Distribution Agreement

Signature:

____________________________	______________
Derun Xia	Date

Approval Sheet

Dynamic prediction of survival status in patients undergoing cardiac catheterization using a joint modeling approach

By

Derun Xia

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

Yi-An Ko, PhD

(Thesis Advisor)

José Binongo, PhD

(Reader)

Abstract Cover Page

Dynamic prediction of survival status in patients undergoing cardiac catheterization using a joint modeling approach

By

Derun Xia

B.S., Nanjing Medical University, 2016

Thesis Committee Chair: Yi-An Ko, PhD

An abstract of

A thesis submiited to the Faculty of the

Rollins School of Public Heath of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Heath

In Biostatistics

2023

Abstract

Dynamic prediction of survival status in patients undergoing cardiac catheterization using a joint modeling approach

By Derun Xia

Background: Traditional cardiovascular disease risk factors have a limited ability to precisely predict patient survival outcomes. To better stratify the risk of patients with established coronary artery disease (CAD), it is useful to develop dynamic prediction tools that can update the prediction by incorporating time-varying data to enhance disease management.

Objective: To dynamically predict myocardial infarction (MI) or cardiovascular death (CV-death) and all-cause death among patients undergoing cardiac catheterization using their electronic health records (EHR) data over time and evaluate the prediction accuracy of the model.

Methods: Data from 6119 participants were obtained from Emory Cardiovascular Biobank (EmCAB). We constructed the joint model with multiple longitudinal variables to dynamically predict MI/CV-death and all-cause death. The cumulative effect and slope of longitudinally measured variables were considered in the model. The time-dependent area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the discriminating capability, and the time-dependent Brier score was used to assess prediction error.

Results: In addition to existing risk factors including disease history, changes in several clinical variables that are routinely collected in the EHR showed significant contributions to adverse events. For example, the decrease in glomerular filtration rate (GFR), body mass index (BMI), high-density lipoprotein (HDL), systolic blood pressure (SBP) and increase in troponin-I increased the hazard of MI/CV-death and all-cause death. More rapid decrease in GFR and BMI (corresponding to decrease in slope) increased the hazard of MI/CV-death and all-cause death. More rapid increase in diastolic blood pressure (DBP) and more rapid decrease in SBP increased the hazard of all-cause death. The time-dependent AUCs of the traditional Cox proportional model were higher than those of the joint model for MI/CV-death and all-cause death. The Brier scores of the joint model were also higher than those of the Cox proportional model.

Conclusion: Joint modeling that incorporates longitudinally measured variables to achieve dynamic risk prediction is better than conventional risk assessment models and can be clinically useful. The joint model did not appear to perform better than a Cox regression model in our study. Possible reasons include data availability, selection bias, and quality uncertainty in the EHR. Future studies should address these issues when developing dynamic prediction models.

Keywords: Dynamic prediction, longitudinal variable, cardiovascular disease, joint model, risk prediction

Dynamic prediction of survival status in patients undergoing cardiac catheterization using a joint modeling approach

By

Derun Xia

B.S., Nanjing Medical University, 2016

Thesis Committee Chair: Yi-An Ko, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2022

Table of contents

1 INTRODUCTION

Globally, cardiovascular disease (CVD) is the major cause of death, accounting for 31% of deaths. CVD incidence and prevalence continue to climb in the United States, despite a drop in death rates[1]. By 2030, it is expected that 44% of American adults will suffer from at least one form of CVD[2]. CVD is associated with significant economic and health costs [3]. The development of precise risk assessment tools and cost-effective preventative and treatment strategies is an unmet need. Traditional cardiovascular disease risk factors only account for the likelihood of developing coronary artery disease (CAD), but they are less effective at predicting patient survival outcomes [4, 5]. To better stratify the risk of patients with established CAD, it is essential to develop dynamic diagnostic tools in order to enhance disease management.

Traditional models use baseline information but typically ignore longitudinal changes in risk markers, missing potential impact on risk assessment [6, 7]. Ideally, forecasting may be more precise if change in marker values over time is also considered. Dynamic prediction utilizes time-dependent marker data obtained during a patient's follow-up to provide updated, more precise survival probability predictions. Electronic health record (EHR) data provides a rich source of clinical information on a large and diverse population, making it cost-effective and ideal for studying rare diseases and subpopulations. Meanwhile, EHR data also emphasizes the time-dependent characteristics of health events, as it records patient data longitudinally over time. This longitudinal data can provide valuable insights into disease progression, treatment response, and long-term outcomes, and can be used to identify patterns and trends in health outcomes over time. Therefore, EHR data is essential for healthcare professionals and researchers seeking to make accurate and informed dynamic predictions about future health outcomes.

Recently, many new methods have utilized longitudinal variables to dynamically predict the time-to-event, such as landmarking, joint model, functional principal component analysis (FPCA), and random forest[8-12]. However, these methods have their own disadvantages. By restricting the analysis to a subset of the data, landmarking can result in a loss of information and decreased statistical power; [13] Many machine learning algorithms are black-box models, making it difficult to understand the underlying relationships between the predictors and the outcome[14].

Joint models are suitable for dynamically predicting outcomes using EHR data[15, 16]. EHR data is often collected over time, and joint models can handle both time-varying and time-invariant variables, making them well suited for modeling these data. Moreover, longitudinal joint models can handle missing data frequently occur in the EHR system in a principled way, making it possible to use all available information. The results of the joint model are also straightforward and can be represented graphically to illustrate the strength of the link between survival outcomes and longitudinal variables, such as the hazard ratio[17]. The predictive results of the combined model have the potential to help physicians make precise and timely medical decisions.

Our aim is to develop a joint model to dynamically predict the adverse events including MI/CV-death and all-cause death among patients undergoing cardiac catheterization using their EHR data over time and to evaluate the prediction accuracy of the model. We illustrate a method to develop individualized dynamic prediction models based on the the progression of longitudianl variables. In addition, we will compare these results with a traditional Cox regression model that uses only baseline covariates.

