Exercises Survival Analysis Lecture 11

Marta Fiocco & Hein Putter

1 Introduction

The objective today is to recall and use together a number of concepts treated over the last couple of weeks. Since these concepts have been practiced before the computer practical will be less directive and exercises will be broader.

We will be using a data set from the EBMT with chronic myeloid leukemia (CML) patients receiving allogeneic stem cell transplantation (SCT) with peripheral blood. Transplantations from 2001 onwards were selected. Focus will be first on prognostic factors predicting relapse-free survival (time to either relapse or death, whichever occurs first). We will build a prognostic score and critically assess its performance.

The data is in the SPSS data `cml`.

```
> library(foreign)
> cml <- read.spss("cml.sav",to.data.frame=TRUE)
> head(cml)
```

```
         rfs rfsstat year      ric      agecl4     ditrcl4
1  2.2678718       0 2007  reduced  > 50 years > 12 months
2 59.1290058       0 2004 standard 31-50 years > 12 months
3 14.8562038       1 2005  reduced  > 50 years > 12 months
4 39.4412490       0 2006 standard 31-50 years > 12 months
5  3.7140509       1 2004  reduced  > 50 years 6-12 months
6  0.2958094       1 2003 standard  > 50 years 6-12 months
                     don3        fememalematch       agvh agvhstat
1                    <NA> other combinations  2.2678718        0
2              HLA id sib other combinations 59.1290058        0
3              HLA id sib other combinations 16.3023829        0
4              HLA id sib other combinations  3.2539030        1
5 matched unrelated donor other combinations  5.2917009        0
6              HLA id sib                m-f  0.2958094        0
```

2 Exploratory analysis

The outcome is relapse-free survival. In the `cml` data it is present in the time variable `rfs` (time to relapse or death or censoring in months), and in the status variable `rfsstat` (1 is relapse or death, 0 is censored).

Exercise 1 — Make a Kaplan-Meier survival curve for the whole data. You may use existing functions in the *survival* package. What is the 5-yrs probability of relapse or death?

Exercise 2 — The prognostic factors are in `year` (year of transplantation), `ric` (standard or reduced conditioning), `agecl4` (age in four classes), `ditrcl4` (interval between diagnosis and transplant, also in four classes), `don3` (donor type), and `femalematch` (gender mismatch between donor and recipient). The covariate `year` is continuous, the rest is categorical. For each of these prognostic factors make a frequency table and test whether they are predictive of RFS. Use univariate Cox regressions and the score test to assess overall significance of each of the covariates. First subtract 2000 from `year`

Exercise 3 — Make a multivariate Cox model with the variables that were trend-significant ($p < 0.10$ by the score test).

Exercise 4 — Use the function *`cox.zph`* to check whether the proportional hazards assumption holds.

Exercise 5 — Suppose that the proportional hazards assumption does not hold for age. What would be the way to proceed? Perform this alternative analysis. What is the difference between the present model and the original Cox model of Exercise **??**?

Exercise 6 — Calculate and plot the model-based survival curves for an individual with interval diagnosis-transplant equal to 10 months and transplanted in 2002, for each of the four age classes, using the stratified Cox model of Exercise **??**.

Exercise 7 — Do the same, but now based on the original Cox model of Exercise **??**. Comment on the differences between the present survival curves and those obtained in Exercise **??**.

Exercise 8 — We will use the original, proportional hazards, Cox model from now on. Calculate the prognostic index given by the model (given by $\hat{\beta}^\top Z_i$) for each individual i. Make a histogram and calculate the mean and standard deviation.

Exercise 9 — Is this analysis correct? Motivate your answer.

Exercise 10 — Perform a time-dependent Cox regression analysis using the time-dependent covariate $X(t)$ as defined above. What is the hazard ratio for RFS of aGvHD with respect to no aGvHD? Compare it with the answer of Exercise **??**.

Exercise 11 — Calculate and plot model-based RFS survival curves for patients with and without aGvHD.

Exercise 12 — Check the proportional hazards assumption. Is it satisfied? Show estimated survival curves with and without aGvHD without using the Cox model.

Exercise 13 — An alternative is to perform a landmark analysis. Construct two groups based on whether or not aGvHD occurred before 100 days. How large are these groups? Compare with the group sizes in Exercise **??** and comment on the differences. Perform a log-rank test comparing RFS *among all individuals still at risk after 100 days* between these two groups. Also perform a Cox regression using the same subset of the data and the same grouping variable. What is the result and how does it compare with the results of Exercises **??** and **??**.

Exercise 14 — Make Kaplan-Meier survival curves for these two groups, again using the same subset of the data and the same grouping variable. Would you say, judging from the figure, that the proportional hazards assumption is satisfied? Perform a formal test for the proportional hazards assumption.

Exercise 15 — Finally, we can try to combine the prognostic model based on `year` (continuous), `agecl4` and `ditrcl4` with aGvHD, using the landmark data set. Decide whether or not to use aGvHD as a stratifying variable or not. Comment on the results.