Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Multivariate Longitudinal Data for Survival Analysis of Cardiovascular Event Prediction

Sharmin Akter

RADI 6336641

Introduction

- CVD – Cardiovascular Diseases
- Risks for CVD:
  - high blood pressure
  - high low-density lipoprotein (LDL) cholesterol
  - Diabetes
  - smoking and secondhand smoke exposure
  - Obesity
  - unhealthy diet
  - physical inactivity

Literature Review

- Less than 8% of prediction models in studies published from 2009 to 2016 included longitudinal data as time-varying covariates
- Most studies were using baseline data and only 8 of them had time varying covariates
- Lots of ML were used but ML classifiers cannot predict the time to event, do not account for censoring, and need to be re-trained for each prediction time

Literature Review

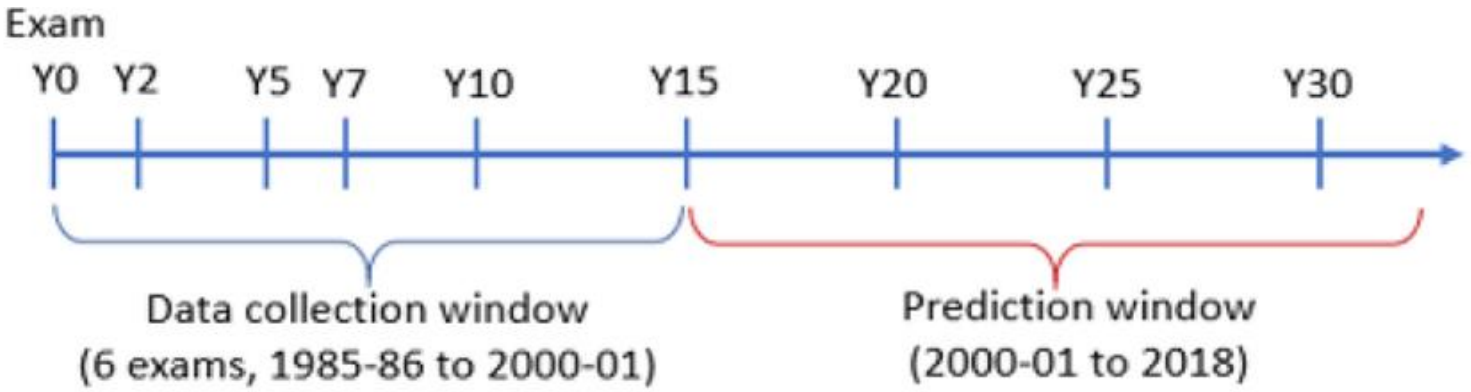
- Commonly used ML were Random Survival Forest (RSF), DeepSurv, and Nnet-survival, many cannot directly process the time series of repeated measures as input
- Recently introduced were Dynamic-DeepHit and MATCH-Net but their utilities need to be externally validated in medical applications

Aim of the study

- to evaluate the utility of multivariate longitudinal data for survival analysis of incident CVD prediction in young adults
- compare the predictive value among those strategies and against baseline models
- Identify the top performing models

