To be more specific, we employ receiver operating characteristic (ROC) curve and the area under the ROC curves (AUC) to assess the discrimination ability of the proposed model, the validation performance is assessed using the expected Brier score (BS). Moreover, we adopt the methodology approach proposed by Blanche *et al.* [34] to quantify and compare the predictive accuracy of different models.

Area under the ROC curves

Following the definition in previous section, for a given cut value $c \in [0,1]$, the time dependent sensitivity and specificity are $P\{\pi_i(t'|t) > c | D(t,t') = 1, T_i^* > t\}$ and $P\{\pi_i(t'|t) \leq c | D(t,t') = 0, T_i^* > t\}$ respectively, where $D(t,t')$ is an indicator function equals to 1 when a new event happen during time interval (t,t'] and equals to 0 otherwise. Therefore, for probability $p \in [0,1]$, the ROC curves will be $ROC_t^{t'}(p) = TP_t^{t'}[FP_t^{t'}]^{-1}(p)$, where $TP_t^{t'}$ denotes the true positive rate, $FP_t^{t'}$ denotes the false positive rate [56]. With the defined time-dependent sensitivity and specificity, we calculate a standard "concordance" summary: the time-dependent Area Under Curve (AUC), the formula is as following [57]:

$$AUC(t,t') = \int_0^1 ROC_t^{t'}(p)dp.$$

With standard numerical integration methods, we can estimate the time-dependent AUC straightforwardly, and it serves as a criteria to assess the global discrimination ability of the proposed model.

Dynamic Brier score

By extending the Brier score (BS) defined in survival models to joint model framework [58], we are able to define the expected BS for dynamic prediction as $BS(t,t') = E[(D(t'|t) - \pi(t'|t))^2]$, here $D(t'|t)$ denotes the observed event status, which equals to 1 if subjects experience a new event during $(t, t]$ and equals to 0 otherwise. As the dynamic expected Brier score is a mean squared error, we can express it as

$$BS(t, t') = E[(E[D(t'|t)] - \pi(t'|t))^2] + E[(D(t'|t) - E[D(t'|t)])^2].$$

The first term in aforementioned equation measures how close the predictions are to expected event status $E[D(t'|t)]$, in other word, evaluates how well the models predict the observed data. The second term is an aggregation of resolution and uncertainty and does not depend on the distribution of the predictions. Both AUC and BS are interesting statistics when assessing the predictive accuracy, and they complement each other when quantifying the overall performance of the dynamic prediction probability $\pi(t'|t)$.

Comparing Dynamic Predictive Accuracy

While AUC and BS are well studied in quantifying the overall prediction performance of a specific model, comparing these two indexes between different models is getting more interest. By comparing the AUC and BS, it enables researchers to select appropriate model which is useful enough regarding prediction purposes. Blanche *et al.* [34] proposed a methodology approach to compute confidence regions of AUC and BS, also tests for the difference of them in different models. The detailed derivations can be found in the publication. As a consequence of the derivations, denoting either AUC(t, t') or BS(t, t') as $\theta(t, t')$, let $\hat{\theta}(t, t')$ denote the corresponding estimator and $\Delta\hat{\theta}(t, t')$ denote the estimator of the difference of AUC or BS between two prognostic models at given landmark time t. We are able to obtain the pointwise confidence interval

$$\Delta\hat{\theta}(t,t') \pm z_{1-\alpha/2}\frac{\hat{\sigma}_{\Delta,t,t'}}{\sqrt{n}},$$

here $z_{1-\alpha/2}$ is the critical value of standard normal distribution and $\hat{\sigma}^2_{\Delta,t,t'}$ is the empirical estimator of variance, which can be consistently estimated by the influence function following formulas in Blanche *et al.* [34]. Therefore, testing for comparison of two prediction accuracy measurements can also be derived accordingly.

2.4 Simulation Setting

In this section, we examine the performance of parameter inference and prediction of the proposed model via simulation. Consider binary covariates for x_{i1}, x_{i2} which equals to 0 or 1 with probability 0.5. We generate the longitudinal measure $Y_i(t_{ij})$ at time t_{ij} of subject i as $y_i(t_{ij}) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 t_{ij} + u_{i1} + u_{i2} t_{ij} + e_i(t_{ij})$, where the error term ϵ_i follows normal distribution, i.e. $e_i(t_{ij}) \sim N(0, \sigma_e^2)$, $i = 1, \ldots, I$, and $j = 1, \ldots, m_i$. For recurrent event process, we assume non-informative censoring, the censoring time C_i is sampled from uniform distribution $Uniform(9, 10)$ with 50% censoring rate. Let T_{ik} be the kth recurrent event times from study onset (time 0) for subject i, $k = 0, \ldots, n_i$, where n_i denotes the number of recurrent events. Let $r_i(t)$ denote the intensity of the recurrent process. Random intercept u_{i1} and random slope u_{i2} follows multivariate normal distribution with mean 0. And the covariance matrix is denoted by Σ_u, with σ_1^2, σ_2^2 denote the variance for random effects u_{i1}, u_{i2} respectively while ρ represents the correlation between two random variables. On the other hand, v_i is the random variable only associated with recurrent process and independent of u_{i1} and u_{i2}, we assume that it follows a normal distribution with variance denoted by σ_v^2. Our simulation setting is as follows:

$$y_i(t_{ij}) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 t_{ij} + u_{i1} + u_{i2} t_{ij} + e_i(t_{ij}) = f_i(t_{ij}) + e_i(t_{ij})$$

$$r_i(t_{ik}) = r_0 \exp\{\beta_1 z_i + \nu f_i(t_{ij}) + v_i\}$$

The inference results is based on 200 simulations with training sample size equal to 600, on average, there are 2 recurrent events per subject. Table 2.1 displays the bias of estimated parameter with true value, standard deviation of posterior mean (SD), standard error (SE), coverage probability (CP), and root mean squared error (RMSE) of inference results. The simulation results without using Parallel EP-MCMC method is summarized in left panel of Table 2.1. To evaluate the performance of Parallel EP-MCMC method under our proposed model setting, we randomly divide the training dataset into 3 subsets and apply the EP-MCMC method following aforementioned procedures in Section 2.3.4. We present the results in Table 2.1 right panel. The results suggest that under simulation settings, parameters estimated by our proposed model have minor bias with true value, small RMSE, and the coverage probability (close to 0.95) does not show evidence of biased estimation or overly conservative standard error estimates. After the inference is done, we apply our proposed dynamic prediction method in validation

	Results without Parallel MCMC					Results with Parallel MCMC				
Parameter	BIAS	SD	SE	CP	RMSE	BIAS	SD	SE	CP	RMSE
α_0=-1.000	0.001	0.169	0.184	0.960	0.169	0.010	0.175	0.184	0.955	0.175
α_1=-0.200	-0.007	0.235	0.244	0.955	0.235	-0.011	0.241	0.244	0.960	0.240
α_2=0.800	0.000	0.037	0.038	0.955	0.037	0.000	0.037	0.038	0.955	0.037
σ_1=1.500	-0.005	0.123	0.138	0.970	0.123	-0.003	0.123	0.138	0.970	0.123
σ_2=0.150	0.008	0.025	0.030	0.980	0.026	0.007	0.025	0.030	0.985	0.026
σ_v=0.200	0.001	0.039	0.041	0.960	0.039	0.001	0.040	0.041	0.955	0.040
ρ=0.400	0.006	0.183	0.244	0.995	0.183	0.008	0.184	0.245	0.995	0.183
σ_e=5.000	-0.001	0.068	0.065	0.940	0.067	0.000	0.065	0.065	0.950	0.065
β_1=-0.12	0.003	0.025	0.025	0.955	0.025	0.002	0.025	0.025	0.960	0.025
ν=0.750	-0.001	0.039	0.042	0.980	0.039	-0.001	0.038	0.042	0.980	0.038

Table 2.1: Parameter estimates in simulation

dataset, which contains 200 new subjects who has never been used in training. Results for dynamic prediction in validation dataset are presented in Table 2.2. Here AUC is the average of all AUC calculated within given time interval for 200 simulation times. The last column represents the average bias of the estimated event-free probability and the true probability that calculated using true value of parameters. The estimated AUC and BS are acceptable rate, and bias are all under 0.005. In conclusion, the simulation results suggest that our proposed model gives close estimation of parameter vector, also has a good prediction performance.

t	t'	AUC	BS	Bias
5	6	0.746	0.014	0.002
	7	0.756	0.016	0.003
	8	0.759	0.017	0.004
	9	0.761	0.020	0.004
6	7	0.788	0.016	0.003
	8	0.788	0.017	0.004
	9	0.789	0.019	0.005
7	8	0.814	0.018	0.003
	9	0.812	0.021	0.004

Table 2.2: Time-dependent AUC in simulation

2.5 ALLHAT Study Application Results

In this section, we apply the proposed joint model to the motivating ALLHAT study. Considering the recurrent non-fatal cardiovascular disease events (composite CHD, stroke, and heart failure) as our primary survival outcome, we select SBP as our longitudinal biomarker which showed significant association with time to CVD events in previous literatures. The longitudinal process and recurrent event process are related by sharing an underlying latent function, an association parameter serves as an measurement of the relationship between longitudinal biomarker and recurrent CVD events.

A number of potential risk factors of cardiovascular disease are assessed and evaluated in previous related studies, including age, gender, diabetes, hypertension, history of cardiovascular disease, alcohol consumption, tobacco use, family history of cardiovascular disease, environmental factors, etc. [44, 59]. We pre-select the following risk factors to be included in our analysis upon availability of the ALLHAT data: trial randomization group (0 if chlorthalidone, 1 if doxazosin), age (in year), gender (0 if female, 1 if male), race (0 if white and others, 1 if black), diabetes (0 if no, 1 if yes), history of MI or Stroke at baseline (0 if no, 1 if yes), history of coronary revascularization (0 if no, 1 if yes), antihypertensive treatment before trial (0 if no, 1 if yes). These covariates are separately tested to be significant either the linear mixed effect model or the recurrent event model. After excluding subjects with pending death confirmation pending and those with missing data in aforementioned covariates, we obtain the analytical dataset with total 19,804 subjects. Among these subjects, 16,811 subjects are cardiovascular disease events free until the end of follow-up period, 2,179 subjects experienced 1 event, and 814 subjects experienced more than 1 events during the follow-up of the study.

Due to the fact that ALLHAT study is one of the largest cardiovascular disease studies in U.S. and our analysis include around 20,000 subjects, running MCMC sampler to get posterior samples of model parameters in the whole dataset can be very time consuming. In order to speed up the computation, we adopt the parallel MCMC approach in Section 2.3.4. To employ EP-MCMC method, we divide the analytical dataset into 10 partitions, each partition includes around 2,000 randomly sampled subjects. In each of the 10 subsets, we run two parallel MCMC chains with overdispersed initial values and each chain is ran for 5,000 iterations. The first 3,000 iterations are discarded as burn-in and the inference of all parameters are based on remaining 2,000 iterations from each chain. We employ piece-wise constant baseline hazard function in our analysis. To ensure we have enough observations in each piece-wise interval, we construct intervals by every 1/4th quantile of the CVD events time in ALLHAT dataset. Good mixing and convergence properties of the MCMC chains are ensured in trace plots and $\hat{R}$ for each parameter is below 1.01.

Parameters estimation using aforementioned method are presented in Table 2.3. For longitudinal sub-model, the results suggests that there is significant different in randomization drug group, on average, the systolic blood pressure for subjects in doxazosin group is 2.796 units (95% CI [2.428, 3.143]) higher than chlorthalidone group. Other risk factors including age, gender, race, and baseline anti-hypertensive drug usage are found to have significant effects on the systolic blood pressure. Specifically, 1 year increase in age increases the systolic blood pressure by 0.136 units (95% CI [0.111, 0.162]) on average. When compared to female subjects, male tend to have 1.2 units (95% CI [0.808, 1.584]) lower systolic blood pressure. On average, black subjects have 2.388 units (95% CI [2.008, 2.773]) higher systolic blood pressure when compared to others. Subjects who took anti-hypertensive drugs before baseline have 1.119 units (95% CI [0.469, 1.760]) higher systolic blood pressure than others. The results are consistent with the primary ALLHAT publications [60]. For recurrent event sub-model, we find that subjects with diabetes history have significantly higher risk to develop new cardiovascular disease events (RR= 1.578; 95% CI [1.418, 1.742]). Similar effects can be found in subjects with MI or stroke (RR= 1.941; 95% CI [1.740, 2.160]), and subjects with history of coronary revascularization (RR= 2.175; 95% CI [1.923, 2.479]). Moreover, the association parameter ν is statistically significant, indicating the longitudinal systolic blood pressure measurements

	Mean	2.5%	97.5%	SD
longitudinal submodel				
Intercept	2.688	1.962	3.412	0.393
Trt(doxazosin)	2.796	2.428	3.143	0.186
Age(years)	0.136	0.111	0.162	0.013
Male	1.200	0.808	1.584	0.209
Black	2.388	2.008	2.773	0.196
BL Antihypertensive drug (Yes)	1.119	0.469	1.760	0.339
σ_1	12.159	11.947	12.364	0.107
σ_2	3.333	3.222	3.441	0.056
ρ	-0.563	-0.584	-0.541	0.011
σ_e	12.416	12.363	12.470	0.028
recurrent event submodel				
Diabetes (Yes)	0.456	0.349	0.555	0.054
History of MI or Stroke	0.663	0.554	0.770	0.054
History of coronary revascularization	0.777	0.654	0.908	0.065
Association	0.016	0.011	0.020	0.002

Table 2.3: Parameters estimation results

are positively associated with the risk of recurrent CVD event with a rate at 0.006 (95% CI [0.001, 0.010]).

To assess and compare the prediction performance, we calculate time-dependent AUC and BS of our proposed model (Model 1) and a simple recurrent event model (Model 2). We present the estimated AUC and BS of different prediction time windows in Table 2.4. The result suggests that our proposed model have higher AUC and lower BS than simple recurrent event model in every prediction interval, suggesting that using longitudinal biomarker improves the prediction of patient's risk of developing a new recurrent event. The results also indicate that using more longitudinal and disease information tend to improves the model prediction performance. For later prediction starting time, updated SBP measurements are used in predicting risk of having a recurrent event, therefore the prediction performance is improved.

To compare the dynamic prediction accuracy curves of these two prognostic models, we employ the testing method proposed by Blanche *et al.* [34]. To illustrate the comparison results, we plot the estimated AUC of Models 1 and 2 (upper panels) as well as the difference of mean AUC from these two prognostic models (lower panels) in Figure 2.2. The first graph of upper and lower panels displays the estimated AUC and difference of AUC respectively for prediction landmark time year 1 with four prediction time windows considered (0.25 year, 0.5 year, 0.75 year, and 1 year). Other graphs are constructed in a similar fashion, but with different prediction landmark time (year 2, year 3, and year 4). We find that for all landmark times, AUC from our proposed model (Model 1; solid black line) is higher than it from simple recurrent event model (Model 2; solid gray line), and the corresponding difference of AUC is above 0 for every prediction time window. For prediction after year 1 visit, the estimated difference of AUC from Model 1 and 2 for 0.5, 0.75, and 1 year prediction window are significant at 0.05 level. Similarly, for prediction made after year 2 visit, there is a significant difference of estimated AUC between Model 1 and 2 for 0.75 and 1 year prediction window at 0.05 significance level. On the other hand, the differences of AUC estimated from predictions made after year 3 and year 4

		Model 1		Model 2	
t(year)	Δt(month)	AUC	BS	AUC	BS
1	3	0.665	0.015	0.556	0.015
	6	0.676	0.027	0.556	0.028
	9	0.675	0.038	0.557	0.040
	12	0.668	0.049	0.558	0.051
2	3	0.660	0.015	0.572	0.015
	6	0.653	0.028	0.566	0.029
	9	0.658	0.041	0.559	0.052
	12	0.651	0.041	0.553	0.052
3	3	0.706	0.017	0.652	0.017
	6	0.695	0.034	0.603	0.035
	9	0.676	0.049	0.604	0.052
	12	0.658	0.065	0.598	0.070
4	3	0.744	0.020	0.639	0.021
	6	0.696	0.067	0.616	0.069
	9	0.685	0.059	0.576	0.063
	12	0.643	0.077	0.576	0.080

Table 2.4: Time-dependent AUC and BS for proposed joint model (Model 1) and reference recurrent event model (Model 2)

from these two models are not significant at 0.05 level.

