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       femalematch       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?

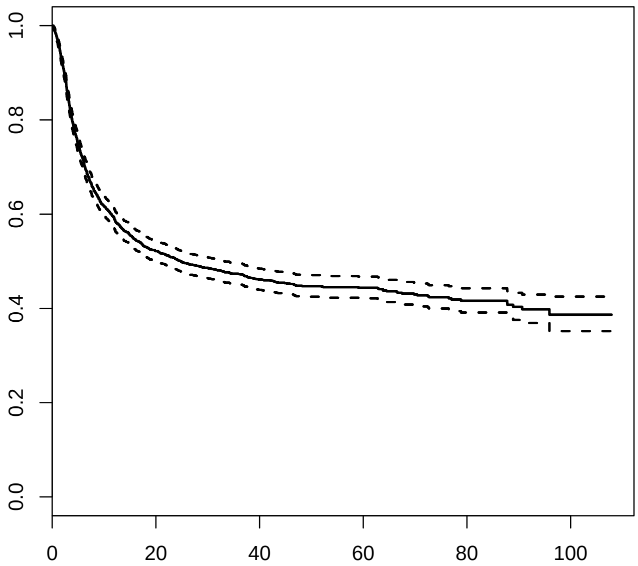

Figure 1: Kaplan-Meier survival curve

Answer — The function *survfit* can be used. The result is in Figure 1.

```
> library(survival)
> plot(survfit(Surv(rfs,rfsstat)~1,data=cml),lwd=2,mark.time=FALSE)
```