2 METHODS

2.1 Study design and participants

Data used in this analysis were obtained from Emory Cardiovascular Biobank (EmCAB), an ongoing prospective registry of patients undergoing cardiac catheterization, which was established to identify novel factors associated with the pathobiological process and treatment of cardiovascular disease. Detailed information on EmCAB study protocols, including participant inclusion and exclusion criteria have been described [18].

In our study, 6119 participants who underwent cardiac catheterization, enrolled 2004—2021, were included. At enrollment, patients are interviewed to collect information on demographic characteristics, medical history, detailed family history, medication usage, and health behaviors (alcohol/drug use) prior to cardiac catheterization. Each patients had a 1- and 5-year follow-up phone calls for any adverse events, including myocardial infarction (MI) and cardiovascular (CV).

We selected longitudinal variables in the EHR data for which at least 90% of the patients had more than one observation, including blood pressure measurements, BMI, and labs (see below). Outliers and extreme values were reviewed and removed if necessary. In case of multiple measurements of the same variable within a day, we reduced the number of observations by using the median value for analysis.

The study was approved by the institutional review board (IRB) at Emory University (Atlanta, Georgia, USA) and is renewed annually. All participants provided written informed consent at the time of enrolment.

2.2 Statistical Analysis

Baseline characteristics of the study participants were summarized using mean ± standard deviation (SD) or median (interquartile range [IQR]) for continuous variables and frequencies and percentages for categorical variables.

We developed a joint longitudinal-survival modeling framework to focus on dynamic prediction of the future risk of MI or CV death and all-cause death. The joint model takes into account multiple longitudinal measures and their slopes, and the prediction of future risk can be updated based on multiple longitudinal measures as well as other baseline characteristics. The survival time was calculated from the enrollment to the time of MI/CV-death, all-cause death or censoring.

The joint model consists of two sub-models. The survival sub-model takes the form of a Cox proportional hazards model with baseline covariates including age, gender, race, education, and history of hypertension, smoking, diabetes, hypercholesterolemia, revascularization and heart failure. The longitudinal sub-model describes the evolution of the repeated measures over time with the main effects from observation time (in years), age, gender, and race. The longitudinal variables included estimated glomerular filtration rate (eGFR), body mass index (BMI), high-density lipoprotein (HDL), low-density lipoprotein (LDL), cardiac troponin-I, diastolic blood pressure (DBP), systolic blood pressure (SBP), and hemoglobin A1c (HbA1c). Random effects were used to capture the between-subject variation. For all longitudinal variables, the slope coefficients of observation time and intercepts vary randomly across individuals. We expanded the time effect in the longitudinal sub-model using a spline basis matrix to capture possibly the nonlinear subject-specific trajectories.

Let $\mathrm{y_{ij}}(\mathrm{t})$ denote observation of the j-th measurement (j = 1, …, n_i, where n_i is the number of observations for subject i) for the i-th subject (i=1,…, N) at time t. The following linear mixed model can be used model a longitudinally measured variable:

$$\mathrm{y_{ij}}(t) = \mathrm{m_{ij}}(t) + \mathrm{\varepsilon_{ij}}(t) = x_{ij}^{\mathbf{T}}(t)\mathrm{\beta} + z_{ij}^{\mathbf{T}}(t)b_i + \mathrm{\varepsilon_{ij}}(t) \quad (1)$$

$x_{ij}^{\mathbf{T}}(t_{ij})\beta$ is the fixed-effect and $z_{ij}^{\mathbf{T}}(t_{ij})b_i$ is the random-effects. $\varepsilon_{ij}(t_{ij})$ donates measurement error.

Given the eGFR as an example:

$$\begin{aligned}\mathrm{eGF}R_i(t) = \mu + \theta_{0i} + (\mathrm{\beta_1} + \theta_{1i})B_n(t, \mathrm{\lambda_1}) + (\mathrm{\beta_2} + \theta_{2i})B_n(t, \mathrm{\lambda_2}) + (\mathrm{\beta_3} + \theta_{3i})B_n(t, \mathrm{\lambda_3}) \quad (2)\\ + \mathrm{\beta_4} * \mathrm{age_i} + \mathrm{\beta_4 Gender_i} + \mathrm{\beta_6 Black_i} + \mathrm{\varepsilon_i}(t)\end{aligned}$$

$$\theta_i \sim N(0, \mathrm{\tau}^2),$$

$$\varepsilon_i(t_{ij}) \sim N(0, \theta^2),$$

$$\theta_i \perp\perp \epsilon_{ij}$$

$$\begin{aligned}\mathrm{m_i}(t) = \mu + \theta_{0i} + (\beta_1 + \theta_{1i})B_n(t, \lambda_1) + (\beta_2 + \theta_{2i})B_n(t, \lambda_2) + (\beta_3 + \theta_{3i})B_n(t, \lambda_3) + \beta_4 \quad (3)\\ * age_i + \beta_4 Gender_i + \beta_6 Black_i\end{aligned}$$

The $\mu + \theta_{0i}$ is the patient-specific intercept μ_0 is the overall intercept and θ_{0i} is the subject-specific difference from μ. The matrix represents a spline basis matrix for a natural cubic spline of time that has two internal knots, resulting in three degrees of freedom. These knots were placed at the 33.3% and 66.7% percentiles of the follow-up time points. The $\beta_u + \theta_{ui}$ is the subject specific slope for the u-th basic function of a spline with knots λ_u

The hazard function is:

$$\mathrm{h_i(t) = h_0(t)exp(\gamma^T\omega_i} + \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{\alpha_k m_{ik}(t))} \qquad (4)$$

$\mathrm{h_0(t)}$ was the baseline hazard function. ω_i is the baseline covariate. We have *K* multiple longitudinal variables, the α_k linked the k-th (k=1,…, K) linear mixed model and Cox regression model and assuming the hazard at time t was dependent on the longitudinal trajectory, $\mathrm{m_{ik}(t)}$, through the estimated value at time t. When the α_k is significant, it indicated that there is an association between the k-th longitudinal variable and the longitudinal measures and time to event. And the exp (α_k) was the hazard ratio for one unit increase in the $\mathrm{m_{ik}(t)}$ at time t for k-th longitudinal variable. We also include the time-dependent slopes and the cumulative effects of longitudinal variables in the model.