We select two subjects who experienced different occurrences of cardiovascular disease events to illustrate our subject-level prediction results. Conditioning on their available measurements, we predict their future longitudinal trajectory as well as the probability of developing a new event in given time interval. With 11 recorded longitudinal SBP measurements, subject 127 (upper panels) experienced total 5 recurrent CVD events before year 4. Subject 15 (lower panels) had 13 recorded SBP measurements and only had one CVD event during year 2 and year 3. Figure 2.3 presents the estimated SBP from year 1 to year 5 for both subjects. When only 1 year information is used in prediction,

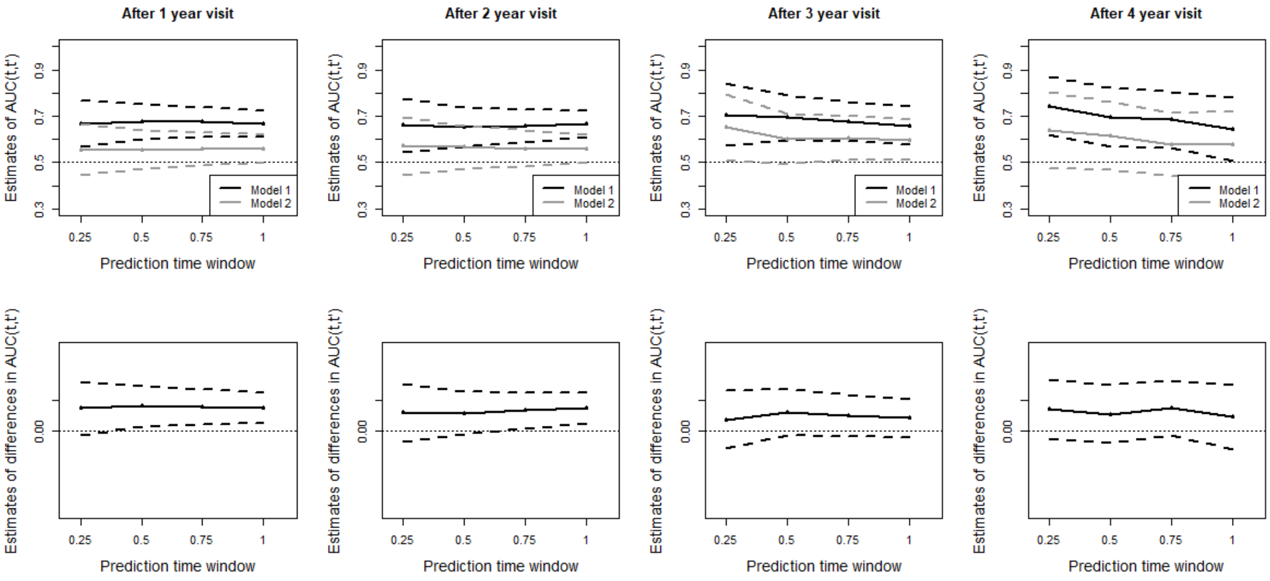


Figure 2.2: Compare AUC for proposed joint model (Model 1) and reference recurrent event model (Model 2)

the 95% confidence bands of both subjects are relatively wide, by using more follow-up data, the 95% confidence band is narrower. Most of the actual SBP measurements fall into the prediction confidence interval.

Besides the prediction of future trajectory of SBP, predicting the probability of having a new CVD event in future is more of clinical interest. We illustrate the prediction patterns of aforementioned two subjects in Figure 2.4. The upper penal of Figure 2.4 presents the event-free probability from different landmark time to year 5 for subject 127, while the lower penal presents the probability for subject 15. Based on 1 year information, subject 15 has a lower event risk between year 2 and year 5 when compared to subject 127 (the first plot of upper and lower panels). When year 2 data become available, the prediction dynamically updates and the risk of having new CVD event decreases as

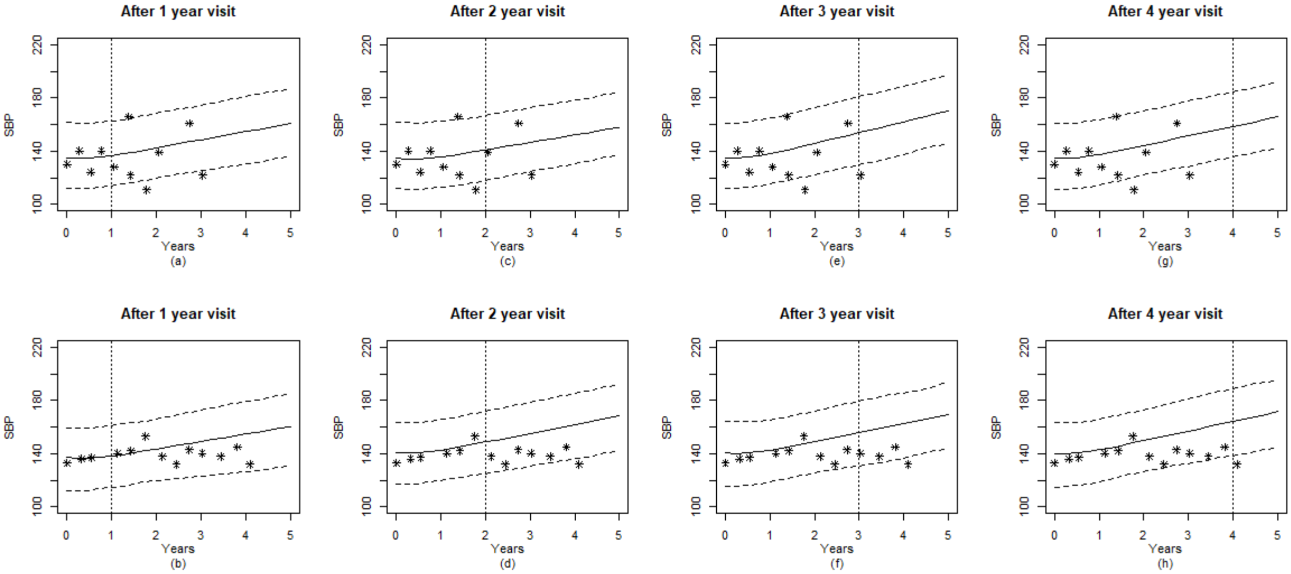


Figure 2.3: Predicted SBP measurements for Subject 127 (upper panels) and Subject 15 (lower panels) in ALLHAT study. Dashed lines are the 2.5% and 97.5% percentiles. The dotted vertical line represents the time of prediction t.

no event occurred for both of the subjects (the second plot of upper and lower panels). However, for subject 127, predicted event risk after year 3 and year 4 increases sharply as multiple CVD events occurred during this time interval (the third and last plots of upper panels). Clinicians may closely monitor the subject and take personalized treatment based on the prediction. A pattern which is similar to longitudinal trajectory prediction is that when more data come into available, the prediction becomes more precise as the confidence band narrows down.

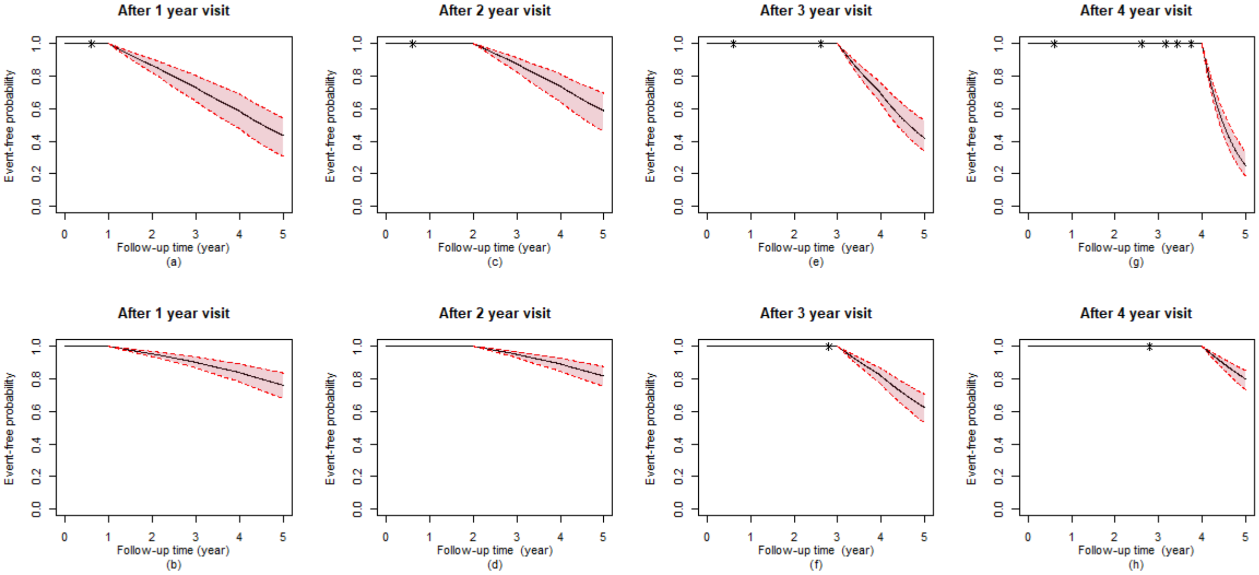


Figure 2.4: Predicted CVD event free probability for Subject 127 (upper panels) and Subject 15 (lower panels) in ALLHAT study. Dashed lines are the 2.5% and 97.5% percentiles.

2.6 Discussion

In this article, we propose a joint model of a longitudinal outcome and a recurrent event process and apply our proposed model in ALLHAT dataset. The longitudinal process and recurrent event process are correlated via shared latent function, an association parameter is set to model the relationship between these two process. The simulation study indicates that the coefficients of covariates in both submodel and the association parameter estimated using our proposed model have small bias to true value and appropriate coverage probability. Dividing the simulated dataset to training set and test set, we conduct personalized prediction for each simulated sample in test set, and compute their probability of having an new event in given time interval. The prediction performance is assessed by time-dependent Area Under Curve and dynamic Brier score. The Area Under Curve and Brier score estimated using our proposed method are at acceptable rate and mean survival probability bias are all below 0.005, indicating our proposed prediction approach has good performance. We then apply our proposed model in ALLHAT study. To address the computation issue caused by the large sample size of ALLHAT study, we employ parallel MCMC method to do inference of parameters parallelly, which significantly reduce our computing time. In every given prediction time interval, our proposed model outperforms the simple recurrent event model in regards to AUC, which indicates that utilizing subject-specific longitudinal information helps improve prediction accuracy of having a new event in near future.

There are some possible future directions that we would like to pursue. One limitation of our proposed model is that we treat death as non-informative censoring. This assumption might not be true as death process can be related to longitudinal biomarker or recurrent events. Joint model for the three correlated outcomes was proposed by recent work in related area [32, 36]. Adding another submodel for death process, Liu & Huang [32] proposed a joint random effects model with correlation between longitudinal process, recurrent event process and death process modeled by shared random effects. Moreover, Krol *et al* [36] not only proposed a trivariate joint model for longitudinal data, recurrent events and a terminal event, but also extended the dynamic prediction for death process. However, to our best knowledge, no work has been done regarding dynamic prediction for recurrent events. In the future work, we will extend the proposed model to take death process into account while predicting future occurrence of recurrent events.

Another limitation of our work is that we only consider predicting future CVD events utilizing under joint model framework, while the longitudinal trajectory is modeled via linear mixed-effect submodel. Other methods including Principal Analysis by Conditional Expectation (PACE) can be utilized to model longitudinal data without making linear assumptions. In future studies, we plan to compare the prediction ability of these two methods. As the occurrence of CVD events are usually related with multiple biomarkers (e.g. serum cholesterol level, serum creatinine), including only longitudinal blood pressure measurements in prediction model may results in relatively low AUC in our analysis. Multivariate Functional Principal Component Analysis (MFPCA) method proposed by Happ *et al.* [61] is a promising way to incorporate multiple longitudinal biomarkers in predicting recurrence of CVD events. By leveraging other longitudinal outcomes, we may be able to improve the prediction performance in future studies.

Moreover, numerous GWAS studies have been conducted in CVD area, incorporating genetic information in our prediction framework may improve prediction accuracy. ALLHAT study also collected genetic information of participants during follow-up, how to efficiently utilize these GWAS data to do the dynamic prediction warrants further investigation.

Acknowledgements

Sheng Luo's research was supported in part by the National Institute of Neurological Disorders and Stroke under award number R01NS091307. The authors acknowledge the Texas Advanced Computing Center (TACC) for providing high-performing computing resources.

Appendix

We can derive $\pi_i(t'|t)$ as follows:

$$
\begin{aligned}
&\pi_i(t'|t) \\
&= \int\int P(T_{i,n_i+1} \le t' | T_{i,n_i+1} > t, \boldsymbol{u}_i, v_i, \boldsymbol{y}_i(t), \boldsymbol{\theta}) P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&= \int\int P(T_{i,n_i+1} \le t' | T_{i,n_i+1} > t, \boldsymbol{u}_i, v_i, \boldsymbol{\theta}) P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&= \int\int \frac{P(t < T_{i,n_i+1} \le t' | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})}{P(T_{i,n_i+1} > t | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})} P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&= \int\int \frac{P(T_{i,n_i+1} > t | \boldsymbol{u}_i, v_i, \boldsymbol{\theta}) - P(T_{i,n_i+1} > t' | \boldsymbol{u}_i, \boldsymbol{\theta})}{P(T_{i,n_i+1} > t | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})} \\
&\quad \cdot P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&= 1 - \int\int \frac{P(T_{i,n_i+1} > t' | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})}{P(T_{i,n_i+1} > t | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})} P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&= 1 - \int\int \frac{S(t' | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})}{S(t | \boldsymbol{u}_i, v_i, \boldsymbol{\theta})} P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&= 1 - \int\int \exp\left[- \int_t^{t'} r_i(s | \boldsymbol{u}_i, v_i, \boldsymbol{\theta}) ds \right] P(\boldsymbol{u}_i | T_{i,n_i+1} > t, \boldsymbol{y}_i(t), \boldsymbol{\theta}, v_i) d\boldsymbol{u}_i P(v_i | T_{i,n_i+1} > t, \boldsymbol{\theta}) dv_i \\
&\approx \frac{1}{M} \sum_{m=1}^{M} 1 - \exp\left[- \int_t^{t'} r_i^{(m)}(s | \boldsymbol{u}_i^{(m)}, v_i^{(m)}, \boldsymbol{\theta}^{(m)}) ds \right].
\end{aligned}
$$

Here $\boldsymbol{\theta}^{(m)}$ is the mth sample ($m = 1, \ldots, M$, where M is the number of post burn-in samples) of parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{u}_i^{(m)}$, $\boldsymbol{v}_i^{(m)}$ is the mth sample of random effects $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ respectively.

Chapter 3

Article 2: Predicting time to PD progression using multiple longitudinal outcomes in PPMI study

Predicting time to H&Y stage reaching 3 using multiple longitudinal outcomes in PPMI study

3.1 Introduction

Parkinson's disease (PD) is a progressive, chronic neurodegenerative disease that affects patients' movement. In the absence of a curative treatment, there is an critical need to identify significant and well-validated biomarkers associated with PD progression [10]. Moreover, there is an increasing interest in utilizing these disease related markers to build prognostic model of PD progression. Previous literature studied a broad range of motor and non-motor symptoms which are clinically correlated with evolution of Parkinson's disease [7–9]. However, due to substantial heterogeneity of different symptoms and subjects, it is hard to give accurate prognoses of disease progression. A lot of effort has been made to build prediction models of PD outcomes based on single or multiple related markers.