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`

Answer — Here is an example of code. Note that we use `cml$cov <- factor(cml$cov)` below. This gets rid of unused factor levels.

```
> cml$year <- cml$year - 2000
> covs <- c("year","ric","agecl4","ditrcl4","don3","femalematch")
> for (i in 1:length(covs)) {
+   cat("\n\nVariable:",covs[i],":\n\n")
+   cml$cov <- cml[,covs[i]]
+   if (i != 1) cml$cov <- factor(cml$cov)
```

```
+   print(table(cml$cov))
+   c1 <- coxph(Surv(rfs,rfsstat) ~ cov, data=cml)
+   print(summary(c1))
+ }
Variable: year :


  1   2   3   4   5   6   7   8
306 368 347 359 298 298 215 183
Call:
coxph(formula = Surv(rfs, rfsstat) ~ cov, data = cml)

  n= 2374, number of events= 1141

       coef exp(coef) se(coef)     z Pr(>|z|)
cov 0.03609   1.03675  0.01448 2.492   0.0127 *
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

    exp(coef) exp(-coef) lower .95 upper .95
cov     1.037     0.9646     1.008     1.067

Concordance= 0.521  (se = 0.009 )
Rsquare= 0.003   (max possible= 0.999 )
Likelihood ratio test= 6.18  on 1 df,   p=0.01293
Wald test            = 6.21  on 1 df,   p=0.01272
Score (logrank) test = 6.21  on 1 df,   p=0.01267



Variable: ric :


standard  reduced
    1666      708
Call:
coxph(formula = Surv(rfs, rfsstat) ~ cov, data = cml)

  n= 2374, number of events= 1141

              coef exp(coef) se(coef)     z Pr(>|z|)
covreduced 0.07112   1.07370  0.06444 1.104     0.27

           exp(coef) exp(-coef) lower .95 upper .95
covreduced     1.074     0.9314    0.9463     1.218

Concordance= 0.503  (se = 0.007 )
Rsquare= 0.001   (max possible= 0.999 )
Likelihood ratio test= 1.21  on 1 df,   p=0.272
```

```
Wald test            = 1.22  on 1 df,   p=0.2698
Score (logrank) test = 1.22  on 1 df,   p=0.2697



Variable: agecl4 :


 < 20 years 20-30 years 31-50 years  > 50 years
        53         437        1319         565
Call:
coxph(formula = Surv(rfs, rfsstat) ~ cov, data = cml)

  n= 2374, number of events= 1141

                  coef exp(coef) se(coef)     z Pr(>|z|)
cov20-30 years 0.04097   1.04182  0.22659 0.181  0.85652
cov31-50 years 0.27373   1.31486  0.21697 1.262  0.20709
cov> 50 years  0.63318   1.88359  0.22036 2.873  0.00406 **
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

               exp(coef) exp(-coef) lower .95 upper .95
cov20-30 years     1.042     0.9599    0.6682     1.624
cov31-50 years     1.315     0.7605    0.8594     2.012
cov> 50 years      1.884     0.5309    1.2230     2.901

Concordance= 0.549  (se = 0.008 )
Rsquare= 0.02   (max possible= 0.999 )
Likelihood ratio test= 47.22  on 3 df,   p=3.12e-10
Wald test            = 48.4  on 3 df,   p=1.754e-10
Score (logrank) test = 49.25  on 3 df,   p=1.152e-10



Variable: ditrcl4 :


 < 3 months  3-6 months 6-12 months > 12 months
        69         372         681        1252
Call:
coxph(formula = Surv(rfs, rfsstat) ~ cov, data = cml)

  n= 2374, number of events= 1141

                  coef exp(coef) se(coef)     z Pr(>|z|)
cov3-6 months  0.06241   1.06439  0.20876 0.299   0.7650
cov6-12 months 0.20264   1.22463  0.20068 1.010   0.3126
cov> 12 months 0.45145   1.57058  0.19646 2.298   0.0216 *
---
```

```
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

               exp(coef) exp(-coef) lower .95 upper .95
cov3-6 months      1.064     0.9395    0.7070     1.602
cov6-12 months     1.225     0.8166    0.8264     1.815
cov> 12 months     1.571     0.6367    1.0686     2.308

Concordance= 0.545  (se = 0.008 )
Rsquare= 0.012   (max possible= 0.999 )
Likelihood ratio test= 29.1  on 3 df,   p=2.13e-06
Wald test            = 28.32  on 3 df,   p=3.107e-06
Score (logrank) test = 28.6  on 3 df,   p=2.719e-06



Variable: don3 :


             HLA id sib matched unrelated donor
                   1403                     564
Call:
coxph(formula = Surv(rfs, rfsstat) ~ cov, data = cml)

  n= 1967, number of events= 950
   (407 observations deleted due to missingness)

                              coef exp(coef) se(coef)     z Pr(>|z|)
covmatched unrelated donor 0.09994   1.10510  0.07041 1.419    0.156

                           exp(coef) exp(-coef) lower .95 upper .95
covmatched unrelated donor     1.105     0.9049    0.9627     1.269

Concordance= 0.512  (se = 0.008 )
Rsquare= 0.001   (max possible= 0.999 )
Likelihood ratio test= 1.99  on 1 df,   p=0.1585
Wald test            = 2.01  on 1 df,   p=0.1558
Score (logrank) test = 2.02  on 1 df,   p=0.1556



Variable: femalematch :


other combinations                m-f
              1828                492
Call:
coxph(formula = Surv(rfs, rfsstat) ~ cov, data = cml)

  n= 2320, number of events= 1112
   (54 observations deleted due to missingness)
```

```
         coef exp(coef) se(coef)     z Pr(>|z|)
covm-f 0.09299   1.09745  0.07184 1.294    0.196

       exp(coef) exp(-coef) lower .95 upper .95
covm-f     1.097     0.9112    0.9533     1.263

Concordance= 0.506  (se = 0.006 )
Rsquare= 0.001   (max possible= 0.999 )
Likelihood ratio test= 1.65  on 1 df,   p=0.1994
Wald test            = 1.68  on 1 df,   p=0.1955
Score (logrank) test = 1.68  on 1 df,   p=0.1954
```

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

Answer — The variables were `year` (continuous), `agecl4` and `ditrcl4`.

```
> cml$agecl4 <- factor(cml$agecl4)
> cml$ditrcl4 <- factor(cml$ditrcl4)
> c2 <- coxph(Surv(rfs,rfsstat) ~ agecl4 + ditrcl4 + year, data=cml)
> summary(c2)
```

```
Call:
coxph(formula = Surv(rfs, rfsstat) ~ agecl4 + ditrcl4 + year,
    data = cml)

  n= 2374, number of events= 1141

                      coef exp(coef) se(coef)      z Pr(>|z|)
agecl420-30 years  -0.01547   0.98465  0.22727 -0.068   0.9457
agecl431-50 years   0.20757   1.23069  0.21781  0.953   0.3406
agecl4> 50 years    0.51191   1.66847  0.22252  2.301   0.0214 *
ditrcl43-6 months   0.06532   1.06750  0.20898  0.313   0.7546
ditrcl46-12 months  0.18458   1.20271  0.20083  0.919   0.3581
ditrcl4> 12 months  0.36289   1.43747  0.19742  1.838   0.0660 .
year                0.01440   1.01451  0.01493  0.965   0.3345
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

                   exp(coef) exp(-coef) lower .95 upper .95
agecl420-30 years     0.9847     1.0156    0.6307     1.537
agecl431-50 years     1.2307     0.8126    0.8031     1.886
agecl4> 50 years      1.6685     0.5994    1.0787     2.581
ditrcl43-6 months     1.0675     0.9368    0.7087     1.608
ditrcl46-12 months    1.2027     0.8315    0.8114     1.783
ditrcl4> 12 months    1.4375     0.6957    0.9762     2.117
year                  1.0145     0.9857    0.9853     1.045

Concordance= 0.568  (se = 0.009 )
Rsquare= 0.027   (max possible= 0.999 )
```

```
Likelihood ratio test= 65.66  on 7 df,   p=1.108e-11
Wald test            = 66.03  on 7 df,   p=9.354e-12
Score (logrank) test = 67.13  on 7 df,   p=5.597e-12
```

