Survival analysis is the branch of statistics that aims to quantify the impact of certain trajectories in the time to an event. This event can be the diagnosis of a desease, the development of symptoms , or even death. Amongst others, one big challenge of survival analysis is the inclusion of longitudinal measurements of patients in the prediction of time to an event. This opens the door of personalized risk prediction .

Suppose we have 100 patients with some demographic attributes which are diagnosed with Hypertension, Coronary Artery Disease or Arithmia. Every patient goes to the doctor at random time points after initial diagnosis and reports measurements of SystolicBP, DiastolicBP, TotalCholesterol, LDL, HDL, Triglycerides, HeartRate, CRP, HbA1c and Weight. There are times that although appointment is in place subjects don’t show up at all or have measurements of some trajectories missing,for example a patient shows up for appointment and has measurements of LDL and HDL but not for HbA1c.

Patients are subjects to event occurrence . Events can be Heart attack, Stroke, Hospitalization and Death.

Simulation scheme:

Three dataframes are simulated

One dataframe containing patients’ master data .This dataframe consists of columns :

PatientID, FirstName, LastName, BirthDate, Gender, Ethnicity, BaselineDiagnosisDate, PrimaryDiagnosis, SmokingStatus, AlcoholUse, BaselineBMI, BaselineMedications.

Values of columns are simulated with Faker and random packages .At first a gender is picked randomly between male and female.If male is picked then a male name is picked randomly , otherwise a female name is picked.Rest of the columns are picked disregarding gender pick as follows:

Lastname : picked randomly by faker

BirthDate : picked randomly with faker so that age is between 30 and 80

Ethnicity : randomly between Caucasian,Asian, African American, Hispanic,Other

BaselineDiagnosisDate: picked randomly with faker so that it is in the last 10 years

PrimaryDiagnosis: randomly between Hypertension,Coronary Artery Disease, Arrhythmia

SmokingStatus: randomly between Never, Current, Former

AlcoholUse: Yes,No,Occasional

BaselineBMI: from uniform distribution between 18.5 and 35.0 rounded to 1st decimal

BaselineMedications: randomly between Statins,Beta Blockers,ACE Inhibitors, None

After Patients have been simulated their events dataframe is drawn. Events dataframe contains columns : EventID, PatientID, EventDate, EventType, TreatmentType, MedicationPrescribed, Adherence and Notes .In every patient up to five events occur at random dates. Event dates are ascending. The first event date is being drawn uniformly between initial diagnosis date and today, the second event date is being drawn uniformly between first event date and today, and so on. EventId is the ascending number of each event for the PatientID th patient. EventType is drawn uniformly between Heart Attack, Stroke, Hospitalization and Death. In case of Death no more events happen for this patien (as he is dead). TreatmentType is drawn uniformly between Angioplasty, Medication Adjustment, Bypass Surgery. MedicationPrescribed is drawn uniformly between Statins, Beta Blockers, ACE Inhibitors,None. Adherence is drawn uniformly between Yes, No, Partial . At last faker provided some funny notes of 50 characters for each event.

At last we have a dataframe containing clinical measurements of certain trajectories for each patient . Columns of clinical measurements dataframe are MeasurementID,PatientID,MeasurementDate,SystolicBP,DiastolicBP,TotalCholesterol,LDL,HDL,Triglicerides, HeartRate, CRP, HbA1c and Weight. Five clinical measurements are simulated for each patient. Different parameters of uniform distribution are used to draw each trajectory based of sex of the patient, namely:

Men Women

SystolicBP: random integer between 110, 140 105, 135

DiastolicBP: random integer between 60, 90 65, 95

TotalCholesterol: random integer: between 160, 240 155, 220

LDL: random integer between 100, 160 85, 120

HDL: random integer between 40, 60 55, 75

Triglicerides: random integer between 100, 200 80, 180

HeartRate: random integer between 60 80 70, 90

CRP: random uniform between 1.0, 4.0 1.5, 4.5

HbA1c: random uniform between 5.0, 6.0 4.5, 5.5

Weight: random integer between 60, 100 50, 90

For some patients there are measurements drawn than are subsequent to their date of death. These measurements are discarded. That makes the number of measurements for every patient random too.

Missingness Induction Strategy:

It is standard in longitudinal studies for patients not to show up for appointment. In that case the doctor doesn’t have the information for the patients’ measurements had they come. To simulate this behaviour, rows are being dropped with 20% chance. To achieve this, a number from 0 to 9 is being drawn from the uniform distribution, if the number is below 2 then the row is discarded, hence all the measurements of the patient for that timepoint are being dropped.

Another common situation in longitudinal studies is for the patient to show up for appointment but not completed all the prescribed checks or some of the measurements have not been reported in the system. In that case the patients’ record in the measurements table will have missing values in the form of Nan. Suppose 10 % of the remaining measurements are missing due to one of the reasons above. To achieve this a number between 0-19 is being draw, if the result is below 2 then the patients’ measurement is set to Nan. This is repeated for every numerical column in the patients’ measurements dataframe.

Noise induction Strategy:

Suppose that the tools conducting the measurements have some random error due to wear, or the measurement values are being entered in the incorrectly. To achieve this a random number is being drawn from the normal distribution, it is multiplied by 3 and it is being added to the value of every the numerical measurent in the patients’ measurements table.