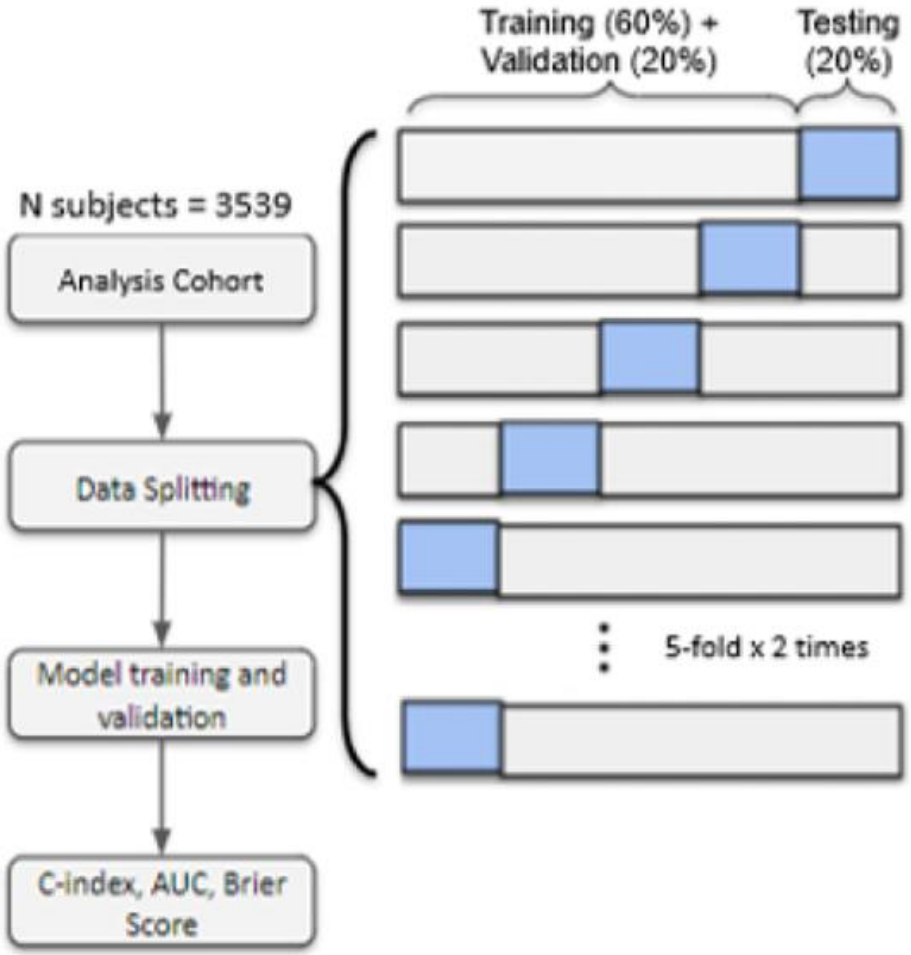
Study Population

- 3539 patients
- 15 year cohort
- Cohort from Coronary Artery Risk Development in Young Adults (CARDIA) study