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

Answer — Lucky for us the proportional hazards assumption does not seem to be violated.

```
> cox.zph(c2)
                         rho   chisq     p
agecl420-30 years  -0.01562 0.28041 0.596
agecl431-50 years  -0.00563 0.03640 0.849
agecl4> 50 years    0.00423 0.02064 0.886
ditrcl43-6 months   0.01528 0.26769 0.605
ditrcl46-12 months  0.01677 0.32235 0.570
ditrcl4> 12 months -0.00120 0.00166 0.967
year               -0.00588 0.04181 0.838
GLOBAL                   NA 5.84753 0.558
```

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 3?

Answer — We could fit a stratified Cox proportional hazards model, stratified by age. This would yield separate baseline hazards for the four age classes and regression coefficients as before for `ditrcl4` and `year`. Here is the code

```
> c2str <- coxph(Surv(rfs,rfsstat) ~ ditrcl4 + year + strata(agecl4), data=cml)
> summary(c2str)
Call:
coxph(formula = Surv(rfs, rfsstat) ~ ditrcl4 + year + strata(agecl4),
    data = cml)

  n= 2374, number of events= 1141

                      coef exp(coef) se(coef)     z Pr(>|z|)
ditrcl43-6 months  0.05662   1.05825  0.20904 0.271   0.7865
ditrcl46-12 months 0.17122   1.18675  0.20089 0.852   0.3941
ditrcl4> 12 months 0.35210   1.42205  0.19743 1.783   0.0745 .
year               0.01544   1.01556  0.01494 1.034   0.3013
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

                   exp(coef) exp(-coef) lower .95 upper .95
ditrcl43-6 months      1.058     0.9450    0.7025     1.594
ditrcl46-12 months     1.187     0.8426    0.8005     1.759
ditrcl4> 12 months     1.422     0.7032    0.9657     2.094
year                   1.016     0.9847    0.9863     1.046
```

```
Concordance= 0.542  (se = 0.014 )
Rsquare= 0.008   (max possible= 0.997 )
Likelihood ratio test= 18.52  on 4 df,   p=0.0009783
Wald test            = 18.11  on 4 df,   p=0.001173
Score (logrank) test = 18.21  on 4 df,   p=0.001122
```

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 5.

Answer — First we make a `newdata` data set containing covariate values (with appropriate factor levels) corresponding with the four patients of the exercise. Subsequently we call `survfit` and make a plot, which is shown in Figure 2.

```
> newdata <- data.frame(year=2,ditrcl4=3,agecl4=1:4)
> newdata$ditrcl4 <- factor(newdata$ditrcl4, levels=1:4, labels=levels(cml$ditrcl4))
> newdata$agecl4 <- factor(newdata$agecl4, levels=1:4, labels=levels(cml$agecl4))
> newdata

  year     ditrcl4      agecl4
1    2 6-12 months  < 20 years
2    2 6-12 months 20-30 years
3    2 6-12 months 31-50 years
4    2 6-12 months  > 50 years

> sf2str <- survfit(c2str, newdata=newdata)

> plot(sf2str, mark.time=FALSE, lwd=2, col=rep(1:4,rep(4,4)),
+   xlab="Months since transplant", ylab="RFS probability")
> legend("topright",levels(cml$agecl4),lwd=2,col=1:4,bty="n")
```

Exercise 7 — Do the same, but now based on the original Cox model of Exercise 3. Comment on the differences between the present survival curves and those obtained in Exercise 6.