The baseline hazard function is represented by $\mathrm{h_0(t)}$. ω_i represents the baseline covariate. The α_k connects the linear mixed model and Cox regression model and assumes that the hazard at time t is dependent on the longitudinal trajectory, represented by $\mathrm{m_{ik}(t)}$, through the estimated value at that time. If the α_k is significant, it indicates that there is a correlation between the longitudinal variable and time to event. The hazard ratio for a one unit increase in $\mathrm{m_i(t)}$ at time t can be calculated as exp(α). The model also considers the time-varying slopes and cumulative effects of the longitudinal variables. The cumulative effects $\frac{\int_0^t m_i(s)ds}{t}$ is the hazard of an event at t is associated with the area under the trajectory up to t.

Joint models for such joint distributions are of the following form. The θ_i is a vector of random effects that explains the interdependencies. $p(.)$ is the density function and $S(.)$ is the survival function.

$$p(y_{ij} \mid \theta_{ij}) = \prod_{k=1}^{n_{ij}} p(y_{ij,k} \mid \theta_{ij}) = \prod_{j} p(y_{ij} \mid \theta_{ij}) \quad (5)$$

$$p(y_i, T_i, \delta_i \mid \theta_i) = \prod_{j} p(y_{ij} \mid \theta_{ij}) p(T_i, \delta_i \mid \theta_i) \quad (6)$$

T_i is the observed event time for patient i and δ_i is the event indicator. The key assumption is that given the random effects, the repeated measurements in each outcome are independent, the longitudinal variables are independent of each other, and longitudinal outcomes are independent of the time-to-event outcome.

The Bayesian approach was adopted for model inference and for dynamic predictions. The key step in prediction for a new subject was to obtain samples of subject's random effects from the posterior distribution given the estimated parameters and previous longitudinal observations (at least one measure). The samples were then used to calculate the predictions for the longitudinal variables' future trajectories and risk of MI/CVdeath and all-cause death. Based on the general framework of joint models presented earlier, we are interested in deriving cumulative risk probabilities for a new subject j^* that has survived up to time point t and has provided longitudinal measurements $\mathcal{Y}_{kj^*}(t) = \{y_{kj^*}(t_{j^*l}); 0 \leq t_{j^*l} \leq t, l = 1, \ldots, n_{j^*}, k = 1, \ldots, K\}$, with K denoting the number of longitudinal outcomes. The probabilities of interest are:

$$\begin{aligned}\pi_{j^*}(u \mid t) &= \Pr\{T_{j^*}^* \leq u \mid T_{j^*}^* > t, \mathcal{Y}_{j^*}(t), \mathcal{D}_n\} \\ &= 1 - \frac{\iint S(u \mid b_{j^*}, \theta)}{S(t \mid b_{j^*}, \theta)} p\{b_{j^*} \mid T_{j^*}^* > t, \mathcal{Y}_{j^*}(t), \theta\} p(\theta \mid \mathcal{D}_n) d b_{j^*} d\theta\end{aligned} \tag{7}$$

where $S(\cdot)$ denotes the survival function conditional on the random effects, and $\mathcal{Y}_{j^*}(t) = \{\mathcal{Y}_{1j^*}(t), \ldots, \mathcal{Y}_{Kj^*}(t)\}$. Combining the three terms in the integrand we can device a Monte Carlo scheme to obtain estimates of these probabilities, namely.

Firstly, we can sample a value $\tilde{\theta}$ from the posterior of the parameters $[\theta \mid \mathcal{D}_n]$ and sample a value $\tilde{b}_{j^*}$ from the posterior of the random effects $[b_{j^*} \mid T_{j^*}^* > t, \mathcal{Y}_{j^*}(t), \tilde{\theta}]$. We then compute the ratio of survival probabilities $S(u \mid \tilde{b}_{j^*}, \tilde{\theta})/S(t \mid \tilde{b}_{j^*}, \tilde{\theta})$. After replicating these steps L times, we can estimate the conditional cumulative risk probabilities by:

$$1 - \frac{1}{L} \sum_{l=1}^{L} \frac{S\left(u \mid \tilde{b}_{j^*}^{(l)}, \tilde{\theta}^{(l)}\right)}{S\left(t \mid \tilde{b}_{j^*}^{(l)}, \tilde{\theta}^{(l)}\right)} \tag{8}$$

and their standard error by calculating the standard deviation across the Monte Carlo samples.

We calculated time-dependent areas under receiver-operating characteristics (ROC) curves (AUCs) and Brier score to assess the performance of the longitudinal marker at different time points over the follow-up period. We predicted the probabilities of MI and CV-death and all-cause of death occurring in the time frame (t, t+Δt], using all measures collected till time t. Then the AUCs were calculated to assess how well the longitudinal marker distinguished the status of patients at time t+Δt. The Brier score is a metric used to assess the precision of a predicted survival function at time t+Δt. It calculates the average squared difference between the observed survival status and the predicted survival probability, with a range of values from 0 to 1. Since the participants were reassessed approximately every year, we selected t at 2, 3 ,4 ,5 and 6 years, and Δt = 1, 2 (years). In general, higher AUCs indicate higher discrimination of the models and lower Brier score indicates worse precision of prediction. For comparison, we also fitted proportional hazards models (Cox model) with baseline measures. We then assessed the predictive performance of these models using time-dependent AUCs and Brier scores.

