In a public health study, participants were enrolled from communities for physical exams, blood tests, and were followed up for years with disease outcomes. We are interested in investigating blood biomarkers for a prevalent and detrimental disease – diabetes.

A portion of the participants were already diagnosed to have diabetes at the time of enrollment (baseline). These participants have "existing_diabetes" outcome as 1. Some participants later on developed diabetes during the follow up period of the study. We call these new development of diabetes "incident_diabetes". For participants with incident diabetes, "diabetes_time" indicates the number of years it took to develop diabetes since baseline enrollment. For participants with existing diabetes, "diabetes_time" is negative and indicates the number of years the diagnosis of diabetes was made prior to baseline enrollment. For participants without diabetes at any time, "diabetes_time" indicates the number of years of follow up time since enrollment.

Some demographic and clinical features are self-explanatory. The list below shows more details for some of the clinical features.

Feature	Note
HDL	high-density lipoprotein
LDL	low-density lipoprotein
trig	triglyceride
SBP	systolic blood pressure
DBP	diastolic blood pressure
lipid_meds	lipid lowering medication
BP_meds	plood pressure medication
exercise	exercise at free-time: 1 [I don't move much]; 2 [moderate movement at least 4 hours a week]; 3 [exercise at least 3 hours a week]; 4 [intense exercise for competitive sports;
fasting	number of hours fasting at blood draw
existing_CHD	existing coronary heart disease
incident_CHD	incident coronary heart disease occurred after baseline enrollment
incident_HF	Incident heart failure

Blood biomarkers are denoted as "mtb_*". When there is a missing value for blood biomarkers, it is usually due to the extremely low abundance of that biomarker in blood samples which could be below the detection limit.

Question 1: Which clinical features and which blood biomarkers are associated with incident diabetes? These clinical features and biomarkers can potentially be risk or protective factors for diabetes development.

Hints: (1) Remove subjects with existing diabetes for this analysis. (2) You can consider incident diabetes as a binary outcome and use logistic regressions; or if you want to try survival analysis,

you can use Cox regressions utilizing the provided time-to-event information. (3) Investigate each individual features and biomarkers and use p-values to justify your findings. (4) For blood biomarkers, use transformation that is robust when there are outliers present. (5) *Demonstrate your findings with visualization.*

Question 2: How can we use blood biomarkers alone to predict the risk of developing diabetes?

Hints: (1) Select the relevant blood biomarkers as features for your classifier. (2) Select and train a ML model to make predictions. (3) Evaluate your predictive model with ROCAUC.

Question 3: Are there any clusters based on blood biomarkers among the participants who developed incident diabetes? Can you identify which blood biomarkers defined these clusters?

Hints: (1) Use the subset of subjects who developed incident diabetes for your unsupervised learning. (2) You can choose to use all or only relevant biomarkers for clustering. (3) Select one approach to identify clusters of these subjects. (4) Identify top blood biomarkers that contributed to the clustering.