Answer — Again `survfit` can be used, and the same `newdata`. This time the survival curves are "parallel", whereas that wasn't the case in Exercise 6.

```
> sf2 <- survfit(c2, newdata=newdata)

> plot(sf2, mark.time=FALSE, lwd=2, col=1:4,
+   xlab="Months since transplant", ylab="RFS probability")
> legend("topright",levels(cml$agecl4),lwd=2,col=1:4,bty="n")
```

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.

Answer — The quickest way to get the individual values of the prognostic index is to use `model.matrix`, see below. Figure 4 shows the histogram.

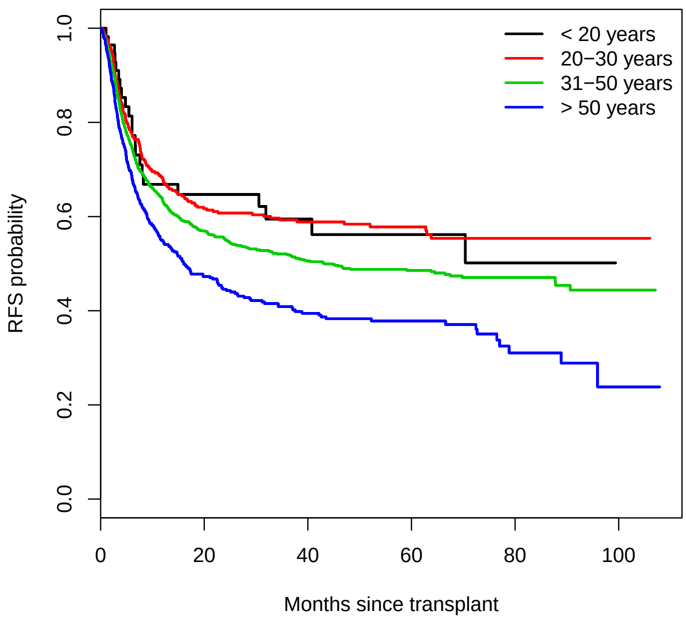

Figure 2: Model-based survival curves for four patients based on the stratified Cox model

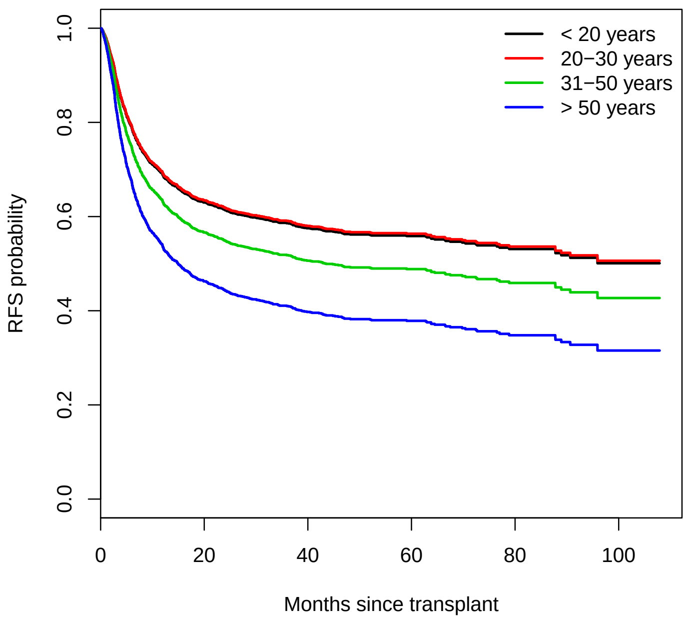

Figure 3: Model-based survival curves for four patients based on the PH Cox model