In addition, we applied the resulting joint models to predict the future longitudinal trajectories and risk of MI /CV-death for new participants. We selected 2 patients to demonstrated initialized dynamic prediction was updated over time as new clinical information became available. The joint model fitting and predictions were achieved using the R Jmbayes2 package[19].

3 RESULTS

Table 1 summarizes the characteristics of the 6119 participants. The median follow-up time was 7.53 years (SD 4.16; range 0.09 - 13.73). The average age at baseline was 62.9 years (SD 12.7; range 18.7 - 99.6), 64.9% were women and 19.6% were black. 3984 (65.1%) patients had a smoking history. Among 6119 participants the average eGFR at baseline was 72.5 mL/min/1.73m2 (SD 24.4; range 2.3 - 175.5). 76.9% of patients had a history of hypertension, 2143 (35.0%) patients had a history of diabetes mellitus, and 1827 (29.9%) patients had a history of hypercholesterolemia. Meanwhile, 1406 (23.0%) patients had a history of myocardial infarction, and 2246 (36.7%) have a history of heart failure. 2986 (48.8%) patients had a history of revascularization.

Figure 1 shows Kaplan-Meier survival curves. The patients having the smoking history, history of heart failure, myocardial infarction, or a history of revascularization had a lower probability of MI/CV-death-free survival than the reference group during 12 years of follow-up.

Table 2 shows estimated hazard ratios from the joint models. Based on the results of the MI/CV-death joint models, age, gender, race, education, history of hypertension, history of myocardial infarction and history of heart failure (measured at baseline), eGFR, BMI, HDL, Troponin-I, and SBP (measured longitudinally) were all significant predictors of the hazard of MI/CV-death. For all-cause death joint model, age, gender, race, education, history of hypertension, history of diabetes mellitus, history of myocardial infarction, history of heart failure, history of revascularization (measured at baseline) and eGFR, BMI, HDL, Troponin-I, and SBP (measured longitudinally) were also significant predictors all-cause death. Compared with the MI/CV-death joint model, the slope of DBP, SBP and HbA1c were new significant predictors of all-cause death.

Table 3 presents the AUCs and Brier scores of Cox regression and the joint models. It shows that the AUCs of the joint models were lower than that of the Cox models. The Brier scores of joint models were higher than Cox regression models, which indicates that the prediction error of joint models was higher than the Cox regression model.

Figure 2 shows the dynamic prediction of all-cause death and MI/CV-death for two patients. As new longitudinal measurements were incorporated into the model, the linear mixed regression models were subsequently updated, and the risk function was simultaneously updated according to cumulative effect (the area under the model divided by the follow-up time). Lastly, the updated survival curve from the prediction time interval presented the predicted survival (event-free) probability.

4 DISCUSSION

In this study, we built a joint model with multiple longitudinal variables to dynamically predict two survival outcomes, MI/CV-death and all-cause death, and found baseline and longitudinal variables from EHR that were significantly associated with survival outcomes. We also compared the discrimination power with the traditional Cox regression model based on time-dependent ROC of AUC. Results showed that the cumulative effect of variables such as eGFR, BMI, HDL and the slope over time was associated with survival outcomes. Based on AUC and Brier score, we did not find better discrimination power and prediction accuracy of joint model compared to Cox regression model.

Several approaches have been developed to accomplish dynamic prediction of conditional survival probabilities based on longitudinal and survival data, including joint models[20], landmark models[21], and random forests[10]. The landmark model takes into account only the most recent available measurement. For random forests, a limitation of RSF landmarking is that the predictions are not linked over time due to the use of independent RSF models at each landmark time result. If the longitudinal variables are extracted from EHR, monitoring of longitudinal variables may not always be organized as fixed follow-up intervals, so estimation of the model may introduce additional uncertainty and bias. The joint model has fewer restrictions on the longitudinal data, which is especially flexible for EHR. A rigidly specified follow-up plan is not required for joint modeling. These characteristics significantly increase the applicability of this joint model.

At each timepoint of longitudinal biomarkers measurement from EHR, our model can offer dynamic subject-specific predictions of MI/CV-death and all-cause of death. For patients who already have CAD, an accurate prediction model for their prognosis that is updated in real-time is crucial. This strategy may direct the frequency of tailored assessments and promote earlier diagnosis, hence improving prognosis and the timing of disease-modifying drug intervention, once accessible. The previous models[8, 22, 23] only considered the biomarkers measured at a few specific time points to the end of treatment. If these models are employed to forecast survival probability and stratify risk groups, individual variations in response to therapies will be disregarded. Although some joint models[24, 25] fulfill the dynamic prediction of cardiovascular risk, they only include one longitudinal biomarker in the model. Our model allows for a more comprehensive consideration of multiple factors associated with survival outcomes. Future prospective studies should also investigate a customized real-time therapy adaptation system to guide the treatment and health care based on the dynamic patterns of the longitudinal risk factors from EHR.

There are some limitations of our study. First, the joint model is computationally intensive, particularly for large datasets, leading to longer training and inference times. In our study, we have more than 6,000 subjects and 170,000 longitudinal records. More computation power is required for consideration of multiple longitudinal variables, their cumulative effects, and the trajectories over time. Second, the discrimination ability of the joint model did not show improvement compared with the Cox model. There could be several reasons. EHR data are not measured at regular intervals. Patients may have many measurements in a short period of time or no measurements for a very long time. As multiple variables are not measured simultaneously, large amount of missing data increases the computational effort, which may lead to inaccuracies in the estimation. Lastly, factors that are not accounted for by the proposed models may influence prediction and prediction performance.