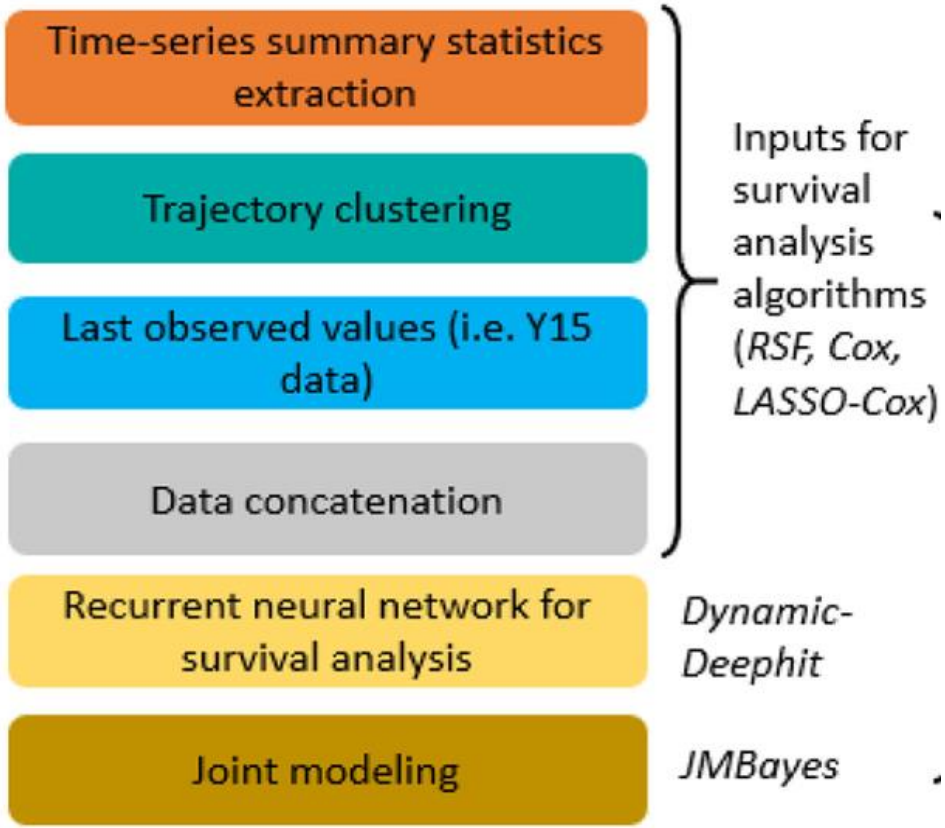


Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Methodology

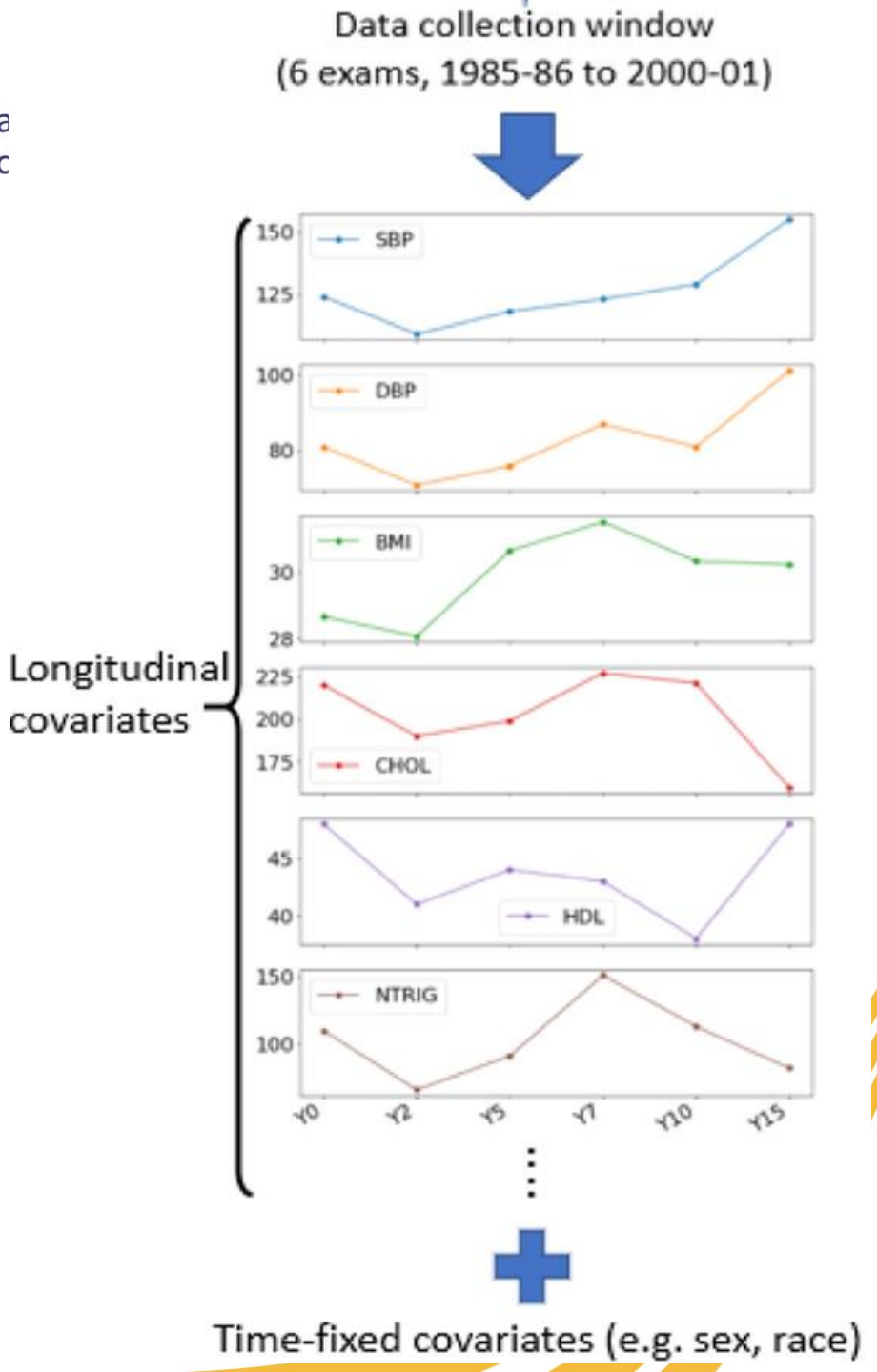


Models


Time-series summary statistics extraction
Trajectory clustering
Last observed values (i.e. Y15 data)
Data concatenation
Inputs for survival analysis algorithms (RSF, Cox, LASSO-Cox)
Recurrent neural network for survival analysis
Dynamic-Deephit
Joint modeling
JMBayes

Variables

- Total: 35 features
  - Time varying (6)
  - Baseline (29)



Results

Table 2 Predictive performance of all models on 35 variables (mean and 95% empirical bootstrap interval)

Strategy	Model	iAUC	C-index	Last AUC
Time-series (TS) massive feature extraction	RSF on TS-extracted features	**0.808 (0.790, 0.826)**	**0.778 (0.757, 0.801)**	0.758 (0.733, 0.784)
	LASSO-Cox on TS-extracted features	0.744 (0.711, 0.781)	0.713 (0.686, 0.739)	0.701 (0.674, 0.727)
Recurrent neural network	Dynamic-DeepHit	0.794 (0.764, 0.825)	0.767 (0.745, 0.789)	**0.762 (0.733, 0.792)**