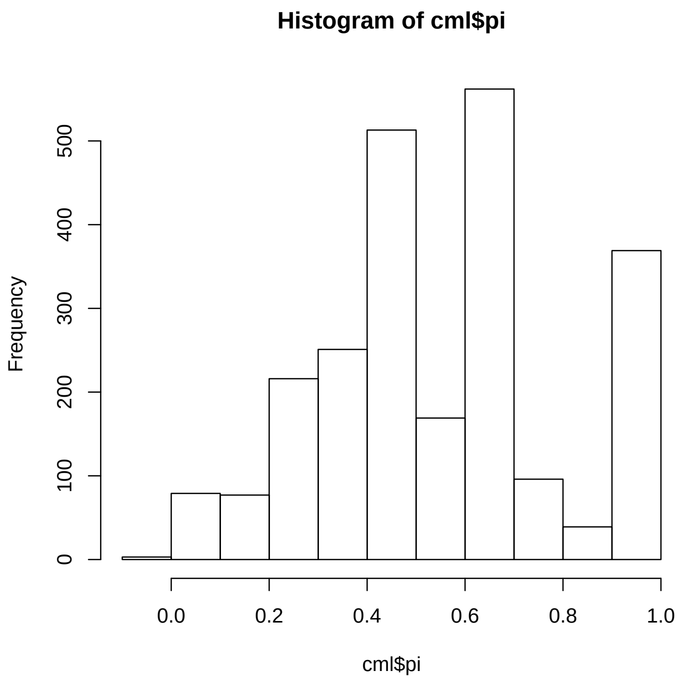

Figure 4: Histogram of the values of the prognostic index

```
> mm <- model.matrix(c2)
> cml$pi <- mm %*% c2$coef
> summary(cml$pi)
       V1
 Min.   :-0.001061
 1st Qu.: 0.390636
 Median : 0.584865
 Mean   : 0.548132
 3rd Qu.: 0.671292
 Max.   : 0.990031
> hist(cml$pi)
```

Exercise 9 — Divide the population into three equally sized risk groups defined by the prognostic index. Store it in a categorical variable with levels "Low risk", "Medium risk" and "High risk". Make Kaplan-Meier plots for each of the three sub-populations defined by these risk groups. Perform a univariate Cox regression with only this covariate.

Answer — Here is our code. The plot is shown in Figure 5.

```
> quantile(cml$pi, probs = seq(0, 1, 1/3))
          0%    33.33333%    66.66667%         100%
```

```
-0.001060928  0.420962419  0.642482999  0.990031226
> cml$risk <- cut(cml$pi, breaks=quantile(cml$pi, probs = seq(0, 1, 1/3)),
+   labels=c("Low risk","Medium risk","High risk"))
> table(cml$risk)
   Low risk Medium risk   High risk
        829         783         759
> c3 <- coxph(Surv(rfs,rfsstat) ~ risk, data=cml)
> print(summary(c3))
Call:
coxph(formula = Surv(rfs, rfsstat) ~ risk, data = cml)

  n= 2371, number of events= 1140
   (3 observations deleted due to missingness)

                   coef exp(coef) se(coef)     z Pr(>|z|)
riskMedium risk 0.28334   1.32756  0.07418 3.819 0.000134 ***
riskHigh risk   0.52989   1.69874  0.07367 7.193 6.34e-13 ***
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

                exp(coef) exp(-coef) lower .95 upper .95
riskMedium risk     1.328     0.7533     1.148     1.535
riskHigh risk       1.699     0.5887     1.470     1.963

Concordance= 0.562  (se = 0.009 )
Rsquare= 0.022   (max possible= 0.999 )
Likelihood ratio test= 52.16  on 2 df,   p=4.706e-12
Wald test            = 51.74  on 2 df,   p=5.815e-12
Score (logrank) test = 52.62  on 2 df,   p=3.741e-12
> plot(survfit(Surv(rfs,rfsstat)~risk,data=cml),col=1:3,lwd=2,mark.time=FALSE)
> legend("topright",levels(cml$risk),lwd=2,col=1:3,bty="n")
```

3 Effect of acute Graft-versus-Host Disease

One possible serious side-effect of stem cell transplantation is acute graft-versus-host-disease, acute GvHD or aGvHD in short. This is a reaction of the immune cells in the donor blood (graft) against the host, resulting in possibly life-threatening complications. The objective of this section is to study the effect of aGvHD on relapse-free survival. The complicating factor is that aGvHD is a time-dependent covariate $X(t)$, taking value 0 for t below the time of aGvHD and 1 for t beyond the time of aGvHD (if it occurs).

A common way of analyzing this type of covariates is by constructing two groups of patients, one with and one without aGvHD, as if these were known from the start, and comparing RFS between these two groups.