REFERENCES

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR *et al*: **Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association**. *Circulation* 2019, **139**(10):e56-e528.
2. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, de Ferranti SD, Floyd J, Fornage M, Gillespie C *et al*: **Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association**. *Circulation* 2017, **135**(10):e146-e603.
3. Trogdon JG, Finkelstein EA, Nwaise IA, Tangka FK, Orenstein D: **The economic burden of chronic cardiovascular disease for major insurers**. *Health Promot Pract* 2007, **8**(3):234-242.
4. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB: **General cardiovascular risk profile for use in primary care: the Framingham Heart Study**. *Circulation* 2008, **117**(6):743-753.
5. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J, 3rd: **Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study**. *Ann Intern Med* 1961, **55**:33-50.
6. Amor AJ, Serra-Mir M, Martinez-Gonzalez MA, Corella D, Salas-Salvado J, Fito M, Estruch R, Serra-Majem L, Aros F, Babio N *et al*: **Prediction of Cardiovascular Disease by the Framingham-REGICOR Equation in the High-Risk PREDIMED Cohort: Impact of the Mediterranean Diet Across Different Risk Strata**. *J Am Heart Assoc* 2017, **6**(3).
7. Chia YC, Gray SY, Ching SM, Lim HM, Chinna K: **Validation of the Framingham general cardiovascular risk score in a multiethnic Asian population: a retrospective cohort study**. *BMJ Open* 2015, **5**(5):e007324.
8. Sayadi M, Zare N, Attar A, Ayatollahi SMT: **Improved Landmark Dynamic Prediction Model to Assess Cardiovascular Disease Risk in On-Treatment Blood Pressure Patients: A Simulation Study and Post Hoc Analysis on SPRINT Data**. *Biomed Res Int* 2020, **2020**:2905167.
9. Keogh RH, Seaman SR, Barrett JK, Taylor-Robinson D, Szczesniak R: **Dynamic Prediction of Survival in Cystic Fibrosis: A Landmarking Analysis Using UK Patient Registry Data**. *Epidemiology* 2019, **30**(1):29-37.
10. Pickett KL, Suresh K, Campbell KR, Davis S, Juarez-Colunga E: **Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker**. *BMC Med Res Methodol* 2021, **21**(1):216.
11. Lin X, Li R, Yan F, Lu T, Huang X: **Quantile residual lifetime regression with functional principal component analysis of longitudinal data for dynamic prediction**. *Stat Methods Med Res* 2019, **28**(4):1216-1229.
12. Campbell KR, Martins R, Davis S, Juarez-Colunga E: **Dynamic prediction based on variability of a longitudinal biomarker**. *BMC Med Res Methodol* 2021, **21**(1):104.
13. Ferrer L, Putter H, Proust-Lima C: **Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment**. *Stat Methods Med Res* 2019, **28**(12):3649-3666.
14. Gong X, Hu M, Zhao L: **Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis**. *Clin Transl Sci* 2018, **11**(3):305-311.
15. Wulfsohn MS, Tsiatis AA: **A joint model for survival and longitudinal data measured with error**. *Biometrics* 1997, **53**(1):330-339.
16. Andrinopoulou ER, Rizopoulos D, Jin R, Bogers AJ, Lesaffre E, Takkenberg JJ: **An introduction to mixed models and joint modeling: analysis of valve function over time**. *Ann Thorac Surg* 2012, **93**(6):1765-1772.
17. Rizopoulos D, Taylor JM, Van Rosmalen J, Steyerberg EW, Takkenberg JJ: **Personalized screening intervals for biomarkers using joint models for longitudinal and survival data**. *Biostatistics* 2016, **17**(1):149-164.

18. Ko YA, Hayek S, Sandesara P, Samman Tahhan A, Quyyumi A: **Cohort profile: the Emory Cardiovascular Biobank (EmCAB)**. *BMJ Open* 2017, **7**(12):e018753.

19. Dimitris Rizopoulos GP, Pedro Miranda Afonso: **JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data**. In., Version 0.3-0

https://CRAN.R-project.org/package=JMbayes2 edn; 2022.

20. Ibrahim JG, Chu H, Chen LM: **Basic concepts and methods for joint models of longitudinal and survival data**. *J Clin Oncol* 2010, **28**(16):2796-2801.

21. Liu Q, Tang G, Costantino JP, Chang C-CH: **Landmark Proportional Subdistribution Hazards Models for Dynamic Prediction of Cumulative Incidence Functions**. In.; 2019: arXiv:1904.09002.

22. Suchy-Dicey AM, Wallace ER, Mitchell SV, Aguilar M, Gottesman RF, Rice K, Kronmal R, Psaty BM, Longstreth WT, Jr.: **Blood pressure variability and the risk of all-cause mortality, incident myocardial infarction, and incident stroke in the cardiovascular health study**. *Am J Hypertens* 2013, **26**(10):1210-1217.

23. Haring R, Teng ZY, Xanthakis V, Coviello A, Sullivan L, Bhasin S, Murabito JM, Wallaschofski H, Vasan RS: **Association of sex steroids, gonadotrophins, and their trajectories with clinical cardiovascular disease and all-cause mortality in elderly men from the Framingham Heart Study**. *Clin Endocrinol* 2013, **78**(4):629-634.

24. Posch F, Ay C, Stoger H, Kreutz R, Beyer-Westendorf J: **Longitudinal kidney function trajectories predict major bleeding, hospitalization and death in patients with atrial fibrillation and chronic kidney disease**. *Int J Cardiol* 2019, **282**:47-52.

25. de Kat AC, Verschuren WM, Eijkemans MJ, Broekmans FJ, van der Schouw YT: **Anti-Mullerian Hormone Trajectories Are Associated With Cardiovascular Disease in Women: Results From the Doetinchem Cohort Study**. *Circulation* 2017, **135**(6):556-565.

APPENDIX

Table 1: Baseline characteristics of Emory Biobank participants stratified by gender.

