The dataset for this project is *pneudata.csv*. The data are simulated; no real patient data are contained in this dataset. The research question is:

Do statins protect against all-cause mortality after a diagnosis of pneumonia?

The data contain information on 7265 patients who had a diagnosis of incident pneumonia during the study period (July 1995 to January 2007). Whether the patient had been prescribed statins, and some basic demographic information is available, along with information on comorbidities and prior use of other medications.

The outcome was death within 6 months following diagnosis of pneumonia.
We will use the **subset** of data described in the table below.

List of key variables:

Table 1: Key variables in the artificial dataset

id	Patient ID
statin	Treatment: Prescription of statin (1=yes, 0=no)
death	Outcome variable: death within 6 months (1=yes, 0=no)
age_diag	Age at pneumonia diagnosis (years)
agecat	Categorised age at pneumonia diagnosis
male	Gender (1=male, 0=female)
smoke	Smoking status (1=never, 2=ex, 3=current, 4=unknown)
alcohol	Alcohol consumption (1=no, 2=Ex, 3=current (quantity unknown) 4=occasional, 5=moderate, 6=excessive, 7=unknown)
bmi	Body Mass Index (kg/m2; 1=<20, 2=20-25, 3=>25, 4=unknown)
diabetes	Prior diagnosis of diabetes
cvd	Prior diagnosis of cardiovascular disease
heartfail	Prior diagnosis of heart failure
dementia	Prior diagnosis of dementia
cancer	Prior diagnosis of cancer
hyperlipid	Prior diagnosis of hyperlipidaemia
aspirin	History of use of aspirin

Analysis track:

1) Tabulate percentages of prescribed statin and death in the population. What is the crude measure of association of statin with respect to mortality? How do you interpret it?

2) Summarise the others variables (age at diagnosis, male, diabetes, hyperlipid…) by statin group. Are these characteristics *similar* across the treatment (statin) groups? If there are any differences, this could depend on the study design in your opinion?

3) If you want to obtain a measure of the **effect** of the exposure to statin on mortality, *adjusted* for possible confounders, which kind of methods you can use?

4) Try **at least two** different approaches that we discussed in the course and compare the adjusted measures of association obtained with respect to the crude one.

5) **Optional:** are you able to find a machine learning (ML) algorithm that you can use to predict the probability to be prescribed with statin and use it to adjust for confounding?