Survival analysis – part 1
Practical in R

Getting started:

- **Log in to moodle http://www.ucl.ac.uk/moodle** using your UCL username and password
- Find the course: **Research Methods for Quantitative Data**
- Go to **Survival analysis** and download the dataset **elsa_cf.sav** to your workspace.
- Start ***R studio***; load the data

```
setwd("N:/…")
elsa_cf <- load ("elsa_cf.Rdata")
```

For this practical on survival analysis, we are interested in determinants of mortality in older adults who take part in the English Longitudinal Study of Ageing. Data from the ELSA study can be found on the UK Data Service website. Specifically, we want to assess the link between cognitive function and mortality risk in this aging population, and if this relationship exists, whether it reflects confounding by lifestyle, socioeconomic status or chronic disease (in particular cardiovascular disease).

Participants in the ELSA study were drawn from respondents to the Health Survey for England. They have a face-to-face interview every two years of the study and a nurse assessment every 4 years. The first assessment (wave 1) was performed in 2002/2003. Through linkage with the Office for National Statistics mortality data, we have data on vital status with date of death over the follow-up. For this specific analysis, the follow-up ends in March 2013.

A cognitive function score, composite of memory, executive functioning and processing speed was measured at baseline (wave 1). All other potential predictors of mortality present in this dataset are measured at baseline. All variable names from wave 1 ends with a "1". This includes:

- **cf1**: cognitive function score composite of memory, executive functioning score and processing speed. Possible range 0-164.
- **alcohol1**: daily consumption of alcohol 0=no, 1=yes
- **cigst1** : smoking status: 0=never 1= former 2=current smoker
- **educ1**: Higher qualifications 0= high (e.g., university degree or higher), 1= intermediate (secondary school), 2= low (completed no more than compulsory schooling)
- **totwq5_bu1**: Quintile of wealth 1=highest 5=lowest
- **physact1:** Physical inactivity at w1 (0=active 1=inactive)
- **sex:** Sex of participant 0=male; 1=female
- **age1**: Age at w1
- **cancer1**: Ever been diagnosed with cancer w1
- **chd1**: Ever been diagnosed with coronary heart disease at w1

We also have the following indicators:

- **iintdtm1:** Month of date of interview w1 (entry in study)
- **iintdty1**: Year of date of interview w1 (entry in study)

Outcome variables are as follows:

- **dodmnth:** Month of date of death
- **dodyr:** Year of date of death
- **dead**: Mortality status 0=alive; 1=dead
- **time**: Survival time in years (time from entry in study up to death or censoring)

We will use the same dataset for both sessions. In this first part, the objectives are:

- To learn how to describe time-to-event data containing censored data
- To draw Kaplan-Meier survival curves
- To compare survival between groups of interest: gender, cognitive function score

2. Getting a sense of the data

Let's take a few minutes to describe our data.

Example syntax:

```
# Get list of variables in the dataset 'elsa_cf'
names(elsa_cf)

# Get summary statistic for 'age1'
summary(elsa_cf$age1)
# Histogram separately for people with an event and survivors #
#Create a subset#
elsadead <-subset(elsa_cf, elsa_cf$death==1)
elsalive <-subset(elsa_cf, elsa_cf$death==0)
# Histogram in each subset#
qplot(elsadead$time, geom = 'histogram', bins = 40)
qplot(elsalive$time, geom = 'histogram', bins = 40)

# Tabulate the outcome 'death'
table(elsa_cf$death)
```

- How many mortality cases were observed over the follow-up period?
- What is the mean, minimum and maximum follow-up time?
- What is the median, interquartile range, minimum and maximum of follow-up time in participants who were censored? Do we have a lot of lost to follow-up?
- What is the median, interquartile range, minimum and maximum of follow-up time in participants who died?

3. Survival and Kaplan-Meier curves

R will not provide summary statistics for survival as Stata and SPSS do. We will have to calculate it by hand.

Remember that to calculate the number of person-years, each person contributes to the number of years they were followed-up for and the number of person-years is simply the sum of all person-years present in the study.
The formula to calculate incidence rate: number of deaths/person-years over the follow-up period.

The following command will define a survival object and return summary

```
km <- survfit(Surv(time, death)~ 1)
km
```

- How many person-years were contributing to the survival analysis?
- What is the incidence rate over the follow-up?
- What is the median follow-up? If you are unable to answer this question, what could be the reason for that?

Drawing a Kaplan-Meier curve will help us visualizing the probability of surviving at every time point over the entire follow-up. The following command draws the curve:

```
plot (km, lty=1, lwd=2, xlab="Time", ylab="Survival
      Probability", col=rainbow(1))
```

- Draw Kaplan-Meier survival curve.
- Do you understand now why R could not give you an estimate of the median follow-up? Interpret this finding in the context of the study.

4. Comparison between groups

- It is commonly reported that women live longer than men in the general population. Are there differences in survival between men and women in our study?

You can first assess the differences graphically by plotting at KM survival curve stratified by sex.

```
kmsex <- survfit( Surv(time, death)~ strata(sex), data=elsa_cf,
conf.type="log-log")

plot(kmsex, lty=1, lwd=1, xlab="Time", ylab="Survival
Probability", col=rainbow(2))

legend("bottomleft", c("Men", "Women") , lty=1, lwd=1,
col=rainbow(2) )
```

- What test can you perform to check if the difference is significant? What is the result for gender?

```
survdiff(Surv(time, death) ~ sex, rho=0)
```

How can you interpret this result?

Note: Remember these are crude (i.e. non-adjusted) estimates. We can plot the adjusted survival estimators but not perform an adjusted test. That will be for the next session!

5. Exercise

The main purpose of this study is to test for the potential effect of cognitive impairment on mortality. The Kaplan-Meier curves and tests available are useful to compare groups, i.e. can only deal with **categorical** variables. Our measure of cognitive function is a continuous score.

- How can we compare groups based on their cognitive score? Create an appropriate categorical variable.
- Draw the graph that compares these groups and run the appropriate test. Are there differences in survival according to cognitive function?
- Interpret these results and their limitations.

Note: To create quantiles in R:

- First ask R to return quantile values (here tertiles)
- Cut your data according to these values

```
quantile(elsa_cf$cf1, prob=c(0.33, 0.66))

elsa_cf$t_cf1 <-cut(elsa_cf$cf1, breaks=c(0, 42, 52, 194))
```