Baseline variable	All (N=6119)	Female (N= 2145)	Male (N=3974)
Age (years)			
Mean (SD)	62.892 (12.662)	62.650 (13.522)	63.023 (12.172)
Range	18.645 - 99.627	18.645 - 98.795	20.879 - 99.627
Race			
Caucasian White	4661 (76.2%)	1510 (70.4%)	3151 (79.3%)
African American Black	1202 (19.7%)	563 (26.3%)	639 (16.1%)
Hispanic	51 (0.8%)	19 (0.9%)	32 (0.8%)
Asian	101 (1.7%)	23 (1.1%)	78 (2.0%)
Native American	7 (0.1%)	4 (0.2%)	3 (0.1%)
Pacific Islander	2 (0.0%)	0 (0.0%)	2 (0.1%)
Other	93 (1.5%)	25 (1.2%)	68 (1.7%)
Black			
Yes	1202 (19.6%)	563 (26.2%)	639 (16.1%)
History of Hypertension			
Yes	4704 (76.9%)	1650 (76.9%)	3054 (76.8%)
History of Diabetes Mellitus			
Yes	2143 (35.0%)	743 (34.6%)	1400 (35.2%)
History of hypercholesterolemia			
Yes	4292 (70.1%)	1424 (66.4%)	2868 (72.2%)
Ever smoker			
Yes	3984 (65.1%)	1231 (57.4%)	2753 (69.3%)
History of myocardial infarction			
Yes	1406 (23.0%)	396 (18.5%)	1010 (25.4%)
History of heart failure			
Yes	2246 (36.7%)	797 (37.2%)	1449 (36.5%)
Highest Level of Education			
Elementary or Middle School	191 (3.1%)	73 (3.4%)	118 (3.0%)
Some High School	565 (9.2%)	243 (11.3%)	322 (8.1%)
High School Graduate	1698 (27.7%)	664 (31.0%)	1034 (26.0%)
Some College	1417 (23.2%)	547 (25.5%)	870 (21.9%)
College Graduate	1242 (20.3%)	366 (17.1%)	876 (22.0%)
Graduate Education or Degree	1006 (16.4%)	252 (11.7%)	754 (19.0%)
Significant CAD			
Yes	3237 (72.9%)	884 (63.1%)	2353 (77.4%)
N-Miss	1679	744	935
Normal catheterization			
Yes	691 (14.7%)	300 (19.6%)	391 (12.4%)
N-Miss	1424	615	809

eGFR			
Mean (SD)	72.514 (24.388)	72.298 (26.347)	72.630 (23.265)
Range	2.333 - 175.481	2.333 - 175.481	2.365 - 154.769
History of revascularization			
Yes	2986 (48.8%)	801 (37.3%)	2185 (55.0%)

eGFR, estimated glomerular filtration rate; Significant coronary artery disease (CAD) is defined as at least one artery with 50% or more stenosis based on angiogram findings.

Table 2: Estimation of hazard ratio for joint model of MI/CV-death and all-cause death

	MI/CV-death				All-cause death			
	Coefficient	Hazard Ratio	95% CI	P-value	Coefficient	Hazard Ratio	95% CI	P-value
Baseline variables:								
Age	0.012	1.012	(1.003, 1.021)	**0.009**	0.027	1.027	(1.019, 1.037)	**0.000**
Male	-0.440	0.644	(0.542, 0.763)	**0.000**	-0.366	0.693	(0.589, 0.816)	**0.000**
Black	0.289	1.335	(1.128, 1.601)	**0.002**	0.076	1.079	(0.914, 1.264)	0.353
Highest Level of Education (Elementary or Middle School as the reference group)	.	.	.	.				
Some High School	-0.008	0.992	(0.697, 1.434)	0.947	-0.027	0.974	(0.723, 1.335)	0.868
High School Graduate	-0.014	0.986	(0.717, 1.395)	0.904	-0.073	0.929	(0.713, 1.237)	0.585
Some College	0.011	1.011	(0.728, 1.43)	0.967	-0.169	0.845	(0.644, 1.131)	0.249
College Graduate	-0.352	0.703	(0.504, 1.005)	**0.052**	-0.287	0.751	(0.561, 1.012)	0.062
Graduate Education or Degree	-0.365	0.695	(0.487, 1.003)	**0.051**	-0.464	0.629	(0.469, 0.853)	**0.007**
History of Hypertension	0.223	1.250	(1.054, 1.497)	**0.009**	0.248	1.281	(1.096, 1.508)	**0.001**
History of Diabetes Mellitus	0.156	1.169	(0.999, 1.35)	**0.052**	0.205	1.228	(1.075, 1.408)	**0.002**
History of hypercholesterolemia	-0.027	0.973	(0.83, 1.138)	0.756	-0.226	0.798	(0.699, 0.914)	**0.001**
Ever smoker	0.128	1.136	(0.988, 1.314)	0.079	0.192	1.211	(1.063, 1.377)	**0.005**
History of myocardial infarction	0.220	1.246	(1.072, 1.454)	**0.004**	0.016	1.016	(0.879, 1.177)	0.818
History of heart failure	0.436	1.546	(1.349, 1.774)	**0.000**	0.515	1.673	(1.481, 1.887)	**0.000**
History of revascularization	0.202	1.224	(1.057, 1.431)	**0.007**	0.039	1.039	(0.914, 1.182)	0.550
Longitudinal variables:								