Trajectory clustering	RSF on trajectory clustering data	0.793 (0.772, 0.816)	0.741 (0.721, 0.76)	0.725 (0.705, 0.744)
Data concatenation	RSF on concatenated data	0.797 (0.778, 0.817)	0.766 (0.745, 0.788)	0.751 (0.725, 0.779)
Joint modeling	JMBayes	Did not converge		
Last observed values	RSF on Y15 data	0.793 (0.773, 0.812)	0.750 (0.729, 0.77)	0.731 (0.705, 0.76)
	Cox on Y15 data	0.778 (0.758, 0.804)	0.75 (0.733, 0.769)	0.728 (0.705, 0.752)
	Cox on Y15 data	0.793 (0.772, 0.818)	0.748 (0.73, 0.763)	0.727 (0.707, 0.745)
Reference (Y0 data)	RSF on Y0 data	0.754 (0.73, 0.777)	0.721 (0.698, 0.743)	0.699 (0.672, 0.726)
	Cox on Y0 data	0.748 (0.724, 0.773)	0.709 (0.686, 0.73)	0.685 (0.654, 0.716)
	LASSO-Cox on Y0 data	0.739 (0.713, 0.768)	0.698 (0.678, 0.717)	0.678 (0.645, 0.711)

The best scores are bolded. iAUC: integrated AUC, LASSO-Cox: Cox Proportional Hazards penalized by LeAst Shrinkage and Selection Operator. *JMBayes* Joint modeling with Bayesian approach, *RSF* Random Survival Forest

