Patient data come from the *Second Manifestations of ARTerial disease* (SMART) study. The SMART study is a study coordinated by University Medical Center Utrecht, in the Netherlands. Many prediction models in the field of cardiovascular disease are developed with data from subjects without clinically manifest atherosclerosis. The aim of the study is to develop a prediction model for patients with cardiovascular disease. Of particular interest are the estimates of risk at 1-, 3-, and 5-year on the occurrence of vascular events (stroke, myocardial infarction, or cardiovascular death). We consider here 3873 patients who were enrolled in the study in the period of September 1996 and March 2006. Patients had a clinical manifestation of atherosclerosis (transient ischemic attack, ischemic stroke, peripheral arterial disease, abdominal aortic aneurysm, or coronary heart disease). After written informed consent, they underwent a standardized vascular screening including a health questionnaire for clinical information, laboratory assessment and anthropometric measurements at enrolment. The primary outcome is *any cardiovascular event*, comprising cardiovascular death, nonfatal stroke, and nonfatal myocardial infarction.

Data structure:

variable	**labels**	**category**
TEVENT	time to event (days)	
EVENT	1= EVENT OF INTEREST; 0=CENSORED	
SEX	2= female; 1=male	Demographics
AGE	age (years)	Demographics
DIABETES	diabetes	Classical Risk factor
CEREBRAL	cerebral	Previous symptomatic atherosclerosis
CARDIAC	cardiac	Previous symptomatic atherosclerosis
AAA	abdominal aortic aneurysm	Previous symptomatic atherosclerosis
PERIPH	Peripheral	Previous symptomatic atherosclerosis
STENOSIS	Carotid artery stenosis >50%	Marker of atherosclerosis
SYSTBP	Systolic, automatic (mmHg)	Blood Pressure
DIASTBP	Diastolic, automatic (mmHg)	Blood Pressure
SYSTH	Systolic, by hand (mmHg)	Blood Pressure
DIASTH	Diastolic, by hand (mmHg)	Blood Pressure
LENGTH	height (m)	Characteristics of the patient
WEIGHT	weight (kg)	Characteristics of the patient
BMI	Body Mass Index (in kg/m2)	Classical Risk factor
CHOL	Total cholesterol (mmol/L)	Lipid levels
HDL	High-density lipoprotein cholesterol (mmol/L)	Lipid levels
LDL	Low-density lipoprotein cholesterol (mmol/L)	Lipid levels
TRIG	Triglycerides (mmol/L)	Lipid levels
HOMOC	Homocysteine (µmol/L)	Marker of atherosclerosis
GLUT	Glutamine (µmol/L)	Marker of atherosclerosis
CREAT	Creatinine clearance (mL/min)	Marker of atherosclerosis

IMT	Intima–media thickness (mm)	Marker of atherosclerosis
albumin	1= No; 2= Micro; 3=Macro	Marker of atherosclerosis
SMOKING	1= Never; 2= Former; 3=Current	Classical Risk factor
packyrs	Packyears (years)	Classical Risk factor
alcohol	1= Never; 2= Former; 3=Current	Classical Risk factor

Of note:

In the first years of the study, blood pressure was measured combined with measurement of the distensibility of the carotid artery wall ("SYSTBP" and "DIASTBP" variables). Four years after the start of the study, it was decided to measure blood pressure with the conventional sphygmomanometry as well ("by hand"). This measurement is considered in most current guidelines. Hence, conventional diastolic and systolic measurements ("SYSTH" and "DIASTH" variables) are obvious candidate predictors for our model rather than the automated measurements. Nearly, all patients had at least one type of blood pressure measurement, and there is a correlation between conventional and automatic measures…this could help in a possible imputation of missing values for "SYSTH" and "DIASTH" variables.

Questions:

1) Which kind of study design has been used?
2) Build a descriptive table, comparing patients with the event *versus* patients without the event. Insert also a column with the total population descriptive statistics. Comment about percentages of missing data in the candidate predictors.
3) Perform univariable Cox analyses, of all candidate predictors for your model. Verify proportional hazards assumptions. [**Optional:** For continuous variables, is the linearity effect reasonable?]
4) Build a multivariable Cox model starting from the list of significant predictors at univariable analyses.
5) Evaluate model performance by means of C-statistic (hint: R function rcorr.cens {Hmisc})
6) Represent the estimated model by means of a nomogram (hint: R function nomogram {rms})
7) Internally validate the estimated model (hint: R function validate {rms})
8) **Optional:** Setting aside the interpretability of the model, are you able to find a machine learning algorithm that predicts the risk of event (globally) and at 1-, 3-, and 5-year with a similar (or better) performance than the Cox model?