eGFR (mL/min/1.73 m²)	-0.017	0.984	(0.98, 0.987)	**0.000**	-0.018	0.982	(0.979, 0.986)	**0.000**
eGFR (slope) (mL/min/1.73 m²/year)	-0.049	0.952	(0.921, 0.986)	**0.013**	-0.145	0.865	(0.799, 0.941)	**0.000**
log(BMI) (kg/m²)	-1.422	0.241	(0.156, 0.371)	**0.000**	-1.827	0.161	(0.106, 0.243)	**0.000**
log(BMI) (slope) (kg/m²/year)	-22.499	exp (-22.4991)	(exp(-29.4230), exp(-15.6880)	**0.000**	-95.336	0.000	(0, 0)	**0.000**
HDL (mg/dL)	-0.022	0.978	(0.97, 0.987)	**0.000**	-0.016	0.984	(0.976, 0.992)	**0.000**
HDL (slope) (mg/dL/year)	-0.815	0.443	(0.318, 0.618)	**0.000**	-1.694	0.184	(0.087, 0.417)	**0.000**
LDL (mg/dL)	0.000	1.000	(0.997, 1.004)	0.833	0.000	1.000	(0.996, 1.003)	0.932
LDL (slope) (mg/dL/year)	0.070	1.073	(0.998, 1.149)	0.060	0.187	1.205	(0.942, 1.52)	0.145
Troponion-I (ng/mL)	0.062	1.064	(1.05, 1.079)	**0.000**	0.043	1.044	(1.013, 1.076)	**0.006**
Troponion-I (slope) (ng/mL/year)	0.013	1.013	(0.975, 1.057)	0.519	0.061	1.063	(0.875, 1.302)	0.551
DBP (mmHg)	-0.005	0.995	(0.979, 1.012)	0.496	0.001	1.001	(0.984, 1.017)	0.934
DBP (slope) (mmHg/year)	0.003	1.003	(0.877, 1.153)	0.965	0.483	1.620	(1.077, 2.52)	**0.015**
SBP (mmHg)	-0.014	0.986	(0.978, 0.994)	**0.000**	-0.019	0.981	(0.974, 0.989)	**0.000**
SBP (slope) (mmHg/year)	0.023	1.023	(0.935, 1.115)	0.608	-0.299	0.742	(0.577, 0.949)	**0.019**
HbA1c (%)	-0.010	0.990	(0.894, 1.089)	0.860	-0.057	0.944	(0.839, 1.059)	0.337
HbA1c (slope) (%/year)	-0.808	0.446	(0.09, 1.953)	0.343	-9.308	0.000	(0, 0.004)	**0.000**

eGFR, estimated glomerular filtration rate; BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; DBP, diastolic Blood pressure; SBP, systolic blood pressure; HbA1c, hemoglobin A1c

Table3: Time-dependent AUC and Brier score of joint model and Cox model

			Using information up to t follow-up (years)									
			2		3		4		5		6	
		Δt (years)	Cox model	Joint model	Cox model	Joint model	Cox model	Joint model	Cox model	Joint model	Cox model	Joint model
AUC	MI/CV-death	2	0.616	0.698	0.628	0.694	0.638	0.670	0.631	0.686	0.635	0.657
		1	0.621	0.666	0.637	0.725	0.661	0.662	0.614	0.669	0.627	0.694
	All-cause death	2	0.641	0.693	0.611	0.701	0.651	0.701	0.648	0.706	0.639	0.703
		1	0.657	0.690	0.630	0.691	0.681	0.704	0.671	0.689	0.661	0.702
Brier Score	MI/CV-death	2	0.063	0.060	0.066	0.064	0.070	0.069	0.070	0.064	0.067	0.062
		1	0.034	0.033	0.033	0.032	0.037	0.037	0.039	0.037	0.035	0.032
	All-cause death	2	0.064	0.062	0.065	0.062	0.067	0.066	0.063	0.061	0.057	0.054
		1	0.034	0.034	0.034	0.033	0.034	0.034	0.037	0.037	0.030	0.028

t of the first row means the the dynamic prediction uses the longitudinal measurements up to *t* years. Δt of he third column means the prediction interval of dynamic prediction. The third row represent the model used to dynamically predict the survival outcome. The AUCs were calculated to assess how well the longitudinal marker distinguished the status of patients at time t+Δt. The Brier score is a metric used to assess the precision of a predicted survival function at time t+Δt. Higher AUCs indicate higher discrimination of the models and lower Brier score indicates worse precision of prediction.