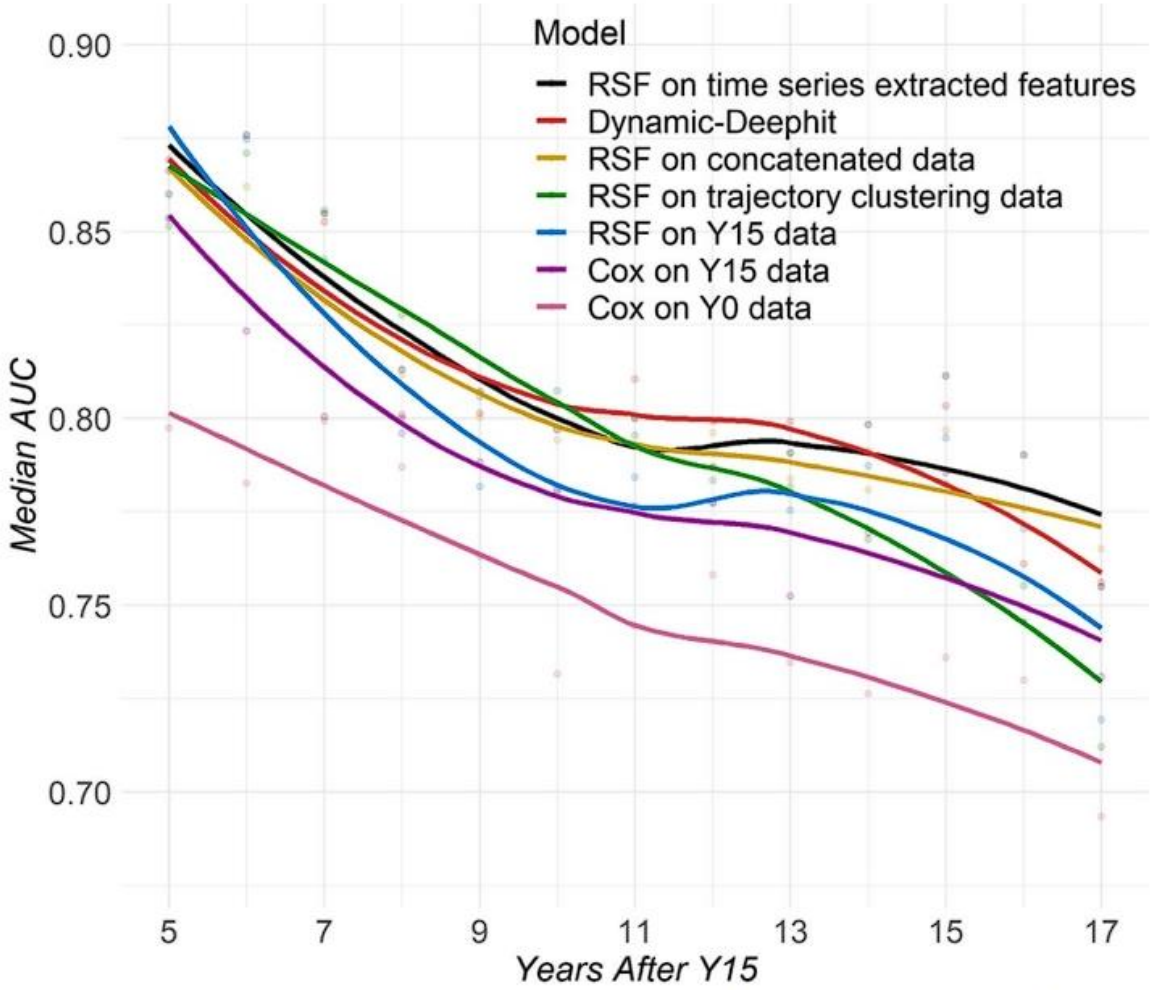


Fig. 2 Model performance over time from different longitudinal modeling strategies. Median time-varying AUC over 10 test sets is shown for all six strategies (the joint model did not converge) plus the reference using only baseline (Y0) data. RSF: Random Survival Forest