As a complex progressive disease, a variety of endpoints have been established to evaluate PD severity and work as a criteria to categorize subjects into different disease stages. Among these measures, Hoehn and Yahr (H&Y) scale has became one of the most commonly and widely used measurement to assess overall PD dysfunction stage [13]. Schrag *et al.* [14] examined the responsiveness of different PD progression outcome measures over time and suggested the H&Y scale to be the most responsive measure. Based on motor functions, the H&Y scale has served as a good endpoint of PD progression in many published studies. The time to development of different H&Y stages from symptom onset or enrollment has been widely studied as a way to evaluate disease progression in past literature [15–18]. Although these literature discussed potential prognostic factors in predicting the H&Y stage progression, only baseline variables were included in the survival model and longitudinal information was not taken into account. Relatively few studies considered modelling longitudinal marker in prediction of PD disease outcome. He and Luo [19] proposed a joint model with multilevel item response theory sub-model for the longitudinal data and Cox propotional hazard sub-model to handle time to terminal event. They assessed the effect of tocopherol on time to initiation of symptomatic therapy in early PD patients. Iddi *et al.* [20] applied a latent time joint mixed-effects model to handle longitudinal outcomes and studied the association between different markers and PD diagnostic category. However, to our best knowledge, no previous research has been done regarding prognostication of time to the H&Y stage transition based on multiple longitudinal markers.

Aiming at developing a prognostic model for time to disease progression measured by H&Y stage transition, we incorporate longitudinal information from varies of clinical markers by employing cutting-edge multivariate functional principal analysis (MFPCA) method in this study. We apply the proposed model to data from the Parkinson's Progression Marker Initiative (PPMI) study, which collected a broad range of clinical markers that were frequently measured for untreated PD patients. External validation is conducted in The Longitudinal and Biomarker Study in PD (LABS-PD) to assess the prediction performance of the established prognostic model across different PD studies. A set of prognostic index is estimated to calculate subject-level risk scores and can be updated when new measurements come into available. The proposed approach enables physicians to make clinical decision based on prognosis from enriched information, moreover, identifying PD patients with higher risk of disease progression according to expected prognostic risk score helps clinicians give adaptive treatment to targeted patients.

3.2 Patients

The Parkinson's Progression Marker Initiative is a multicenter study of patients across North America, Europe, Israel, and Australia. Detailed information regarding study design, inclusion and exclusion criteria, and study protocols can be fould at `https://www.ppmi-info.org/`. Aiming at identify one or more markers of progression for Parkinson's disease (PD), the PPMI study recruited 423 newly diagnosed PD cases, who must be untreated for PD at the time of enrollment. Clinical measurements, imaging data and biological samples are collected longitudinally over a period of 6 years. For PD patients, subjects were scheduled to visit sites every 3 months from enrollment during their first year, and follow-up visits were conducted every 6 months. The LABS-PD program, whose main goal is developing biomarkers that measure risk for PD progression, enrolled total 537 PD subjects at baseline [62]. Multiple motor and non-motor markers were evaluated annually during study follow up period. Before initiating the study, all sites of PPMI study and LABS-PD have been approved by the institutional review board, and all study participants from both studies were given a written informed consent for research.

Our interest is to study the prediction ability of multiple longitudinal risk factors for disease progression. Following definition in study conducted by Müller *et al* [15], we define time to PD progression as time from enrollment to patients reach H&Y stage 3, which will be our primary outcome. Among all PD subjects from PPMI study, 3 of them already reached H&Y stage 3 at baseline and has been excluded in following analysis. We select the candidate longitudinal risk factors that were suggested to have association with PD progression in previous literature [63] and are available in PPMI study: Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS), Montreal Cognitive Assessment (MoCA), total Scale for Outcomes in Parkinson's - autonomic questionnaire (SCOPA-AUT), Modified Schwab and England Activities of Daily Living Scale (SEADL), Symbol Digit Modalities Test (SDM), Geriatric Depression Scale (GDS), Letter Number Sequencing (LNS), Semantic verbal fluency and Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease (QUIP). Cerebrospinal fluid (CSF) biomarkers including total tau (t-tau), phosphorylated tau (p-tau) and amyloid-beta (Aβ_{1-42}), which have been suggested as significant predictor for PD progression

	Progressed to H&Y stage 3	Not progressed to H&Y stage 3	Combined
Total	98	325	423
Age (years)	66.01 (9.40)	60.39 (9.44)	61.70 (9.72)
Education (years)	15.47 (2.74)	15.56 (3.07)	15.54 (2.99)
Male	58 (59.18%)	219 (67.38%)	277 (65.48%)
White	93 (94.90%)	308 (94.77%)	401 (94.80%)
Right handed	87 (88.78%)	288 (88.62%)	375 (88.65%)
Years of PD diagnosis (years)	0.62 (0.61)	0.53 (0.53)	0.55 (0.55)
MDS-UPDRS 1	7.67 (5.01)	4.93 (3.50)	5.57 (4.06)
MDS-UPDRS 2	7.86 (4.48)	5.30 (3.92)	5.89 (4.19)
MDS-UPDRS 3	24.09 (8.46)	19.93 (8.76)	20.89 (8.86)
MoCA	26.81 (2.51)	27.24 (2.25)	27.14 (2.32)
SDM	37.34 (9.49)	42.34 (9.51)	41.18 (9.73)
SEADL	91.84 (6.19)	93.54 (5.74)	93.14 (5.89)
QUIP	0.35 (0.73)	0.34 (0.89)	0.34 (0.86)
SCOPA-AUT	10.63 (5.75)	7.75 (5.25)	8.42 (5.50)
t-tau	49.29 (22.03)	43.36 (16.83)	44.69 (18.28)
p-tau	15.33 (10.30)	15.73 (9.99)	15.64 (10.05)
Aβ_{1-41}	347.36 (103.68)	377.33 (98.55)	370.56 (100.39)

Table 3.1: Descriptive statistics measured at baseline of PPMI participants.

[64–66], are considered as potential risk factors in our analysis. Baseline characteristics including patients' age, gender and years of PD diagnosis are controlled in the prediction model. We present the sample size and descriptive statistics for key baseline variables in Table 3.1.

3.3 Statistical Analysis

To model the time-to-event outcome (e.g. PD progression defined as HY stage reach 3) accounting for multiple available longitudinal biomarkers, we adopt a novel approach integrating features of longitudinal trajectories. By assuming there exists a latent process for observed measurements of each longitudinal biomarker, we employ functional principal component analysis (FPCA) approach to extract the features of each longitudinal process [38]. While the mean trajectory of a longitudinal marker is estimated by entire sample set, a set of subject-specific FPC scores for this biomarker is calculated to summarize the changing pattern in individual level. There are two advantages when utilizing FPC scores to capture longitudinal feature of subjects. First, by employing FPCA method, we do not assume a trajectory model. In most cases, it is hard to specify a proper parametric model for longitudinal process, and a miss-specified model will lead to biased estimation and inaccurate prediction. Second benefit is that while extracting features of longitudinal measurements, we use the observed values only, which means we allow subjects to have different missing visits from one to another, and no need to impute the missing values while calculating FPC scores. Moreover, with the eigenfunctions and eigenvalues obtained from FPC analysis, we are able to reconstruct the longitudinal process and estimate the value of biomarker at given time point. This feature of FPCA enables us to handle irregular missing visits among study subjects while provides a practical way to estimate missing values. However, one potential drawback of using FPCA on multiple markers separately is that, it is likely for the markers to have significant correlations between each other, which could cause the FPC scores computed from candidate markers become highly correlated and fail the independent covariates assumption in subsequent analysis. To solve this problem, We adopt multivariate functional principal component analysis (MFPCA) method proposed by Happ *et al* [61] to properly address this issue. Unlike using separate FPCA which fail to assess the joint variation of markers, MFPCA captures the joint variation of markers and addresses the potential correlation directly while estimating the matrix of covariances. As consequence, the MFPC scores derived from candidate markers are uncorrelated and more parsimonious when compared with separate FPC scores. To model the survival outcome integrating the impact of multiple longitudinal markers, we use the MFPC scores to represent the feature of trajectories and put them in the Cox-PH model as new risk factors.

3.3.1 MFPCA

To illustrate Multivariate FPCA method, we first introduce the univariate FPCA approach. The principal component model was first proposed by James *et al.* [67]. Let $X_i(t)$ denote the underlying latent longitudinal process for subject i, $i = 1, \cdots, n$. We denote the total follow up time of the study as τ. The observed survival time is denoted by T_i, which is the minimum of subject's survival time S_i and censoring time C_i. Let $Z_i(t) = Z_{i1}(t), \cdots, Z_{ir}(t)$ denote vector of candidate covariates valued at time t that have significant effects on longitudinal process, and $V_i(t) = V_{i1}(t), \cdots, V_{is}(t)$ represent the covariates valued at time t associated with recurrent event processes. Let $\mu(t)$ be the overall mean effect of longitudinal process without considering covariates $Z_i(t)$, then the mean function of longitudinal process is: $\mu_i(t) = \mu(t|Z_i) = \mu(t) + \beta Z_i(t)$, where β denotes the coefficients of covariates associated with longitudinal data. Define $G(s,t) = cov(X_i(s), X_i(t))$, and denote orthogonal eigenfunctions as $\phi_k, k = 1, 2, \cdots$, non-decreasing eigenvalues as $\lambda_k, k = 1, 2, \cdots$, then we can write $G(s,t) = \sum_k \lambda_k \phi_k(s) \phi_k(t)$. From Karhunen-Loéve theory [68], individual trajectory $X_i(t)$ can be expressed as $X_i(t) = \mu_i(t) + \sum_k \xi_{ik} \phi_k(t)$, here coefficients $\xi_{ik} = \int_0^\tau \{X_i(t) - \mu_i(t)\} \phi_k(t) dt$ are uncorrelated random variables with mean zero and variances $E\xi_{ik}^2 = \lambda_k$. By supposing that G can be approximated by first finite terms in eigen-decomposition, we can truncate the eigenfunctions and model the trajectory using first K leading principal components, $X_i(t) = \mu_i(t) + \sum_{k=1}^K \xi_{ik} \mu_k(t)$. Let Y_{ij} be the observation of longitudinal process at time t_{ij}, we can express the longitudinal observation Y_{ij} as:

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{n=1}^{M} \xi_{in} \phi_n(t_{ij}) + \epsilon_{ij}.$$

where ϵ_{ij} is measurement error with mean zero and variance σ^2 and is independent of ξ_{ik}. Note that we can model the eigenfunctions using expansions of a set of smooth basis functions, such as B-splines or regression splines.

The principal analysis by conditional estimation (PACE) algorithm was proposed by

Yao, Müller & Wang [38] to estimate the mean function $\mu(t_{ij})$, covariance function $G(s,t)$, eigenfunction ϕ_n and FPCA scores $\boldsymbol{\xi}_n$, where $n = 1, \cdots, M$. The PACE method was demonstrated to be powerful when applied to sparse longitudinal data with noise. Basically, the estimated mean function $\hat{\mu}(t_{ij})$ is obtained by a one-dimensional kernel smoother, following the estimation of mean function, the covariance matrix $\hat{G}$ can be estimated by a two-dimensional kernel smoother with pairwise products $\{Y_{ij} - \hat{\mu}(t_{ij})\}\{Y_{il} - \hat{\mu}(t_{il})\}$. Thus we are able to obtain eigenfunctions $\hat{\phi}_n$ and eigenvalues $\hat{\lambda}_n$, $n = 1, \cdots, M_j$, by spectral decomposition of covariance matrix $\hat{G}$. Based on the estimation results, the FPCA scores for longitudinal trajectory of subject i can be estimated by $\xi_{i,n} = \int (Z_i(t) - \mu(t))\phi_n(t)$. Other than integration, Yao, Müller & Wang [38] proposed another way to estimate FPCA scores by assuming $\xi_{i,n}$ and ϵ_{ij} are independent. Define $\Sigma_{Y_i} = cov(Z_i, Z_i) + \sigma^2 I$, it can be estimated following details given in Yao, Müller & Wang's paper [38].Thus, we can estimate univariate FPCA scores based on its conditional expectation, specifically:

$$\hat{\xi}_{i,n} = \hat{E}(\xi_{i,n}|Y_i) = \hat{\lambda}_n \hat{\phi}_n^{'}(t_{ij}) \hat{\Sigma}_{Y_i}^{-1}(Y_i(t_{ij}) - \hat{\mu}(t_{ij})).$$

MFPCA method proposed by Happ & Greven [69] provides a way to extract features of multiple longitudinal outcomes while capturing the joint variation between different outcomes. Let $X(t)$ denote the vector that combined p different functions, i.e, $X(t) =$

$(X^{(1)}(t), \cdots, X^{(j)}(t), \cdots, X^{(p)}(t))$, where $j = 1, \cdots, p$. Following the assumptions for univariate FPCA, multivariate Karhunen-Loéve theory give a representation as:

$$\begin{aligned} X(t) &= (X^{(1)}(t), \cdots, X^{(j)}(t), \cdots, X^{(p)}(t)) \\ &= \left(\sum_{m=1}^{\infty} \rho_m \psi_m^{(1)}(t), \cdots, \sum_{m=1}^{\infty} \rho_m \psi_m^{(j)}(t), \cdots, \sum_{m=1}^{\infty} \rho_m \psi_m^{(p)}(t) \right) \end{aligned}$$

here multivariate FPC score $\boldsymbol{\rho} = (\rho_1, \rho_2, \cdots, \rho_m, \cdots)$ are random variables with zero mean, ψ_m is an orthonormal basis of Hilbert space. The eigenvalue ν_m represents the amount of variability in X that explained by corresponding multivariate functional principal components ψ_m, while the multivariate FPC scores ρ_m serves as weights of ψ_m. As non-decreasing eigenvalues ν_m goes to 0 when $m \to \infty$, we can truncate the Karhunen-Loéve expression using first M components of ψ_m, M can be selected using model selection techniques (e.g. AIC, backward selection, proportion of variance explained, etc.). Next we give our approach for estimating multivariate FPCA based on properties given in section 3 of Happ & Greven. [69].

1. For each element $X^{(j)}, j = 1, \cdots, p$, estimate univariate FPCA based on observation for N subjects, as a result, eigenfunctions $\hat{\phi}_m^{(j)}$ and FPC score $\hat{\xi}_{i,m}^{(j)}$ can be estimated. Here, $i = 1, \cdots, N$, $m = 1, \cdots, M_j$ is the truncation parameter of element $X^{(j)}$.

2. Let $M_+ = \sum_{j=1}^{p} M_j$, define matrix $B \in \mathbb{R}^{N \times M_+}$, each row of B is a vector of estimated FPC scores of subject i for p-dimensional functional covariates, $B_{i,\cdot} = (\hat{\xi}_{i,1}^{(1)}, \cdots, \hat{\xi}_{i,M_1}^{(1)}, \cdots, \hat{\xi}_{i,1}^{(p)}, \cdots, \hat{\xi}_{i,M_p}^{(p)})$. Hence we can get the estimated Z matrix,

$\hat{Z}_{M^+ \times M^+} = (N-1)^{-1} B^\top B.$

3. We then perform spectral decomposition of $\hat{Z}$ matrix, resulting estimated eigenvalues $(\hat{v}_1, \cdots, \hat{v}_m, \cdots, \hat{v}_{M_+})$ and eigenvectors $(\hat{c}_1, \cdots, \hat{c}_m, \cdots, \hat{c}_{M_+})$.