Exercise 10 — Perform this analysis. How large are the groups? Make a plot with the Kaplan-Meier survival curves of these groups. What is the p-value of

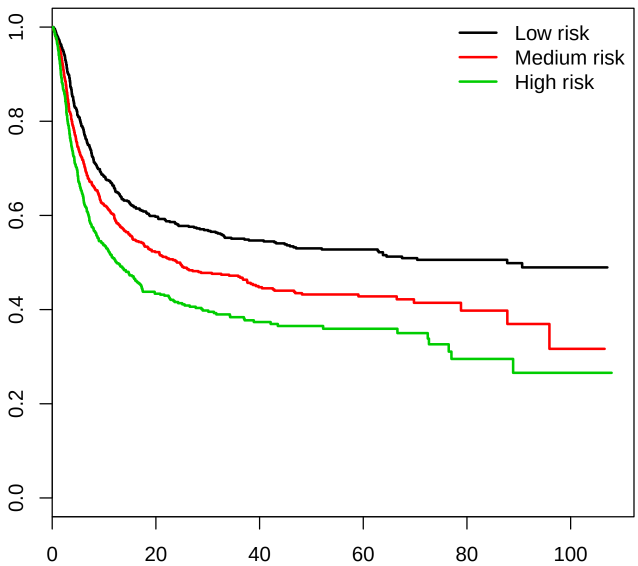

Figure 5: Kaplan-Meier survival curves for each of the risk groups

the log-rank test comparing the groups. What is the hazard ratio for RFS of aGvHD with respect to no aGvHD?

Answer — The plot is shown last in Figure 6.

```
> table(cml$agvhstat)

   0    1
1538  760

> survdiff(Surv(rfs,rfsstat) ~ agvhstat, data=cml)

Call:
survdiff(formula = Surv(rfs, rfsstat) ~ agvhstat, data = cml)

n=2298, 76 observations deleted due to missingness.

              N Observed Expected (O-E)^2/E (O-E)^2/V
agvhstat=0 1538      694      764       6.4      20.8
agvhstat=1  760      411      341      14.3      20.8

 Chisq= 20.8  on 1 degrees of freedom, p= 5.09e-06

> ca <- coxph(Surv(rfs,rfsstat) ~ agvhstat, data=cml)
> print(summary(ca))

Call:
coxph(formula = Surv(rfs, rfsstat) ~ agvhstat, data = cml)

  n= 2298, number of events= 1105
   (76 observations deleted due to missingness)

           coef exp(coef) se(coef)     z Pr(>|z|)
agvhstat 0.2833    1.3276   0.0623 4.548 5.41e-06 ***
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

         exp(coef) exp(-coef) lower .95 upper .95
agvhstat     1.328     0.7533     1.175       1.5

Concordance= 0.537  (se = 0.007 )
Rsquare= 0.009   (max possible= 0.999 )
Likelihood ratio test= 20.13  on 1 df,   p=7.217e-06
Wald test            = 20.68  on 1 df,   p=5.414e-06
Score (logrank) test = 20.82  on 1 df,   p=5.037e-06

> plot(survfit(Surv(rfs,rfsstat)~agvhstat,data=cml),col=1:2,lwd=2,mark.time=FALSE)
> legend("topright",c("No aGvHD","aGvHD"),lwd=2,col=1:2,bty="n")
```

Exercise 11 — Is this analysis correct? Motivate your answer.

Exercise 12 — Perform a time-dependent Cox regression analysis using the time-dependent covariate $X(t)$ as defined above. What is the hazard ratio for

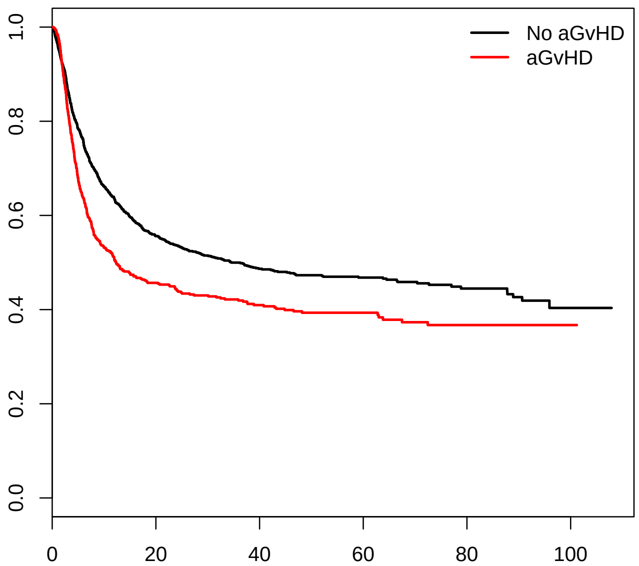

Figure 6: Kaplan-Meier survival curves for aGvHD and no aGvHD

RFS of aGvHD with respect to no aGvHD? Compare it with the answer of Exercise 10.