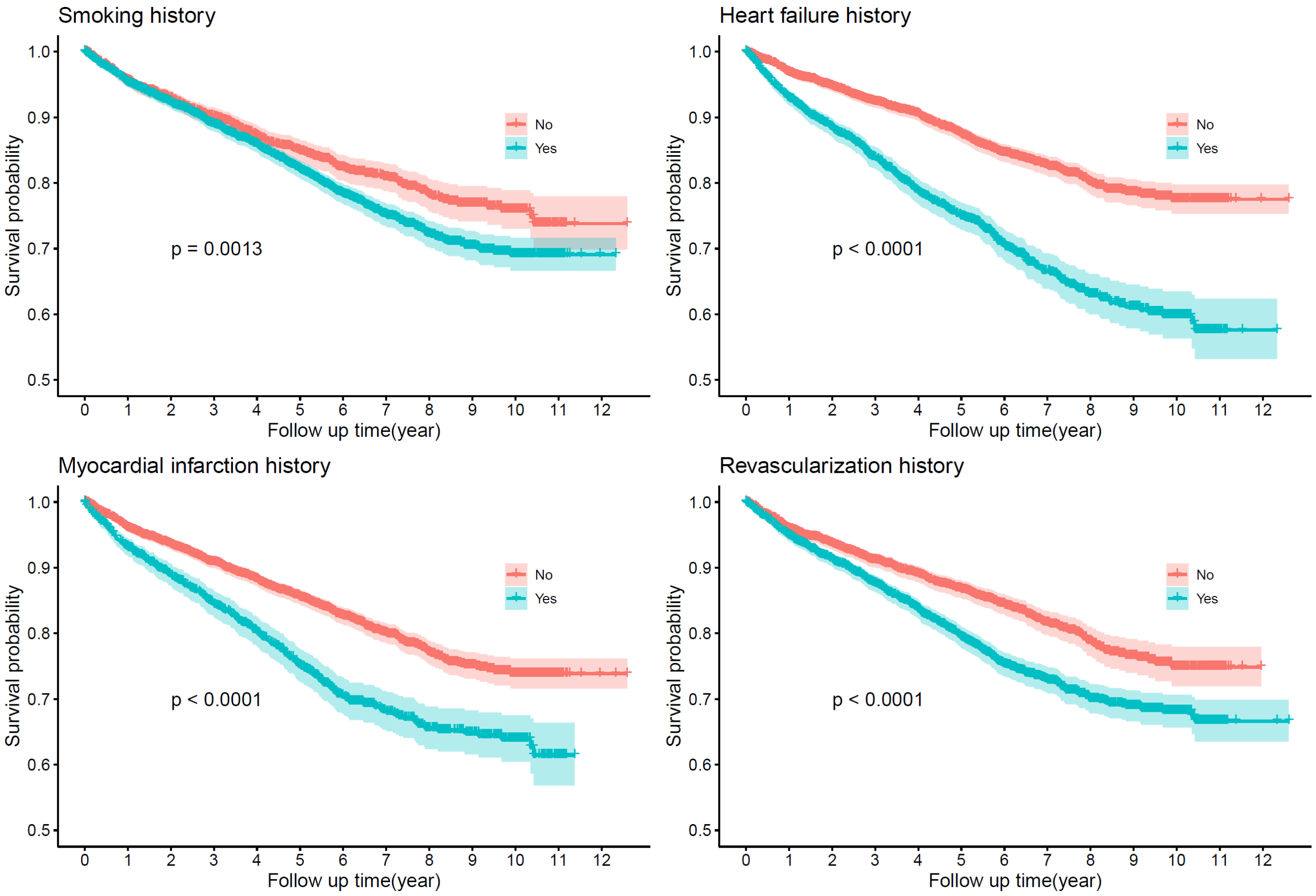

Figure 1: Kaplan Meier plot of MI/CV-death stratified by smoking history, heart failure history, myocardial infarction history, and revascularization history

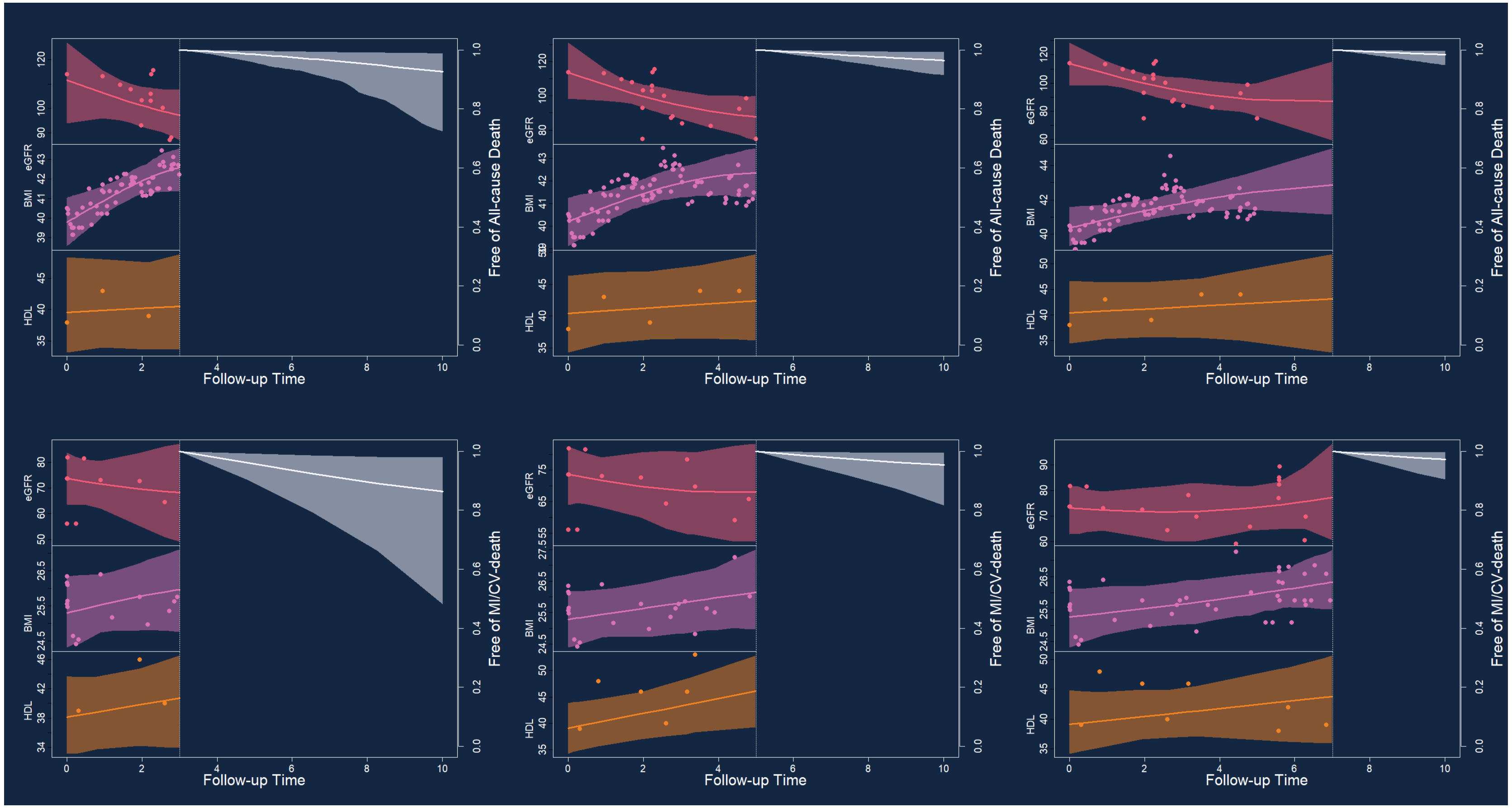

Figure 2: The dynamic prediction of MI/CV-death and all-cause death probabilities for 2 different patients during follow-up

The horizontal axis shows the number of years in follow-up, with a vertical dotted line indicating the time of longitudinal variables. The left-hand vertical axis displays eGFR, BMI, and HDL, with observed values denoted by stars and a solid line showing the longitudinal trajectory. The right-hand vertical axis presents the mean survival probability estimate and 95% confidence interval using dashed lines.