4. Estimation of multivariate eigenfunctions are given as follows:

$$\hat{\psi}_m^{(j)}(t) = \sum_{n=1}^{M_j} [\hat{c}_m]_n^{(j)} \hat{\phi}_n^{(j)}(t),$$

here $[\hat{c}_m]_n^{(j)}$ is the nth$(n = 1, \cdots, M_j)$ element of jth$(j = 1, \cdots, p)$ block of eigenvector $\hat{c}_m(m = 1, \cdots, M_+)$, $\hat{\phi}_n^{(j)}(t)$ is the univariate FPC eigenfunction obtained from step 1. Multivariate FPC scores for each observation can also be calculated using following formula:

$$\hat{\rho}_{i,m} = \sum_{j=1}^{P} \sum_{n=1}^{M_j} [\hat{c}_m]_n^{(j)} \hat{\xi}_{i,n}^{(j)} = B_{i,.} \hat{c}_m,$$

here $i = 1, \cdots, N$, $m = 1, \cdots, M_+$, $\hat{\xi}_{i,n}^{(j)}$ is the estimated univariate FPC score obtained from step 1. Therefore, we are able to express multivariate functional covariates for subject i as $X_i(t)$ in terms of estimated multivariate eigenfunctions and scores,

$$X_i(t) = (X_i^{(1)}(t), \cdots, X_i^{(p)}(t)) = \left(\sum_{m=1}^{M_+} \hat{\rho}_{i,m} \hat{\psi}_m^{(1)}(t), \cdots, \sum_{m=1}^{M_+} \hat{\rho}_{i,m} \hat{\psi}_m^{(p)}(t) \right). \quad (3.1)$$

Following the theory foundation in aforementioned sections, we are able to estimate the MFPC scores, which is included in the proposed Cox regression model 1 to integrate multivariate longitudinal information.

3.3.2 Longitudinal marker selection

As PPMI study collected many longitudinal markers of participants during follow up period, we first investigate the prediction ability of each candidate markers in order to select appropriate markers in disease progression prognostic model. Joint modeling method is employed to simultaneously model time to reaching H&Y stage 3 and candidate longitudinal markers, and assess the association strength of each markers and survival outcome [28, 70]. The joint model consists of two sub-models, one is the linear mixed-effects model for longitudinal data, another is the Cox regression model for survival data. The time metric for both models is years from enrollment. As the intention is to assess the prognostic ability of different markers, we standardize each longitudinal outcome before fitting the model to make them comparable. The association parameter from the fitted joint model is used to capture the relation between longitudinal marker and survival outcome, which in our case, quantifies the prediction ability of time to disease progression for candidate markers. An significant estimation of association parameter meant that the marker accounts for the variation of time to H&Y stage transition to 3. In all models, we control for baseline covariates: gender, race, age, education years, years of PD diagnosis and handedness. The joint modelling results are presented in Table 3.2, with longitudinal markers ranked by the absolute value of association parameter's Z-score. Candidate markers ranked within top 5 are selected to be the longitudinal predictor included in our proposed prognostic model, i.e., MDS-UPDRS 3, MDS-UPDRS 2, QUIP, SDM, and SCOPA-AUT.

Longitudinal marker	N	Association (SE)	P-value	Z-Score
MDS-UPDRS 3	417	1.124(0.011)	<0.001	6.413
MDS-UPDRS 2	417	0.730(0.007)	<0.001	5.577
QUIP	410	-3.026(0.162)	<0.001	-3.226
SDM	410	-0.378(0.004)	0.002	-3.164
SCOPA-AUT	410	0.328(0.008)	0.007	2.715
SEADL	417	-0.361(0.009)	0.025	-2.279
MDS-UPDRS 1	417	0.309(0.009)	0.031	2.120
MoCA	417	-0.244(0.006)	0.060	-1.973
Derived LNS	410	-0.201(0.005)	0.140	-1.455
GDS15	410	0.106(0.008)	0.476	0.715
Semantic Fluency	410	0.031(0.009)	0.856	0.172

Table 3.2: Prediction of time to H&Y stage reaching 3 using longitudinal marker: joint modelling results.

3.3.3 MFPCA based Cox regression model

To predict the time to reaching H&Y stage 3, we specify our proposed Cox regression model integrating the longitudinal information from selected 5 markers by extracting features using MFPCA method (referred as Model 1). To ensure at least 99% of total variation is explained, we select the first 10 MFPC scores as the extracted features of those 5 longitudinal markers. For comparison purpose, we also specify another Cox regression model with only subject's demographic variables and baseline measurements of selected 5 markers (referred as Model 2). Comparing the prediction performance of these two models help us assess the gain from incorporating longitudinal information. Before conducting the analysis, we fit a separate Cox model to select significant baseline demographic variables, age, gender, baseline t-tau, and baseline Aβ_{1-41} are identified as significant predictors. Due to the fact that the exact date of patient transit to H&Y stage 3 is unknown, we define the time to event outcome as time to the first visit that patient was rated as H&Y stage 3 while subjects who did not reach H&Y stage 3 yet are considered as censored. Controlling for common baseline variables, we specify the hazard function for Model 1 (Equation 1) and Model 2 (Equation 2) as following:

$$h_i(t) = h_0(t)\exp\{\beta_1\text{Age}_i + \beta_2\text{Male}_i + \beta_3\text{t-tau}_i + \beta_4\text{A}\beta_i + \sum_{k=1}^{10}\alpha_k\text{MFPCscores}_{ik}\} \tag{3.2}$$

$$\begin{aligned} h_i(t) = h_0(t)\exp\{&\beta_1\text{Age}_i + \beta_2\text{Male}_i + \beta_3\text{t-tau}_i + \beta_4\text{A}\beta_i + \beta_5\text{MDS-UPDRS3}_i \\ &+ \beta_6\text{MDS-UPDRS2}_i + \beta_7\text{QUIP}_i + \beta_8\text{SDM}_i + \beta_9\text{SCOPA-AUT}_i\}. \end{aligned} \tag{3.3}$$

Here, $h_0(t)$ represents the baseline hazard function, $\boldsymbol{\beta}$ denotes the coefficients associated with risk factors, and $\boldsymbol{\alpha}$ denotes the regression coefficients for MFPC scores. To ensure we have sufficient observations for each longitudinal markers to calculate MFPC scores, we exclude subjects only have less or equal than two available longitudinal measurements for any of the markers.

We assess the prediction performance of the previous specified models from two aspect. To assess the global discrimination ability, we employ the integrated area under the time-dependent receiver operating characteristic curve (iAUC) proposed by Uno *et al* [71], which is based on inverse-probability-of-censoring weights. On the other hand, we evaluate the calibration ability of prognostic models by integrated Brier score (BS) [72]. Moreover, while iAUC and BS are well studied in quantifying the overall prediction performance, comparing the time-dependent AUC between the two pre-specified models is also important. We employ the methodology approach proposed by Blanche *et al* [73] to compute confidence region of time-dependent AUC and test for the difference between them in the two aforementioned prognostic models. Regards of prediction, we use 10-

fold cross validation (CV) strategy to avoid overestimation. We repeatedly conduct CV in PPMI dataset for 100 times using different random seed to reduce the variation caused by data splitting. The internal iAUC and BS for both models are calculated to assess the improvement when integrating longitudinal information in prognostication, and the confidence region of difference of time-dependent AUC from the two models is calculated to test for significance. With the intention to establish an index that can be used across studies, we computed the prognostic index (PI) from the MFPCA based Cox regression Model 2. The formula for calculating prognostic index is based on the estimated regression coefficients:

$$\mathrm{PI} = \beta_1 \mathrm{Age}_i + \beta_2 \mathrm{Male}_i + \beta_3 \text{t-tau}_i + \beta_4 \mathrm{A}\beta_i + \sum_{k=1}^{10} \alpha_k \mathrm{MFPCscores}_{ik}. \tag{3.4}$$

We use data from LABS-PD to conduct external validation. Since LABS-PD did not collected all the top 5 ranked longitudinal markers in PPMI (MDS-UPDRS 3, MDS-UPDRS 2, QUIP, SDM, and SCOPA-AUT), we rank the available markers in both studies by their absolute Z-score and select the top 5 to be included in validation prognostic models, i.e., MDS-UPDRS 3, MDS-UPDRS 2, MDS-UPDRS 1, SEADL, MDS-UPDRS 1, MoCA. In external validation session, we refer the two Cox regression models as The Model 1a and Model 2a, which are similar to Equation 1 and Equation 2 respectively, with the only difference that instead of using the optimal 5 longitudinal markers, we use the 5 top ranked markers among available ones in both studies. We exclude the subjects who only measured less or equal than twice for any of these longitudinal markers during follow up. To better compare the prediction performance of proposed models across studies, we first conduct internal validation for Model 1a and Model 2a in PPMI and LABS-PD separately. The external validation is conducted by applying models fitted from PPMI data on LABS-PD data. Prediction performance is evaluated by iAUC and BS. Based on the Model 2a fitted in PPMI data, we establish the formula for calculating prognostic index using regression coefficients. We then apply the formula on LABS-PD data to assess the ability of categorizing risk groups across studies.

3.4 PPMI Study Application Results

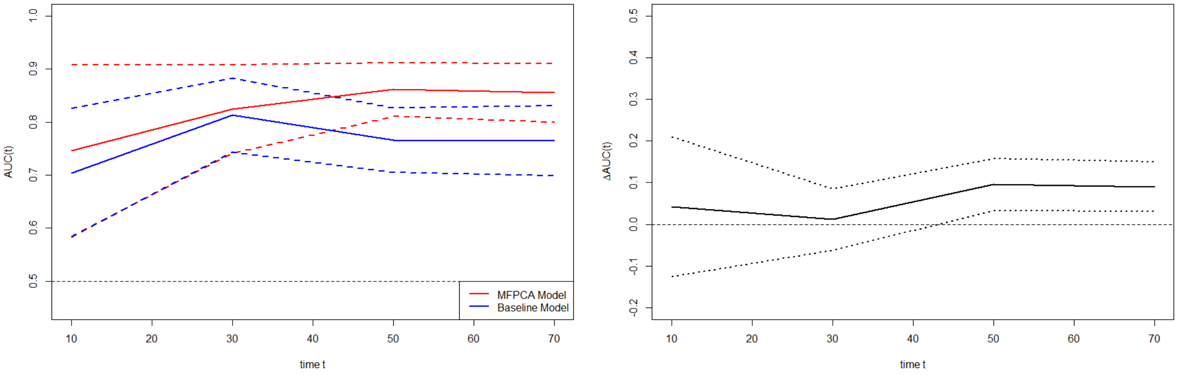
The results of the prognostic models are summarized in Table 3.3. While the the internal prediction performance index (iAUC$_{INT}$) for baseline Cox regression Model 2 on PPMI is 0.758, the internal iAUC for MFPCA based Cox regression Model 1 gets 5 percentage point higher by incorporating longitudinal information of selected markers. Similarly, the internal Brier Score (BS$_{INT}$) for Model 1 is 0.094, which is about 1 percentage point lower than Model 2, indicating the bias between predicted and true risk is smaller in Model 1. To illustrate the comparison results, we plot the estimated time-dependent AUC of MFPCA model (Model 1) and Baseline Model (Model 2) as well as the difference of mean AUC from these two prognostic models in Figure 3.1. From the AUC plot (left panel), We find that the estimated time-dependent AUC is higher for MFPCA model when compared to Baseline model at all landmark time. Moreover, the right panel shows the difference of mean AUC is alway above 0 across time, and the confidence interval does not contain 0 after around 40 months, indicating the AUC difference between two prognostic models is significant when in later prediction time. As the analysis results favor Model 1 in regards of higher iAUC and lower BS, we select Model 1 to establish the prognostic index. Based on the regression coefficients estimated from Model 1, we follow formula in Equation (3) to calculate PI for each subject in PPMI. We then use the PI quartiles (50%, 75%, 100%) to categorize subjects into three risk groups (high, mid and low). For comparison purpose, we also calculate a similar prognostic risk score based on Model 2 and classify subjects into different risk groups according to calculated risk scores. The Kaplan-Meier (K-M) curves for PD progression risk groups categorized from two prognostic models are presented in Figure 3.2. Comparing the Kaplan-Meier curves for the 3 risk groups classified according to PI from Model 1 (left panel) and the curves for risk groups classified based on Model 2 (right panel), we find PI quartiles from Model 1 separate risk groups better in the sense of non over-lapping survival probability confidence band in later follow-up time, which agrees with the difference of estimated AUC plot in Figure 3.1. The finding suggests that incorporating longitudinal information in prognostic model help to identify subjects with higher risk of reaching H&Y stage 3 after enrollment. We then apply the proposed prognostic models in LABS-PD to assess the prediction performance across studies. Due to the availability of longitudinal markers in LABS-PD, we choose the 5 top ranked markers that is collected in both PPMI and LABS-PD studies instead of using the optimal

5 longitudinal markers. The MFPCA based Cox regression model adjusted for data availability is referred as Model 1a, and the corresponding prognostic model only includes baseline measurements is referred as Model 2a. To establish a benchmark for comparison, we conduct 10-fold class internal validation in PPMI and LABS-PD respectively, after that the external validation is carried out in LABS-PD using models fitted in PPMI dataset. Results for internal and external validation are summarized in Table 3.3. The iAUC calculated from Model 1a, which includes the top 5 available markers, is 0.797, about 1 percentage point lower than iAUC for Model 1 including optimal 5 markers. When compared with the baseline Cox regression Model 2a, mean iAUC estimated from Model 1a incorporating longitudinal information is about 5 percentage point higher. In internal validation for LABS-PD study, iAUC for Model 1a is 0.806, which is about 8 percentage higher than Model 2a which only accounts for baseline measurements. For external validation, we fit the proposed models in PPMI, and apply the prognostic models on LABS-PD data. As a result, Model 1a is also favored with higher iAUC and lower BS. The consistent internal and external findings suggests Model 1a has a better prognostic ability regarding prediction of time to H&Y stage reach 3 across studies. When it comes to the comparison between iAUC from internal and external validation in LABS-PD, we find the iAUC of external validation is 8 percentage point lower in Model 1a, while the difference is 2 percentage point in Model 2a. This could be caused by the nature of the two studies. While PPMI only recruited early, untreated PD patients, the 537 subjects in LABS-PD PostCEPT were part of participants from previous conducted PRECEPT

study and not necessarily de-novo patients. Figure 3.3 shows the Kaplan-Meier plot for the high, mid and low risk groups of PD progression in PPMI and LABS-PD respectively. The risk curves for LABS-PD is similar to the curves for PPMI with the top curve for low risk subjects, bottom curve for high risk group and the curve in the middle for mid risk group. The plots indicate the PI formula established in PPMI can be well applied to LABS-PD to classify subjects with different level of progression risk. Moreover, the low risk group in LABS-PD study progresses faster than the same risk group in PPMI, which is consist with the characteristic of LABS-PD subjects that they have longer PD disease history than PPMI subjects and tend to progress in nearer future. With the calculated PI, We are able to estimate the survival probability for subjects in different risk groups based on the Kaplan-Meier curves in Figure 3.3 correspondingly. Also, with the approximated baseline risk function (e.g. using piece-wise constant or splines) we are able to calculate the survival risk of time to H&Y stage reaching 3 for specific subject.