Answer — This requires a bit of data restructuring. The subjects with aGvHD are represented with two lines, one starting at 0 and ending at time of aGvHD with `tcov`=0, the second starting at time of aGvHD and ending at `rfs` with `tcov`=1.

```
> cml1 <- cml2 <- cml[cml$agvhstat==1,]
> cml1$Tstart <- 0
> cml1$Tstop <- cml1$agvh
> cml1$status <- 0
> cml1$tcov <- 0
> cml2$Tstart <- cml2$agvh
> cml2$Tstop <- cml2$rfs
> cml2$status <- cml2$rfsstat
> cml2$tcov <- 1
> cml3 <- cml[cml$agvhstat==0,]
> cml3$Tstart <- 0
> cml3$Tstop <- cml3$rfs
> cml3$status <- cml3$rfsstat
> cml3$tcov <- 0
> cmltd <- rbind(cml1,cml2,cml3)
> catd <- coxph(Surv(Tstart,Tstop,status) ~ tcov, data=cmltd)
> print(summary(catd))
```

```
Call:
coxph(formula = Surv(Tstart, Tstop, status) ~ tcov, data = cmltd)

  n= 2920, number of events= 996
   (366 observations deleted due to missingness)

        coef exp(coef) se(coef)     z Pr(>|z|)
tcov 0.35380   1.42447  0.07158 4.942 7.72e-07 ***
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

     exp(coef) exp(-coef) lower .95 upper .95
tcov     1.424      0.702     1.238     1.639

Concordance= 0.54  (se = 0.006 )
Rsquare= 0.008   (max possible= 0.993 )
Likelihood ratio test= 23.46  on 1 df,   p=1.274e-06
Wald test            = 24.43  on 1 df,   p=7.718e-07
Score (logrank) test = 24.66  on 1 df,   p=6.823e-07
```

Exercise 13 — Calculate and plot model-based RFS survival curves for patients with and without aGvHD.

Answer — The plot is shown in Figure 7.

```
> sfatd <- survfit(catd, newdata=data.frame(tcov=0:1))
```

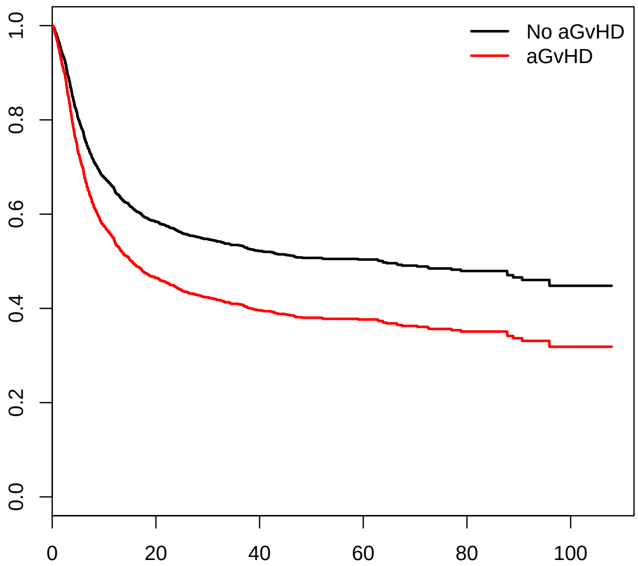

Figure 7: Model-based survival curves for aGvHD and no aGvHD

```
> plot(sfatd,col=1:2,lwd=2,mark.time=FALSE)
> legend("topright",c("No aGvHD","aGvHD"),lwd=2,col=1:2,bty="n")
```

Exercise 14 — Check the proportional hazards assumption. Is it satisfied? Show estimated survival curves with and without aGvHD without using the Cox model.

Answer — The proportional hazards assumption does not seem to be satisfied. The non model-based survival curves are shown in Figure 8. A thought-provoking plot of the violation of the PH assumption is shown in Figure 9.

```
> cox.zph(catd)

        rho chisq      p
tcov -0.146  21.8 3e-06

> sfa <- survfit(Surv(Tstart,Tstop,status) ~ tcov, data=cmltd)
> plot(sfa,col=1:2,lwd=2,mark.time=FALSE)
> legend("topright",c("No aGvHD","aGvHD"),lwd=2,col=1:2,bty="n")

> plot(cox.zph(catd))
```

Exercise 15 — An alternative is to perform a landmark analysis. Construct two groups based on whether or not aGvHD occurred before 100 days. How large