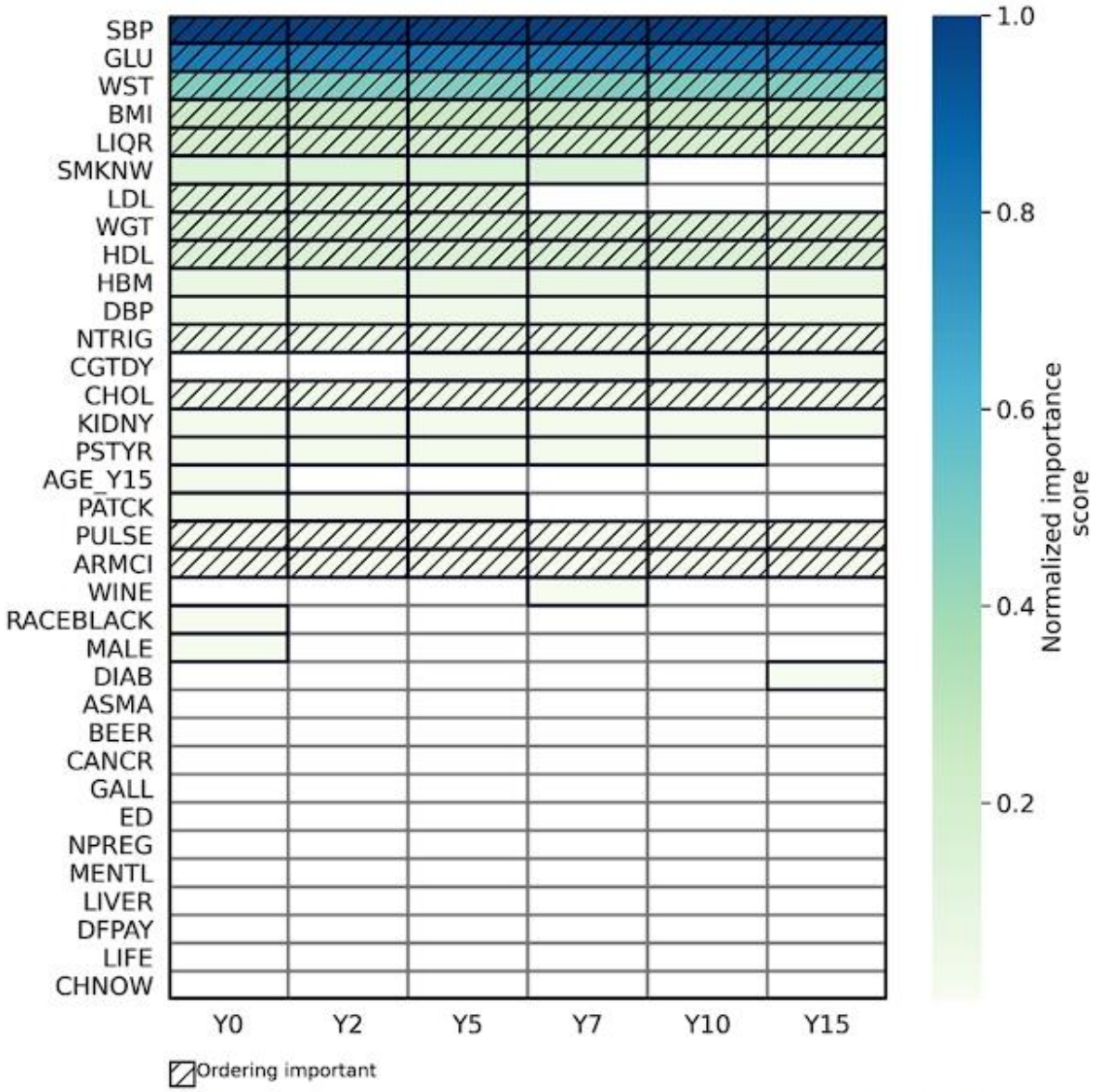


Fig. 5 Model explanation of the best performing model, RSF trained on time series extracted features, using TIME. TIME (Temporal Importance Model Explanation) is a model-agnostic longitudinal explanation method. A cell (box) is colored if it's important, is white if not deemed important by the model. Each row is a variable and shows the most important windows to the model (groups of cells in the same shade of color). The variables are ordered along the y-axis based on the overall importance (darker color = more important). Hatched texture implies the ordering within the window is important to model prediction (i.e., shuffling the variable values at different times within the window affects the model prediction)

Discussion

- RSF on TS performed the best
- longitudinal data improved up to 8% in AUC and C-index, compared to using baseline values alone, and up to 4% compared to using the last observed data
- TIME is one explainability technique that explains RSFTS better, along with the capability of explaining all temporal models using raw time series as input
- The predictors with the strongest association with lowered survival probability included HBM, smoking status, DBP, glucose, LDL, HDL, SBP, and pulse beats

Limitations

- did not exhaustively try all tuning options
- the number of CVD events by the end of follow-up was relatively small
- CARDIA study consists of Black and White participants in the US with baseline data collected in 1985, and thus the results from this work may not be transferable to other populations of different demographic characteristics
- Some models did not perform at all
- TIME evaluation do not imply causality

Reference

- Multivariate longitudinal data for survival analysis of cardiovascular event prediction in young adults: insights from a comparative explainable study | BMC Medical Research Methodology | Full Text (biomedcentral.com)