	Model 1	Model 2	Model 1a		Model 2a	
Study	iAUC$_{INT}$	iAUC$_{INT}$	iAUC$_{INT}$	iAUC$_{EXT}$	iAUC$_{INT}$	iAUC$_{EXT}$
PPMI (n=375)	0.808	0.758	0.797	-	0.752	-
LABS-PD (n=459)	-	-	0.806	0.720	0.721	0.701
	BS$_{INT}$	BS$_{INT}$	BS$_{INT}$	BS$_{EXT}$	BS$_{INT}$	BS$_{EXT}$
PPMI (n=375)	0.094	0.106	0.089	-	0.097	-
LABS-PD (n=459)	-	-	0.184	0.165	0.292	0.172

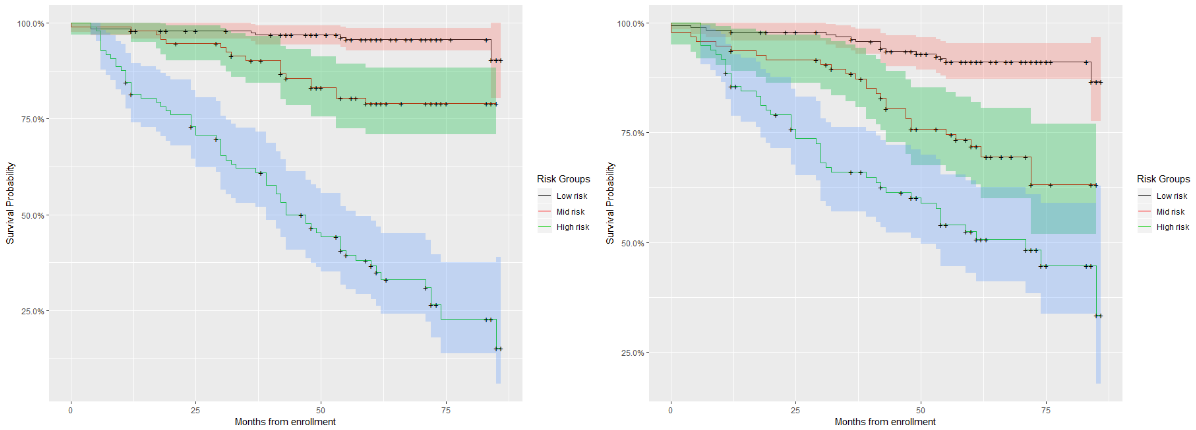
Table 3.3: Comparison of prognostic models in internal validation and external validation. Model 1 is the proposed prognostic model with optimal 5 markers, Model 2 is the corresponding baseline prognostic model. Model 1a and Model 2a are similar models with top 5 ranked available markers.



(a) Estimated AUC plot over time

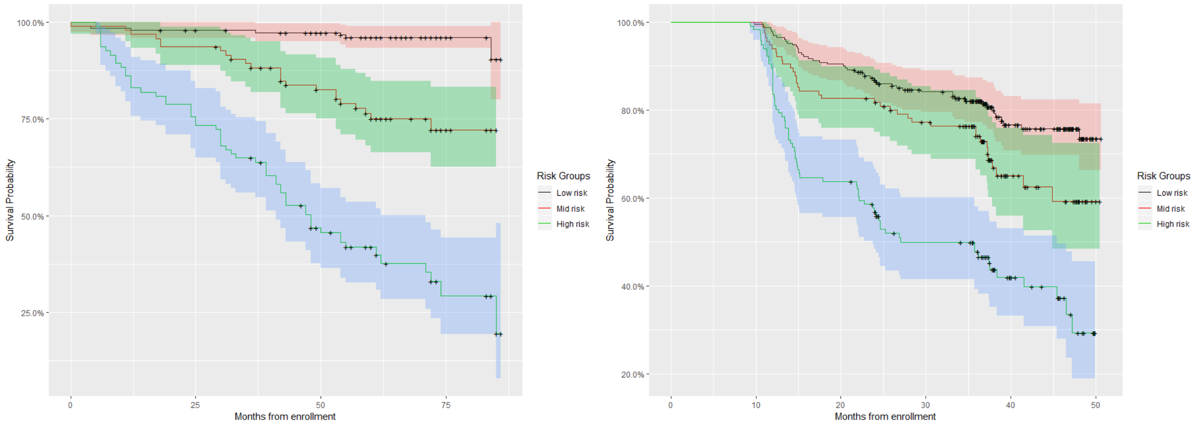
(b) Difference of estimated AUC plot over time

Figure 3.1: Time-dependent AUC plot for Model 1 and Model 2.



(a) Risk groups categorized based on Model 1 (b) Risk groups categorized based on Model 2

Figure 3.2: Kaplan-Meier plot for risk groups of PD progression to H&Y stage 3.



(a) Kaplan-Meier plot for subjects in PPMI (b) Kaplan-Meier plot for subjects in LABS-PD

Figure 3.3: Kaplan-Meier plot for risk groups of PD progression to H&Y stage 3 in PPMI and LABS-PD.

Chapter 4

Article 3: Dynamic prediction of recurrent events using multiple longitudinal outcomes

Dynamic prediction of recurrent events using multiple longitudinal outcomes

4.1 Introduction

Over the past decades, the prevalence of cardiovascular disease (CVD) has motivated systematic investigation into its risk factors, prevention strategy and treatment method. As the first large CVD study with longitudinal follow up visits, Framingham Heart Study (FHS) has been collecting longitudinal clinical outcomes at each schduled exam since 1948 [74]. Due to the emerging evidence of the role of shared familial factors in developing CVD, researchers are interested in study the next generation of participants. FHS investigators recruited the Offspring Cohort, which consists of a group of children of Original Cohort members with coronary disease and the spouses of those children [75]. Starting from 1972, the FHS Offspring Cohort conducted in-person examinations for participants to collect longitudinal measurements including systolic blood pressure (SBP), body mass index (BMI), blood glucose and etc. Although there are numerous literature aiming at identifying risk factors for CVD, a growing interest lay in the prediction of the occurrence of cardiovascular events. As the primary outcome of FHS Offspring Cohort is the CVD events, which have feature of recurrence (i.e. the event of interest often occurs multiple times from the same subject), the commonly used Cox proportional hazard model will not be appropriate, since it ignores the intrinsic correlation between those repeated occurrences within the same subject [21]. To avoid the biased estimated risk, proposing an effective prediction model for the next-occurrence of CVD event is of importance here. Moreover, as the recurrent CVD events tend to have significant association with clinical longitudinal markers, incorporating those rich longitudinal information in prediction models meets the urgent need of improving the prognostic performance. A few of studies has investigated joint analysis of longitudinal marker and recurrent event [28, 32, 36, 37], however, utilizing information from multiple longitudinal markers to predict the risk of re-occurrence of events remains an open question. Motivated by the FHS Offspring data, our study focuses on the joint analysis of multiple longitudinal clinical markers and recurrent CVD events. A prediction model is proposed to estimate the risk of having another recurrent CVD event in given time interval. Moreover, we develop a dynamic prediction framework for the next occurrence of CVD event using subject-specific longitudinal profiles, which enables physicians give personalized and precision treatment to patients.

Joint models (JM) of longitudinal measurements and survival data (including terminal event and recurrent event data) have been a popular method to analyze clinical measurements and event of interest together in the past decades. Among the existing literature, Henderson [28] proposed a joint model of longitudinal data and recurrent event data, with association between two process captured by correlated latent trajectory. Liu and Huang [32] established a JM approach where a more complex setting was considered, i.e. a repeated measures process and a recurrent events process were correlated, both subject to a terminal event. In terms of prediction of terminal event risk under JM framework, Krol *et al* [36] extended the usage of joint model of longitudinal data, recurrent events, and a terminal event to dynamic prediction area. As a result of their work, the probability of having a terminal event in specific time interval is estimated, given previous longitudinal profile and recurrent event history. In recurrent events prediction area, Musoro *et al* [37] employed landmark method to handle the longitudinal data as time-fixed covariate at different landmark time point, and in this way extended dynamic prediction by landmarking to recurrent event data. However, these publications only investigated the prediction ability of single longitudinal outcome. While it is challenging to specify appropriate parameter distributions when modelling multiple longitudinal trajectories, the major limitation occurs in extending the JM approach to incorporate multivariate longitudinal outcomes. In addition to the issue in statistical aspect, the computation intensity becomes another concern when joint model under Bayesian framework is considered, especially with a large number of candidate longitudinal markers.

To avoid involving in the potential issues when extending JM framework, we consider a novel approach which is more flexible, robust and time efficient. Since Yao [38] proposed a nonparametric approach to perform functional principal components analysis (FPCA) on sparse longitudinal data, researchers has extended the usage of FPCA to joint analysis of repeated measurements and survival data [39]. In the two-step approach proposed by Holte *et al* [40], the feature of longitudinal trajectory was extracted and represented by the functional principle component (FPC) scores estimated from separate model, and the estimated FPC scores were included in the survival model as new risk factors to build association between longitudinal trajectory and the survival outcome. Yan *et a* [41] extended the FPCA framework further to dynamic prediction area, their work enables the estimated FPCA scores can be updated as new longitudinal information come into available. However, these studies focused only incorporating information from single longitudinal marker in the survival model, moreover, the Cox model they used to model the terminal event is not suitable when the event of interest has recurrent feature (e.g. CVD events). To our best knowledge, no existing literature has investigated the prediction of recurrent event utilizing information from multiple longitudinal markers. In this article, we develop an novel approach to fit this critical need. In the proposed approach, we first extract the feature of multiple longitudinal trajectories using multivariate FPCA (MFPCA) method, and then fit a recurrent event model with intensity function incorporating the feature scores from longitudinal markers. The dynamic prediction for the risk of next occurrence of recurrent event is also derived to conduct personalized prediction based on subject-specific longitudinal measurements.

The rest of the article is organized as follows. In Section 2, we illustrate the modeling framework with detailed description of MFPCA method that used to extract features of longitudinal outcomes, the recurrent event model with Poission intensity function, and the dynamic prediction derivation. In Section 3, the proposed approach is applied on the motivating FHS Offspring Cohort, and the estimated AUC is compared across candidate models. The simulation results are presented in Section 4 to assess the prediction performance of our proposed model under different scenarios. In the last section, the concluding remarks and future research directions are discussed.

4.2 Methods

In this section, we give a detailed description of the proposed approach to model the recurrent CVD events incorporating information from multiple longitudinal outcomes. Considering a study of N subjects. Let $\{Y_{ijq}\} = \{Y_{iq}(t_{ij})\}$ be the vector of longitudinal observation of the q-th ($q = 1, \cdots, Q$) clinical marker for subject i at time t_{ij}, where $i = 1, \ldots, N$, and $j = 1, \ldots, m_i$. Assuming there exists an underlying latent trajectory $X_{iq}(t)$ for each observed longitudinal measurement Y_{iq}, we denote the corresponding error term as ϵ_{ijq} and re-write the longitudinal observation as $Y_{ij}(t_{ij}) = X_{iq}(t_{ij}) + \epsilon_{ijq}$, where $epsilon_{ijq}$ has zero mean and variance $\sigma^2_{\epsilon_q}$. Let T_{ik} be the recurrent event times from study onset for subject i, $k = 0, \ldots, n_i$, where n_i denotes the total number of recurrent CVD events subject i experienced.

4.2.1 multivariate FPCA method and implement

Let $\boldsymbol{X}_i(t)$ denote the vector that combined q different functions, i.e, $\boldsymbol{X}_i(t) = (X_{i1}(t), \cdots, X_{iq}(t), \cdots, X_{iQ}(t))$, where $q = 1, \cdots, Q$. Before getting into multivariate FPCA, We first illustrate univariate FPCA method for single longitudinal outcome q. We denote the smoothed mean function of $X_{iq}(t)$ as $\mu_q(t)$, and the covariance function which models correlation of longitudinal outcome q's latent trajectory between time t and t' as $\Sigma_q(t,t') = cov\{X_{iq}(t), X_{iq}(t')\}$. Using spectral decomposition, the covariance function can be re-write as $\Sigma_q(t,t') = \Sigma_{n=1}^{\infty}\lambda_{qn}\phi_{qn}(t)\phi_{qn}(t')$, with $\{\lambda_{qn}\}$ represent the non-increasing eigenvalues for outcome q, and $\{\phi_{qn}(t)\}$ denote the corresponding eigenfunctions. The Karhunen-Loéve theory give a representation of $X_{iq}(t)$ as:

$$X_{iq}(t) = \mu_q(t) + \sum_{n=1}^{\infty} \xi_{iqn}\phi_{qn}(t). \tag{4.1}$$

Here, $\{\xi_{iqn}\}$ is the set of uncorrelated FPC scores with 0 mean and variance λ_{qn}. The FPC scores describes how the observed profile of subject i follows the changing pattern $\phi_{qn}(t)$, and can be viewed as the extracted feature of subject i. Moreover, the latent trajectory can be approximated with finite terms as $X_{iq}(t) \approx \mu_q(t) + \sum_{n=1}^{N_q} \xi_{iqn}\phi_{qn}(t)$, where N_q can be selected according to pre-specified percentage of variance explained (PVE). As the actual observed longitudinal data for outcome q is usually only available for irregular visit time and tend to be sparse data, we employ the principle analysis by conditional estimation (PACE) [38] to conduct analysis. We can express the longitudinal observation Y_{ijq} as:

$$Y_{ijq} = X_{iq}(t_{ij}) + \epsilon_{ijq} = \mu_q(t_{ij}) + \sum_{n=1}^{N_q} \xi_{iqn}\phi_n(t_{ij}) + \epsilon_{ijq}. \tag{4.2}$$

The PACE method was demonstrated to be powerful when applied to sparse longitudinal data with noise. We use it to estimate the mean function $\mu_q(t_{ij})$, covariance function $\Sigma_{iq}(t,t')$, eigenfunctions $\{\phi_{qn}\}$ and FPCA scores $\boldsymbol{\xi}_{iqn}$, where $n = 1, \cdots, N_q$. Basically, the estimated mean function $\hat{\mu}_q(t_{ij})$ is obtained by a one-dimensional kernel smoother. Following the estimation of mean function, the covariance matrix $\hat{\Sigma}$ can be estimated by a two-dimensional kernel smoother with pairwise products $\{Y_{ijq} - \hat{\mu}_q(t_{ij})\}\{Y_{ilq} - \hat{\mu}_q(t_{il})\}$. Thus we are able to obtain eigenfunctions $\hat{\boldsymbol{\phi}}_{qn}$ and eigenvalues $\hat{\lambda}_{qn}$, $n = 1, \cdots, N_q$, by spectral decomposition of covariance matrix $\hat{\Sigma}_{iq}$. Based on the estimation results, the FPCA scores for longitudinal trajectory of subject i can be estimated based on its conditional expectation, specifically:

$$\hat{\xi}_{iqn} = \hat{E}(\xi_{iqn}|Y_{iq}) = \hat{\lambda}_{qn}\hat{\boldsymbol{\phi}}_{qn}^{'}\hat{\Sigma}_{Y_{iq}}^{-1}(\boldsymbol{Y}_{iq} - \hat{\boldsymbol{\mu}}_q). \tag{4.3}$$

Since the estimated FPC scores from q longitudinal clinical outcomes can be correlated due to their intrinsic relationship, we use MFPCA method to extract features of multiple longitudinal outcomes while capturing the joint variation between different outcomes [69]. Let $M_+ = \sum_{q=1}^{Q} N_q$, define matrix $B \in \mathbb{R}^{N \times M_+}$, each row of B is a vector of estimated FPC scores of subject i for q-th outcome, $B_{i,\cdot} = (\hat{\xi}_{i11}, \cdots, \hat{\xi}_{i1N_q}, \cdots, \hat{\xi}_{iQ1}, \cdots, \hat{\xi}_{iQN_Q})$. Hence we can get the estimated Z matrix, $\hat{Z}_{M^+ \times M^+} = (N-1)^{-1}B^\top B$. We then perform spectral decomposition of $\hat{Z}$ matrix, resulting estimated eigenvalues $(\hat{v}_1, \cdots, \hat{v}_m, \cdots, \hat{v}_{M_+})$ and eigenvectors $(\hat{c}_1, \cdots, \hat{c}_m, \cdots, \hat{c}_{M_+})$. Estimation of multivariate eigenfunctions are given as $\hat{\psi}_{mq}(t) = \sum_{n=1}^{N_q} [\hat{c}_m]_{nq} \hat{\phi}_{nq}(t)$, here $[\hat{c}_m]_{nq}$ is the nth$(n = 1, \cdots, N_q)$ element of q-th$(q = 1, \cdots, Q)$ block of eigenvector $\hat{c}_m(m = 1, \cdots, M_+)$, $\hat{\phi}_{nq}(t)$ is the univariate FPC eigenfunction obtained from training dataset. Multivariate functional principle component (MFPC) scores for each observation can also be calculated using following formula:

$$\hat{\rho}_{im} = \sum_{q=1}^{Q} \sum_{n=1}^{N_q} [\hat{c}_m]_{nq} \hat{\xi}_{iqn} = B_{i,.} \hat{c}_m, \tag{4.4}$$

here $m = 1, \cdots, M_+$, $\hat{\xi}_{iqn}$ is the estimated univariate FPC score obtained from step 1. We select first $D \leq M^+$ FPC scores based on pre-specified PVE or other criterion such as Akaike information criterion (AIC) to represent the features extracted from multivariate longitudinal trajectories, and use them in the recurrent event model. Therefore, we are able to express the underlying trajectories of q-th longitudinal outcome for subject i in terms of estimated multivariate eigenfunctions and scores:

$$X_{iq}(t) \approx \ \hat{\mu_q}(t) + \Sigma_{m=1}^{D} \hat{\rho}_{im} \hat{\psi}_{qm}(t) \tag{4.5}$$

4.2.2 Recurrent event model

Assuming that the number of recurrent events in non-overlapping time intervals is a poisson process, we are able to model the event process via intensity function. To integrate multivariate longitudinal information, we use the estimated MFPC scores $r\hat{h}o_{im}$ as predictors in the intensity function, to model the relationship between longitudinal outcomes and recurrent events. Following previous notation, the intensity function of recurrent events for subject i is:

$$r_i(t) = r_0(t)\exp\{\boldsymbol{Z}_i^R\boldsymbol{\beta} + \hat{\boldsymbol{\rho}}_i\boldsymbol{\gamma} + v_i\}. \tag{4.6}$$

Here, the covariate vector $\boldsymbol{Z}_i^R$ is the time-independent covariates and $\boldsymbol{\beta}$ is the corresponding coefficient vector, $\boldsymbol{\gamma}$ is the vector of coefficient for the estimated MFPC scores $\{\hat{\rho}_{im}\}$, $m = 1, \cdots, D$, v_i is the random effect follows normal distribution $N(0, \sigma_v)$. We utilize piece-wise constant baseline hazard model to obtain estimators for both fixed effects and random effects, which increase the robustness of our model fitting. The total follow-up time interval $[0, \tau]$ is divided using time knots $\tau_t = (0, \tau_1, ..., \tau_R)$ by quantile of event times, and the baseline hazard vector is denoted as $g = (g_0, g_1, ..., g_{R-1})$. Then we can define the piecewise constant hazard function as $h_0(t) = \sum_{r=0}^{R-1} g_r I_r(t)$, where indicator function $I_r(t) = 1$, if $\tau_r \leq t < \tau_{r+1}$ and 0 if otherwise. Therefore, conditioning on random effect v_i, the likelihood of the recurrent events process for subject i is:

$$\begin{aligned} l_i^R &= \prod_{k=0}^{n_i} r_i(t_{ik})^{\sigma_{ik}} S_i(x_i) \\ &= \prod_{k=0}^{n_i} \Big[r_0(t_{ik}) \exp\big\{\boldsymbol{Z}_i^R\boldsymbol{\beta} + \hat{\boldsymbol{\rho}}_i\boldsymbol{\gamma} + v_i\big\}\Big]^{\delta_{ik}} \cdot \exp\Big[- \int_0^{\tau_i} r_0(t)\exp\big\{\boldsymbol{Z}_i^R\boldsymbol{\beta} + \hat{\boldsymbol{\rho}}_i\boldsymbol{\gamma} + v_i\big\} dt \Big], \end{aligned}$$

where δ_{ik} is the indicator of a recurrent event at time t_{ik} and τ_i is the observed follow-up time. Thus, the full likelihood for subject i is

$$l_i = l_i^R \cdot f(v_i), \tag{4.7}$$

where $f(v_i)$ is the density function of v_i. The unknown parameter vector is $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_v, r_0(t)\}$.

4.2.3 Dynamic prediction

In order to extend the proposed model framework to dynamic prediction of event risk, we randomly partition the dataset into two part, one is the training datasest which is used to build the model, another is the validation dataset to access the prediction performance of proposed model. We first apply our proposed MFPCA approach on training dataset and estimate parameters and MFPC scores $\hat{\boldsymbol{\rho}}_{\boldsymbol{m}}$ using all available longitudinal observations. With the estimated MFPC scores and time-independent covariates, we fit recurrent event model (4.6) in training dataset. After obtaining the estimated parameter vectors $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{r}_0(t)$, we then illstrate the derivation and procedure to conduct prediction in validation dataset. Suppose a new subject L had n_L (e.g., $n_L = 0, 1, 2, \cdots$) recurrent events up to time t, with q longitudinal outcomes $\boldsymbol{Y}_{Lq} = \{Y_{Lq}(t_{Lj}); 0 \leq t_{Lj} \leq t\}$, we fist compute the FPC scores for the q-th outcome using equation (4.3), specifically, $\xi_{Lqn} = \hat{\lambda}_{qn}\hat{\boldsymbol{\phi}}_{qn}^{'}\hat{\Sigma}_{Y_{Lq}}^{-1}(\boldsymbol{Y}_{Lq} - \hat{\boldsymbol{\mu}}_q)$. With the estimated FPC scores for each of the longitudinal outcomes, We then compute the MFPC scores $\hat{\boldsymbol{\rho}}_L$ for subject L following equation (4.5).

Our goal is to predict his/her probability of having the $n_L + 1$ recurrent event before time $t' = t + \Delta t$ (e.g., 1 year), denoted by $\pi_L(t'|t) = P(T_{L,n_L+1} \leq t' | T_{L,n_L+1} > t, \boldsymbol{Z}_L, \hat{\boldsymbol{\rho}}_L, \boldsymbol{\theta})$. The probability $\pi_L(t'|t)$ can be derived as follows:

$$\begin{aligned} &\pi_L(t'|t) \\ &= \int P(T_{L,n_L+1} \leq t' | T_{L,n_L+1} > t, \boldsymbol{Z}_L, \hat{\boldsymbol{\rho}}_L, \boldsymbol{\theta}, v_L) P(v_L | T_{L,n_L+1} > t, \boldsymbol{\theta}) dv_L \\ &\approx \frac{1}{M} \sum_{m=1}^{M} 1 - \exp\left[- \int_t^{t'} r_L^{(m)}(s | v_L^{(m)}, \boldsymbol{Z}_L, \hat{\boldsymbol{\rho}}_L, \boldsymbol{\theta}^{(m)}) ds \right]. \end{aligned} \tag{4.8}$$

When new information come into available, the estimated prediction probability $\pi_L(t'|t)$ can be updated, which leads to dynamic risk prediction that change over time when more data accumulate. We first update the estimated MFPC scores $\hat{\boldsymbol{\rho}}_L$ using enriched multivariate longitudinal information. The updated MFPC scores reflect the trends that altered by new observations. Then we update the estimated recurrent event risk. Here $\boldsymbol{\theta}^{(m)}$ is the mth sample ($m = 1, \ldots, M$, where M is the number of post burn-in samples) of parameter vector $\boldsymbol{\theta}$. For random effect, $\boldsymbol{v}_L^{(m)}$ is the mth sample of $\boldsymbol{v}_L$. The term $r_L^{(m)}$ denotes the intensity function from poission process conditioning on mth copy of $\boldsymbol{\theta}$ and corresponding random effect $\boldsymbol{v}_L$. To approximate the event probability $\pi_L(t^{'}|t)$, we need to obtaining samples for random effect $\boldsymbol{v}_L$. The posterior samples of v_L is drew from its posterior distribution $P(v_L|T_{L,n_L+1} > t, \boldsymbol{\theta})$ as follows:

$$\begin{aligned} P(v_L|T_{L,n_L+1} > t, \boldsymbol{\theta}^{(m)}) &= \frac{P(T_{L,n_L+1} > t, v_L|\boldsymbol{\theta}^{(m)})}{P(T_{L,n_L+1} > t|\boldsymbol{\theta}^{(m)})} \propto P(T_{L,n_L+1} > t, v_L|\boldsymbol{\theta}^{(m)}) \\ &= P(T_{L,n_L+1} > t|v_L, \boldsymbol{\theta}^{(m)})P(v_L|\boldsymbol{\theta}^{(m)}), \end{aligned}$$

It is important to access how well our proposed risk prediction model performs. Here, we access the prediction performance in terms of global discrimination ability. To be more specific, we employ receiver operating characteristic (ROC) curve and the area under the ROC curves (AUC) to assess the discrimination ability of the proposed model. Following the definition in previous section, for a given cut value $c \in [0, 1]$, the time dependent sensitivity and specificity are $P\{\pi_i(t'|t) > c|D(t, t') = 1, T_i^* > t\}$ and $P\{\pi_i(t'|t) \leq c|D(t, t') = 0, T_i^* > t\}$ respectively, where $D(t, t')$ is an indicator function equals to 1 when a new event happen during time interval (t,t'] and equals to 0 otherwise. Therefore, for probability $p \in [0, 1]$, the ROC curves will be $ROC_t^{t'}(p) = TP_t^{t'}[FP_t^{t'}]^{-1}(p)$, where $TP_t^{t'}$ denotes the true positive rate, $FP_t^{t'}$ denotes the false positive rate [56]. With the defined time-dependent sensitivity and specificity, we calculate a standard "concordance" summary: the time-dependent Area Under Curve (AUC) [57], the formula is $AUC(t, t') = \int_0^1 ROC_t^{t'}(p)dp$.

4.3 Application to FHS Offspring Cohort

In this section, we apply the proposed dynamic prediction approach to the motivating Framingham Offspring Cohort. The primary survival outcome of FHS is the recurrent general cardiovascular disease event. Scheduled every 4 year, investigators of FHS Offspring Cohort conducted exam on study participants, during which multiple longitudinal clinical markers were measured. The whole follow-up period is from year 1972 to year 2015 with total 9 exams. Our primary goal is to assess the prediction performance of combined multiple longitudinal outcomes on recurrent CVD events. Among the available clinical markers, we select the markers that are longitudinal collected and measured at every exam. Specifically, we include 6 longitudinal clinical outcomes in the analysis: systolic blood pressure (SBP), body mass index (BMI), blood glucose (BG), high-density lipoprotein (HDL), total cholesterol (TC), triglycerides (TG). Evidence in previous related studies has shown significant correlation between these markers and CVD events [76–80]. We also include other relevant time-independent covariates, i.e., age (in year), gender (0 if female, 1 if male), diabetes (0 if no, 1 if yes) and smoking status (0 if no, 1 if yes). After removing incomplete observations with missing data, our analysis focuses on 4221 participants in FHS Offspring Cohort. In the study, the participants were assessed during each exam cycle, and had up to 9 longitudinal observations. As the exact date for conducting exam varies across participants, we refer the average mean visit time for each exam as the corresponding exam time. Among the participants, 734 subjects experienced

1 general CVD events, and 801 subjects experienced more than 1 events, while the rest of the participants are cardiovascular disease events free until the end of follow-up period. We specify the intensity function of our proposed recurrent event model integrating the multivariate longitudinal information in Model 1. In order to ensure at least 99% of total variation is explained, we select the first 10 MFPC scores as the extracted features of those 6 longitudinal markers. To compare the prediction performance of our proposed model with other candidate models, we also specify another simple recurrent event model with only subject's time-independent covariates and baseline measurements of 6 longitudinal outcomes (referred as Model 2). Moreover, to test whether incorporating multiple longitudinal information can improve the prediction ability on recurrent event data, we apply the joint model of single longitudinal outcome (SBP, one of the most relevant clinical marker with CVD events) and recurrent event data (referred as Model 3) on FHS Offspring dataset and compare it with our proposed approach. While controlling for common baseline variables, we specify the intensity function for Model 1 as:

$$r_i(t) = r_0(t)\exp\{\beta_1\mathrm{Age}_i + \beta_2\mathrm{Male}_i + \beta_3\mathrm{Smoke}_i + \sum_{k=1}^{10}\alpha_k\mathrm{MFPCscores}_{ik} + v_i\}, \tag{4.9}$$

and Model 2 as:

$$\begin{aligned} r_i(t) = r_0(t)\exp\{&\beta_1\mathrm{Age}_i + \beta_2\mathrm{Male}_i + \beta_3\mathrm{Smoke}_i + \beta_4\mathrm{SBP}_i + \beta_5\mathrm{BMI}_i \\ &+ \beta_6\mathrm{BG}_i + \beta_7\mathrm{HDL}_i + \beta_8\mathrm{TC}_i + \beta_9\mathrm{TG}_i + v_i\}. \end{aligned} \tag{4.10}$$

The Model 3 specification is as following:

$$\begin{aligned} SBP_i(t) &= \gamma_0 + \gamma_1 \mathrm{Age}_i + \gamma_2 t + u_i + e_i(t) = f_i(t) + e_i(t) \\ r_i(t) &= r_0(t)\exp\{\beta_1 \mathrm{Age}_i + \beta_2 \mathrm{Male}_i + \beta_3 \mathrm{Smoke}_i + \nu f_i(t) + v_i\}. \end{aligned} \tag{4.11}$$

Here, $r_0(t)$ represents the baseline intensity function, $\boldsymbol{\beta}$ denotes the coefficients associated with risk factors, and $\boldsymbol{\alpha}$ denotes the regression coefficients for MFPC scores. In the joint model, u_i denotes the random effect in longitudinal process, and ν is the association parameter. Subjects with less or equal than two available longitudinal measurements for any of the markers are excluded while estimating MFPC scores. The prediction performance of the previous specified models are assessed by calculating the time-dependent Area Under Curve (AUC) at different prediction starting point over the follow-up period. The 10-fold cross validation is conducted to avoid overestimation of the prediction. The values of time-dependent AUC calculated based on the three candidate model are presented in Table 4.1. From the summarized results, we find both three models have acceptable AUCs indicating a good prediction performance. Among the three models, our proposed Model 1 has better prediction performance in regards of highest AUC for each prediction time interval. The AUCs estimated from Model 1 have about 0.05 increase on average when compared to Model 2 while the differences become larger in later prediction time such as Exam 6 and Exam 7. This pattern indicates incorporating longitudinal information in prediction of recurrent CVD events will enhance the accuracy of predicting risk of events, especially when prediction start at later follow-up time. When comparing

		Model 1	Model 2	Model 3
T(year)	T'(year)	AUC	AUC	AUC
13 (Exam 3)	16	0.756	0.717	0.733
	20	0.801	0.765	0.781
	24	0.793	0.754	0.762
	26	0.791	0.759	0.768
	34	0.744	0.733	0.741
	38	0.769	0.768	0.761
16 (Exam 4)	20	0.842	0.799	0.807
	24	0.805	0.745	0.762
	26	0.791	0.745	0.767
	34	0.741	0.734	0.739
	38	0.763	0.727	0.740
20 (Exam 5)	24	0.756	0.731	0.742
	26	0.762	0.752	0.759
	34	0.733	0.728	0.730
	38	0.754	0.733	0.744
24 (Exam 6)	26	0.796	0.720	0.742
	34	0.715	0.661	0.678
	38	0.736	0.701	0.713
34 (Exam 7)	34	0.720	0.695	0.705
	38	0.737	0.725	0.729

Table 4.1: Time-dependent AUC for Model 1 and candidate Model 2 & 3. Model 1 is the proposed model based on MFPCA method, Model 2 is the baseline model, and Model 3 is the joint model with single longitudinal marker.

Model 1 and Model 3, we notice an average of 0.03 increase in AUCs for our proposed Model 1, suggesting the gain of prediction accuracy when include multivariate longitudinal information instead of considering single longitudinal outcome. As a result, we select Model 1 as our final model and conduct dynamic prediction in the validation dataset. In order to illustrate the personalized dynamic prediction, we select two subjects who have different patterns of longitudinal profiles and experienced different occurrences of cardiovascular disease events. Conditioning on the available observations of 6 clinical markers, we predict their longitudinal trajectory from different prediction starting time. Figure 4.1 and Figure 4.2 shows how the predictions of longitudinal outcomes are updated over time for Subject 103 and Subject 122 respectively. From the left panel to the right panel of these two figures, the predicted trajectories are closer to the true observed values, also, the 95& confidence interval becomes more narrower, suggesting with more follow-up data the prediction improves. In Figure 4.3, we present the predicted probability of developing a new event in given time interval. Subject 103 did not experiance any CVD event during the follow-up period, while subject 122 (lower panels) had 2 CVD events between Exam 3 and Exam 7. At the beginning of prediction, although both Subject 103 and subject 122 did not experience any CVD event, the estimated event-free probability of Subject 122 is lower than Subject 103, which consist with his/her worse longitudinal profiles (i.e. significantly higher SBP, BMI, BG). The results suggest that Subject 122 has a higher risk of developing re-occurrence CVD events and should be closely monitored.

4.4 Simulation Study

In this section, we evaluate the performance of our proposed approach via simulation, in which the simulated data mimic the application dataset. The primary goal of the simulation study is to evaluate predictive performance of our proposed model under different scenarios, and compare it with multivariate joint modelling approach. Focusing on q longitudinal outcomes, $q = 1, 2, 3$, we use the linear mixed-effects models to generate simulated longitudinal data, which is:

$$Y_{iq}(t) = X_{iq}(t) + \epsilon_{iq}. \tag{4.12}$$

Here, $X_{iq}(t)$ denotes the latent trajectory of the q-th longitudinal outcome, and ϵ_{iq} is the error term which follows normal distribution. Since in practise, the longitudinal trajectories could follow different distributions, we consider three different scenarios of latent trajectory $X_{iq}(t)$ in the simulation:

1. Linear Model $X_{iq}(t_{ij}) = a_q + b_{q1}Z_i + b_{q2}t_{ij} + u_{iq}$

2. Exponential Model $X_{iq}(t_{ij}) = c_q + \exp(a_q t_{ij}) + u_{iq}$

3. Quadratic Model $X_{iq}(t_{ij}) = a_q(t_{ij} - b_q)^2 + c_q k + u_{iq}$

Let $r_i(t)$ denote the intensity of the recurrent process, we then assume the recurrent event intensity function of subject i as:

$$r_i(t) = r_0(t) \exp\{z_1\gamma_1 + \sum_{q=1}^{3} \alpha_q X_{iq}(t) + v_i\}. \tag{4.13}$$

We assume non-informative censoring for recurrent event process, the censoring time C_i is sampled from uniform distribution with 50% censoring rate. Let T_{ik} be the kth recurrent event times from study onset (time 0) for subject i, $k = 0, \ldots, n_i$, where n_i denotes the number of recurrent events. Here, α_q represents the association parameter between q-th longitudinal outcome and the recurrent event process. Random effect from longitudinal outcomes is denoted by u_{i1}, u_{i2} and u_{i3}, we assume them follow multivariate normal distribution with mean 0. The covariance matrix is denoted by Σ_u, with $\sigma_1^2, \sigma_2^2, \sigma_3^2$ denote the variance for random effects u_{i1}, u_{i2}, u_{i3} respectively, and $\rho_{12}, \rho_{23}, \rho_{13}$ represents the correlation between the three random effects. On the other hand, v_i is the random variable only associated with recurrent process and independent of u_{i1}, u_{i2} and u_{i3}, we assume that it follows a normal distribution with variance denoted by σ_v^2. For each of the 200 simulated datasets, the total sample size is 600, with randomly selected 400 subjects as the training dataset and the remaining 200 subjects for validation dataset. On average, there are 2 recurrent events per subject. In each scenario, we conduct the model inference in the training dataset using our proposed model based on MFPCA method, and the true joint model which is used to simulate the data. Moreover, since it is always

			MFPCA Model 1	True JM	Miss-specified JM
Scenario	T(year)	T'(year)	AUC	AUC	AUC
Linear Model	4	2	0.737	0.756	0.612
		4	0.752	0.761	0.614
	5	2	0.756	0.759	0.621
		4	0.753	0.763	0.634
	6	2	0.778	0.788	0.639
		4	0.782	0.792	0.641
Exponential Model	4	2	0.775	0.780	0.601
		4	0.781	0.785	0.605
	5	2	0.798	0.810	0.610
		4	0.802	0.813	0.612
	6	2	0.809	0.815	0.623
		4	0.812	0.820	0.629
Quadratic Model	4	2	0.721	0.741	0.593
		4	0.725	0.745	0.602
	5	2	0.742	0.754	0.603
		4	0.738	0.749	0.604
	6	2	0.761	0.770	0.611
		4	0.759	0.768	0.615

Table 4.2: Time-dependent AUC for 3 scenarios

challenging to specify appropriate distribution for longitudinal process in applications, assessing the performance of joint model with miss-specified latent trajectories is also worth investigation. After fitting the model, for each scenario, we conduct dynamic prediction for subjects in validation dataset using three methods: our proposed model, joint model with true latent trajectories, and joint model with miss-specified latent trajectories. The time-dependent AUCs are calculated for different prediction time interval, and serve as a criteria to evaluate prediction performance. Results for dynamic prediction in validation dataset are presented in Table 4.2. The AUC here is the average of all AUC calculated within given time interval for 200 simulation times. From the summarized results, although the our proposed model has an average of 0.01 lower AUC than the joint model under true trajectories, it outperforms the joint model when the longitudinal distributions are miss-specified in regards of an average of 0.15 higher AUC. The results suggest that the MFPCA approach is more robust compared to JM approach when the underlying latent trajectory for longitudinal process is unknown. Moreover, the computation cost for multivariate joint model of longitudinal markers and recurrent event is significantly higher than using our proposed MFPCA approach, which makes it hard to implement JM approach when application sample size is large. Overall, our proposed method has a robust predictive performance, also significantly improve computation efficiency.

4.5 Discussion

We integrate longitudinal information from varies of clinical markers in predicting the risk of next occurrence of recurrent event in this study. By employing multivariate functional principle component analysis method, we develop a novel two-step approach to model recurrent cardiovascular disease events with the extracted feature from historical observations of multiple relevant longitudinal clinical markers. Without assuming parametric distribution of longitudinal trajectories, the multivariate functional principle components analysis is a robust way to deal with joint variations between clinical markers and captures the feature of subjects. Among all available candidate markers, we select the optimal 6 markers that are frequently measured and has been demonstrated as risk factors of cardiovascular disease events in previous literature. We specify three prediction models, one is the recurrent event model with estimated MFPC scores capturing feature of multivariate longitudinal trajectories, one is the baseline recurrent event model only accounting for demographic characteristics and baseline measurements of selected markers, another is the joint model of single longitudinal marker (SBP) and recurrent CVD event. The application results suggest that including the historical observations of multivariate longitudinal markers improves prediction performance of recurrent CVD events when compared to only consider baseline covariates or single longitudinal outcome. In addition, we conduct a simulation study to evaluate the performance of our proposed approach under different circumstances. The results indicates that the MFPCA approach is more robust than the parametric multivariate joint model in regards of significantly higher AUCs when the underlying longitudinal trajectories are miss-specified. One limitation of the proposed two-stage approach is that in order to estimate parameters of MFPCA, we will need the subjects in training dataset to have more than two observations for each of the selected longitudinal markers to ensure estimation accuracy. When the sample size is small, excluding subjects without sufficient longitudinal observations may lead to non-universal useful outcome. Another limitation is that when modeling the recurrent event, we treat death as non-informative censoring, which may ignore the correlation between recurrent CVD events and death. We plan to extend the proposed model to take death process into account while predicting future occurrence of recurrent events in future research work. In summary, our study propose a novel approach to conduct personalized prediction of risk of recurrent cardiovascular disease events with multivariate longitudinal clinical markers, which helps physicians to closely monitor high risk participants and give personalized health care.

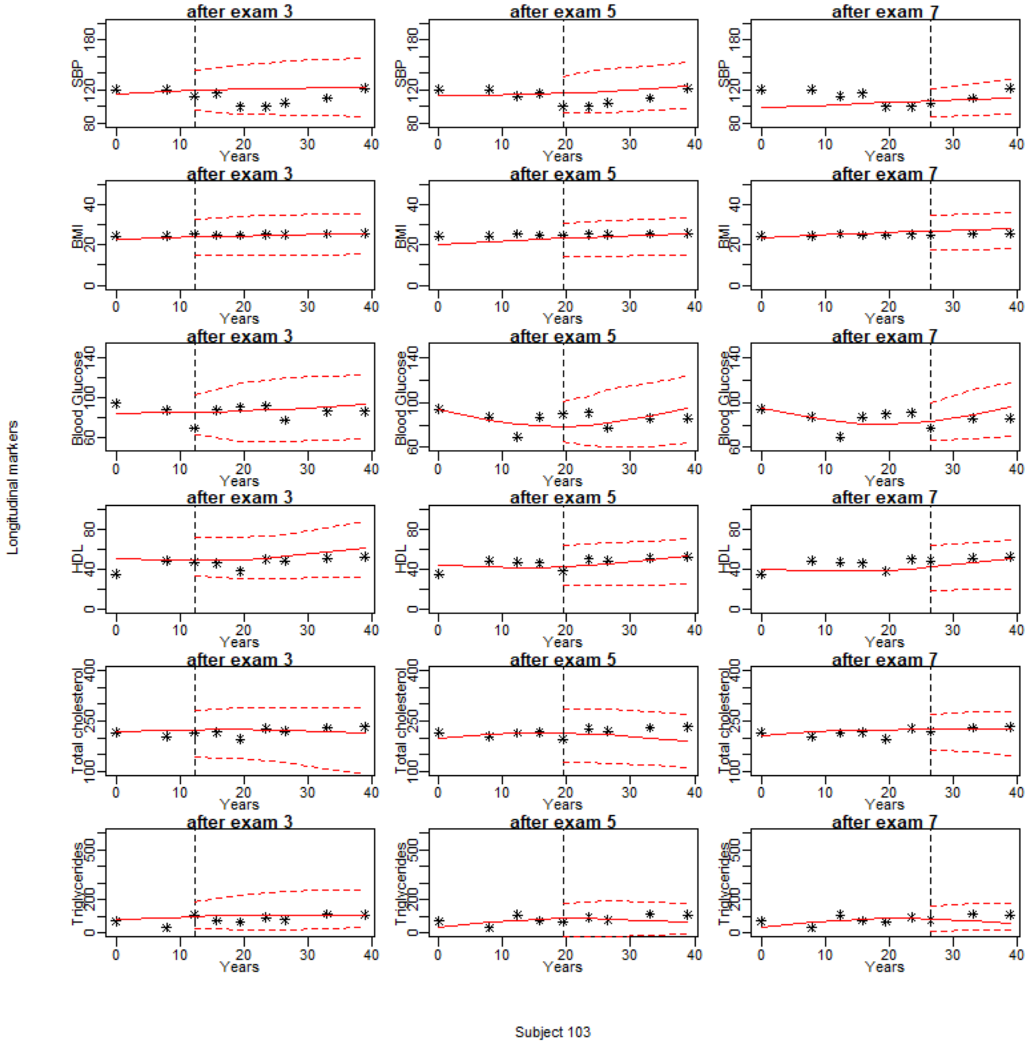


Figure 4.1: Predicted longitudinal trajectories for Subject 103. Dashed lines are the 2.5% and 97.5% percentiles. The dotted vertical line represents the time of prediction T.

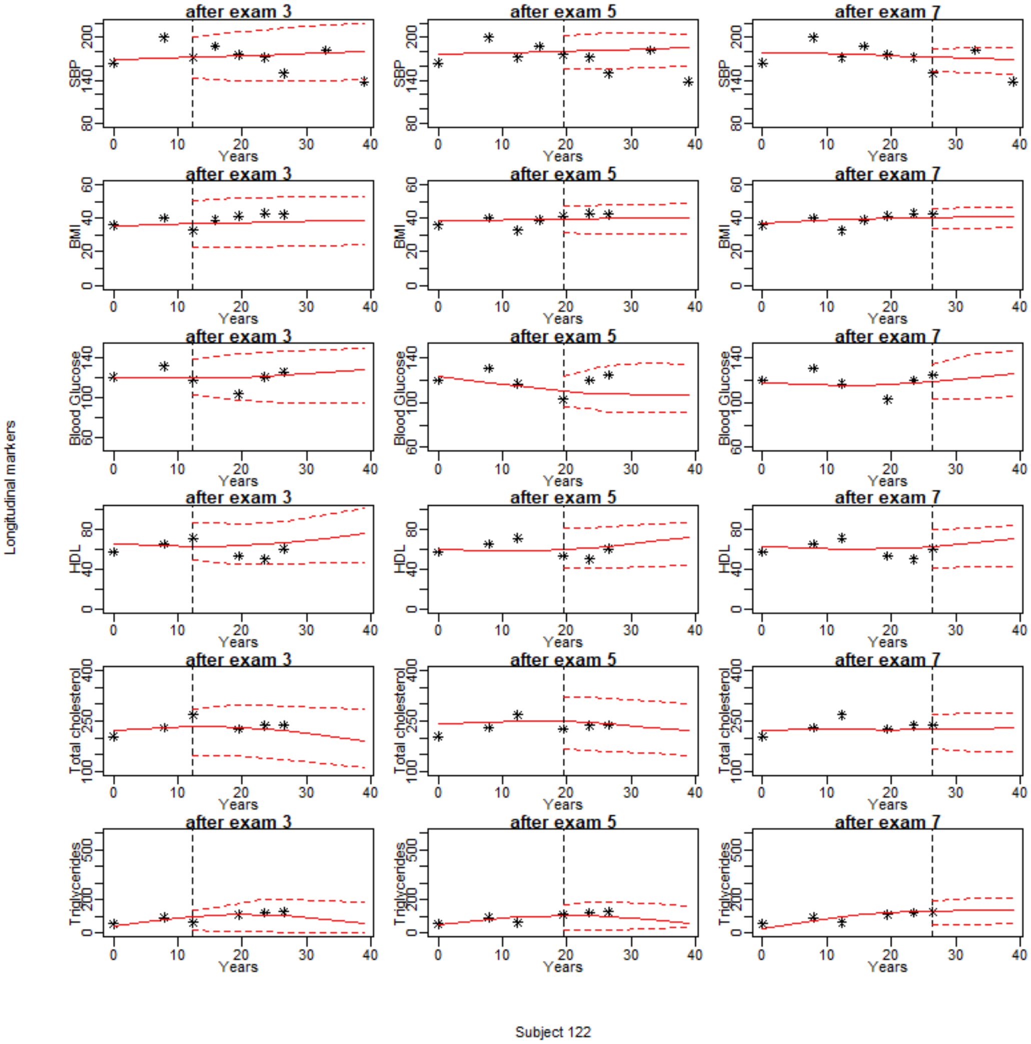


Figure 4.2: Predicted longitudinal trajectories for Subject 122. Dashed lines are the 2.5% and 97.5% percentiles. The dotted vertical line represents the time of prediction T.

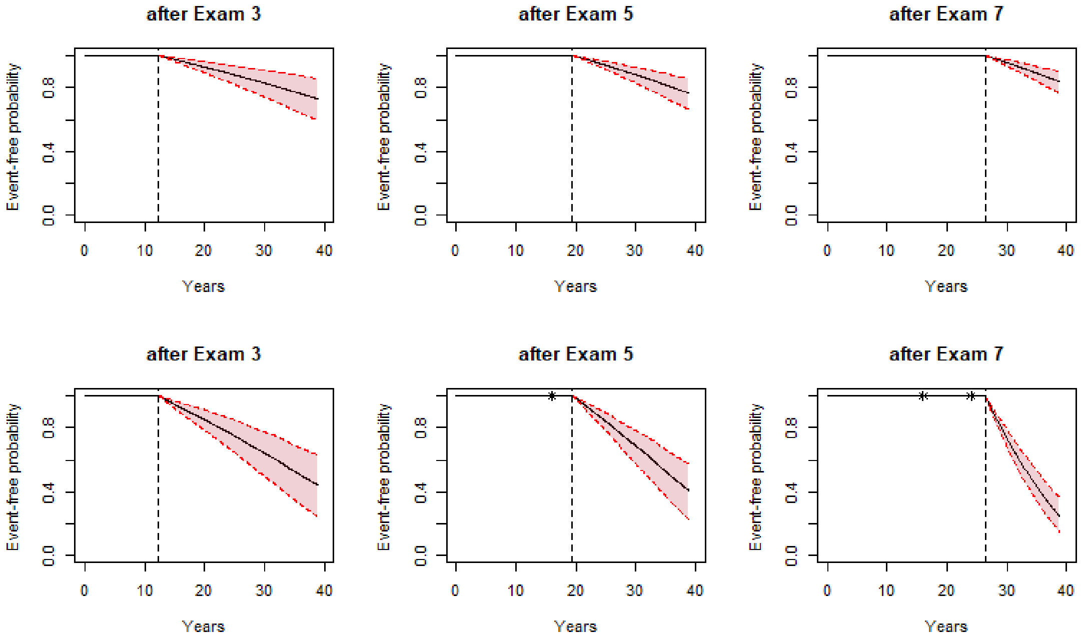


Figure 4.3: Predicted CVD event free probability for Subject 103 (upper panels) and Subject 122 (lower panels) in ALLHAT study. Dashed lines are the 2.5% and 97.5% percentiles.

Bibliography

1. Thayer, J. F., Yamamoto, S. S. & Brosschot, J. F. The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors. *International Journal of Cardiology* **141,** 122–131 (2010).

2. Mozaffarian, D., Benjamin, E., Go, A., Arnett, D., Blaha, M., Cushman, M., Das, S., De Ferranti, S., Després, J., Fullerton, H. & Howard, V. Executive summary: heart disease and stroke statistics - 2016 update. *Circulation* **133,** 447–454 (2016).

3. Trial, P. H. A. Major outcomes in high-Risk hypertensive patients. *Journal of the American Medical Association* **288,** 2981–2997 (2002).

4. Xing, C., Dupuis, J. & Cupples, L. A. Performance of statistical methods on CHARGE targeted sequencing data. *BMC Genetics* **15,** 104 (2014).

5. Collaboration, P. S. *et al.* Cholesterol, diastolic blood pressure, and stroke: 13 000 strokes in 450 000 people in 45 prospective cohorts. *The Lancet* **346,** 1647–1653 (1995).

6. MacMahon, S., Peto, R., Collins, R., Godwin, J., Cutler, J., Sorlie, P., Abbott, R., Neaton, J., Dyer, A. & Stamler, J. Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet* **335,** 765–774 (1990).

7. Antonini, A., Barone, P., Marconi, R., Morgante, L., Zappulla, S., Pontieri, F. E., Ramat, S., Ceravolo, M. G., Meco, G., Cicarelli, G., *et al.* The progression of non-motor symptoms in Parkinson's disease and their contribution to motor disability and quality of life. *Journal of neurology* **259,** 2621–2631 (2012).

8. Barone, P., Antonini, A., Colosimo, C., Marconi, R., Morgante, L., Avarello, T. P., Bottacchi, E., Cannas, A., Ceravolo, G., Ceravolo, R., *et al.* The PRIAMO study: a multicenter assessment of nonmotor symptoms and their impact on quality of life in Parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society* **24,** 1641–1649 (2009).

9. Cooper, J. A., Sagar, H. J., Jordan, N., Harvey, N. S. & Sullivan, E. V. Cognitive impairment in early, untreated Parkinson's disease and its relationship to motor disability. *Brain* **114,** 2095–2122 (1991).

10. Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C. M., Trojanowski, J. Q., *et al.* The Parkinson's progression markers initiative (PPMI)–establishing a PD biomarker cohort. *Annals of clinical and translational neurology* **5,** 1460–1477 (2018).

11. Shoulson, I. & Group, P. S. DATATOP: a decade of neuroprotective inquiry. *Annals of neurology* **44,** S160–S166 (1998).

12. Luo, S. & Wang, J. Bayesian hierarchical model for multiple repeated measures and survival data: an application to Parkinson's disease. *Statistics in medicine* **33,** 4279–4291 (2014).

13. Hoehn, M. M., Yahr, M. D., *et al.* Parkinsonism: onset, progression, and mortality. *Neurology* **50,** 318–318 (1998).

14. Schrag, A., Spottke, A., Quinn, N. P. & Dodel, R. Comparative responsiveness of Parkinson's disease scales to change over time. *Movement Disorders* **24,** 813–818 (2009).

15. Müller, J., Wenning, G., Jellinger, K., McKee, A., Poewe, W. & Litvan, I. Progression of Hoehn and Yahr stages in Parkinsonian disorders: a clinicopathologic study. *Neurology* **55,** 888–891 (2000).

16. Sato, K., Hatano, T., Yamashiro, K., Kagohashi, M., Nishioka, K., Izawa, N., Mochizuki, H., Hattori, N., Mori, H. & Mizuno, Y. Prognosis of Parkinson's disease: time to stage III, IV, V, and to motor fluctuations. *Movement disorders: official journal of the Movement Disorder Society* **21,** 1384–1395 (2006).

17. Zhao, Y. J., Wee, H. L., Chan, Y.-H., Seah, S. H., Au, W. L., Lau, P. N., Pica, E. C., Li, S. C., Luo, N. & Tan, L. C. Progression of Parkinson's disease as evaluated by Hoehn and Yahr stage transition times. *Movement Disorders* **25,** 710–716 (2010).

18. Djaldetti, R., Rigbi, A., Greenbaum, L., Reiner, J. & Lorberboym, M. Can early dopamine transporter imaging serve as a predictor of Parkinson's disease progression and late motor complications? *Journal of the neurological sciences* **390,** 255–260 (2018).

19. He, B. & Luo, S. Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease. *Statistical methods in medical research* **25,** 1346–1358 (2016).

20. Iddi, S., Li, D., Aisen, P. S., Rafii, M. S., Litvan, I., Thompson, W. K. & Donohue, M. C. Estimating the Evolution of Disease in the Parkinson's Progression Markers Initiative. *Neurodegenerative Diseases* **18,** 173–190 (2018).

21. Amorim, L. D. & Cai, J. Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology* **44,** 324–333 (2015).

22. Vaida, F., Xu, R., *et al.* Proportional hazards model with random effects. *Statistics in Medicine* **19,** 3309–3324 (2000).

23. Pepe, M. S. & Cai, J. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88,** 811–820 (1993).

24. Lin, D., Wei, L., Yang, I. & Ying, Z. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62,** 711–730 (2000).

25. Clayton, D. Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* **3,** 244–262 (1994).

26. Faucett, C. L. & Thomas, D. C. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* **15,** 1663–1685 (1996).

27. Wulfsohn, M. S. & Tsiatis, A. A. A joint model for survival and longitudinal data measured with error. *Biometrics,* 330–339 (1997).

28. Henderson, R., Diggle, P. & Dobson, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1,** 465–480 (2000).

29. Han, J., Slate, E. H. & Peña, E. A. Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in Medicine* **26,** 5285–5302 (2007).

30. Crowther, M. J., Abrams, K. R. & Lambert, P. C. Joint modeling of longitudinal and survival data. *The Stata Journal* **13,** 165–184 (2013).

31. Brown, E. R. & Ibrahim, J. G. Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59,** 686–693 (2003).

32. Liu, L. & Huang, X. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58,** 65–81 (2009).

33. Rizopoulos, D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-Event data. *Biometrics* **67,** 819–829 (2011).

34. Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F. & Jacqmin-Gadda, H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71,** 102–113 (2015).

35. Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T. & Sandler, H. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* **69,** 206–213 (2013).

36. Król, A., Ferrer, L., Pignon, J.-P., Proust-Lima, C., Ducreux, M., Bouché, O., Michiels, S. & Rondeau, V. Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000–05 trial. *Biometrics* **72,** 907–916 (2016).

37. Musoro, J., Struijk, G., Geskus, R., ten Berge, I. & Zwinderman, A. Dynamic prediction of recurrent events data by landmarking with application to a follow-up study of patients after kidney transplant. *Statistical Methods in Medical Research* **27,** 832–845 (2018).

38. Yao, F., Müller, H.-G. & Wang, J.-L. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100,** 577–590 (2005).

39. Yao, F. Functional principal component analysis for longitudinal and survival data. *Statistica Sinica,* 965–983 (2007).

40. Holte, S., Randolph, T., Ding, J., Tien, J., McClelland, R., Baeten, J. & Overbaugh, J. Efficient use of longitudinal CD4 counts and viral load measures in survival analysis. *Statistics in medicine* **31,** 2086–2097 (2012).

41. Yan, F., Lin, X., Huang, X., *et al.* Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *The Annals of Applied Statistics* **11,** 1649–1670 (2017).

42. Ornish, D., Brown, S. E., Billings, J., Scherwitz, L., Armstrong, W. T., Ports, T. A., McLanahan, S. M., Kirkeeide, R. L., Gould, K. & Brand, R. Can lifestyle changes reverse coronary heart disease?: The Lifestyle Heart Trial. *The Lancet* **336,** 129–133 (1990).

43. Kannel, W. B. Blood pressure as a cardiovascular risk factor: prevention and treatment. *Journal of the American Medical Association* **275,** 1571–1576 (1996).

44. Kannel, W. B. & McGee, D. L. Diabetes and cardiovascular disease: the Framingham study. *Journal of the American Medical Association* **241,** 2035–2038 (1979).

45. Rizopoulos, D., Hatfield, L. A., Carlin, B. P. & Takkenberg, J. J. Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* **109,** 1385–1397 (2014).

46. Damen, J. A., Hooft, L., Schuit, E., Debray, T. m. P., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C., Chiocchia, V., Roberts, C., *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *British Medical Journal* **353,** i2416 (2016).

47. Iadecola, C. & Gorelick, P. B. Hypertension, angiotensin, and stroke: beyond blood pressure. *Stroke* **35,** 348–350 (2004).

48. Staessen, J. A., Thijs, L., Fagard, R., O'brien, E. T., Clement, D., de Leeuw, P. W., Mancia, G., Nachev, C., Palatini, P., Parati, G., *et al.* Predicting cardiovascular risk using conventional vs ambulatory blood pressure in older patients with systolic hypertension. *Journal of the American Medical Association* **282,** 539–546 (1999).

49. Giorda, C. B., Avogaro, A., Maggini, M., Lombardo, F., Mannucci, E., Turco, S., Alegiani, S. S., Raschetti, R., Velussi, M. & Ferrannini, E. Recurrence of cardiovascular events in patients with type 2 diabetes: epidemiology and risk factors. *Diabetes Care* **31,** 2154–2159 (2008).

50. Gilks, W. R., Best, N. & Tan, K. Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics,* 455–472 (1995).

51. Wang, Z., Broccardo, M. & Song, J. Hamiltonian Monte Carlo methods for Subset Simulation in reliability analysis. *Structural Safety* **76,** 51–67 (2019).

52. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science,* 457–472 (1992).

53. Welling, M. & Teh, Y. W. *Bayesian learning via stochastic gradient Langevin dynamics* in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (2011), 681–688.

54. Quiroz, M., Kohn, R., Villani, M. & Tran, M.-N. Speeding up MCMC by efficient data subsampling. *arXiv preprint arXiv:1404.4178* (2014).

55. Wang, X., Guo, F., Heller, K. A. & Dunson, D. B. *Parallelizing MCMC with random partition trees* in *Advances in Neural Information Processing Systems* (2015), 451–459.

56. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56,** 337–344 (2000).

57. Heagerty, P. J. & Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics* **61,** 92–105 (2005).

58. Proust-Lima, C., Séne, M., Taylor, J. M. & Jacqmin-Gadda, H. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* **23,** 74–90 (2014).

59. Finegold, J. A., Asaria, P. & Francis, D. P. Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. *International Journal of Cardiology* **168,** 934–945 (2013).

60. Kaplan, N. M. & Vidt, D. G. Major cardiovascular events in hypertensive patients randomized to doxazosin versus chlorthalidone. *Current Hypertension Reports* **2,** 431–431 (2000).

61. Happ, C. & Greven, S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association,* 1–11 (2018).

62. Ravina, B., Tanner, C., DiEuliis, D., Eberly, S., Flagg, E., Galpern, W. R., Fahn, S., Goetz, C. G., Grate, S., Kurlan, R., *et al.* A longitudinal program for biomarker development in Parkinson's disease: a feasibility study. *Movement disorders: official journal of the Movement Disorder Society* **24,** 2081–2090 (2009).

63. Latourelle, J. C., Beste, M. T., Hadzi, T. C., Miller, R. E., Oppenheim, J. N., Valko, M. P., Wuest, D. M., Church, B. W., Khalil, I. G., Hayete, B., *et al.* Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *The Lancet Neurology* **16,** 908–916 (2017).

64. Hall, S., Surova, Y., Öhrfelt, A., Zetterberg, H., Lindqvist, D. & Hansson, O. CSF biomarkers and clinical progression of Parkinson disease. *Neurology* **84,** 57–63 (2015).

65. Mollenhauer, B., Caspell-Garcia, C. J., Coffey, C. S., Taylor, P., Shaw, L. M., Trojanowski, J. Q., Singleton, A., Frasier, M., Marek, K., Galasko, D., *et al.* Longitudinal CSF biomarkers in patients with early Parkinson disease and healthy controls. *Neurology* **89,** 1959–1969 (2017).

66. Kang, J.-H., Caspell, C., Coffey, C., Taylor, P., Frasier, M., Marek, K., Trojanowski, J. Q. & Shaw, L. M. Association Between CSF Biomarkers and Clinical Phenotype of Early Parkinson's Disease in the Parkinson's Progression Markers Initiative (PPMI). *ratio* **1,** 42 (2012).

67. James, G. M., Hastie, T. J. & Sugar, C. A. Principal component models for sparse functional data. *Biometrika,* 587–602 (2000).

68. Fukunaga, K. & Koontz, W. L. Application of the Karhunen-Loeve expansion to feature selection and ordering. *IEEE Transactions on Computers* **100,** 311–318 (1970).

69. Happ, C. & Greven, S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* (2016).

70. Rizopoulos, D. *Joint models for longitudinal and time-to-event data: With applications in R* (Chapman and Hall/CRC, 2012).

71. Uno, H., Cai, T., Tian, L. & Wei, L. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537 (2007).

72. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48,** 1029–1040 (2006).

73. Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine* **32,** 5381–5397 (2013).

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

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

76. Hokanson, J. E. & Austin, M. A. Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies. *Journal of cardiovascular risk* **3,** 213–219 (1996).

77. Gordon, D. J., Probstfield, J. L., Garrison, R. J., Neaton, J. D., Castelli, W. P., Knoke, J. D., Jacobs Jr, D. R., Bangdiwala, S. & Tyroler, H. A. High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies. *Circulation* **79,** 8–15 (1989).

78. Lemieux, I., Lamarche, B., Couillard, C., Pascot, A., Cantin, B., Bergeron, J., Dagenais, G. R. & Després, J.-P. Total cholesterol/HDL cholesterol ratio vs LDL cholesterol/HDL cholesterol ratio as indices of ischemic heart disease risk in men: the Quebec Cardiovascular Study. *Archives of internal medicine* **161,** 2685–2692 (2001).

79. Bonora, E. & Muggeo, M. Postprandial blood glucose as a risk factor for cardiovascular disease in type II diabetes: the epidemiological evidence. *Diabetologia* **44,** 2107–2114 (2001).

80. Hao, G., Wang, X., Treiber, F. A., Harshfield, G., Kapuku, G. & Su, S. Blood pressure trajectories from childhood to young adulthood associated with cardiovascular risk: results from the 23-year longitudinal Georgia stress and heart study. *Hypertension,* 116 (2017).