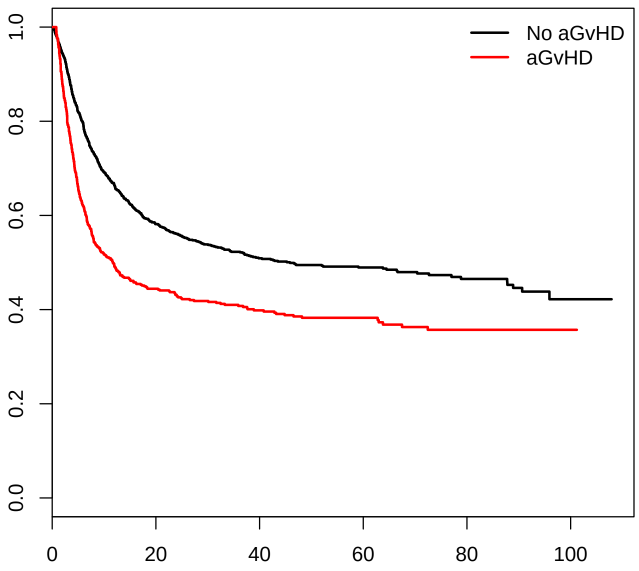

Figure 8: Survival curves for aGvHD and no aGvHD

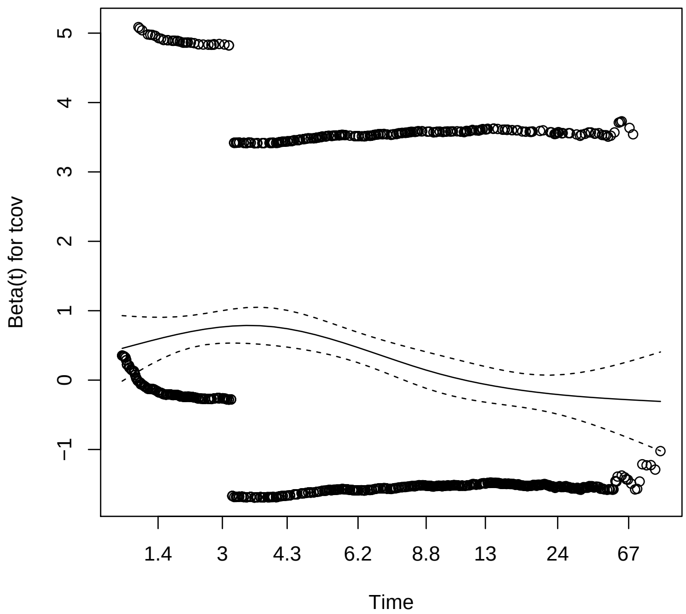

Figure 9: Residuals plot for the time-dependent Cox model

are these groups? Compare with the group sizes in Exercise 10 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 10 and 12.

Answer — The landmark analysis requires first to make the appropriate selection of the data, second to assign the grouping variable.

```
> LM <- 100/(365.25/12)
> LMdata <- cml[cml$rfs > LM,]
> LMdata$group <- 0
> LMdata$group[LMdata$agvhstat==1 & LMdata$agvh<=LM] <- 1
> table(LMdata$group)
```

```
   0    1
1310  577
```

```
> survdiff(Surv(rfs,rfsstat) ~ group, data=LMdata)
```

```
Call:
survdiff(formula = Surv(rfs, rfsstat) ~ group, data = LMdata)

           N Observed Expected (O-E)^2/E (O-E)^2/V
group=0 1310      497      539      3.25      11.3
group=1  577      260      218      8.03      11.3

 Chisq= 11.3  on 1 degrees of freedom, p= 0.000776
```

```
> cLM <- coxph(Surv(rfs,rfsstat) ~ group, data=LMdata)
> print(summary(cLM))
```

```
Call:
coxph(formula = Surv(rfs, rfsstat) ~ group, data = LMdata)

  n= 1887, number of events= 757

         coef exp(coef) se(coef)     z Pr(>|z|)
group 0.25679   1.29278  0.07657 3.354 0.000797 ***
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

      exp(coef) exp(-coef) lower .95 upper .95
group     1.293     0.7735     1.113     1.502

Concordance= 0.537  (se = 0.009 )
Rsquare= 0.006   (max possible= 0.997 )
Likelihood ratio test= 10.93  on 1 df,   p=0.0009482
Wald test            = 11.25  on 1 df,   p=0.0007975
Score (logrank) test = 11.31  on 1 df,   p=0.0007714
```

Exercise 16 — Make Kaplan-Meier survival curves for these two groups, again using the same subset of the data and the same grouping variable. Would you

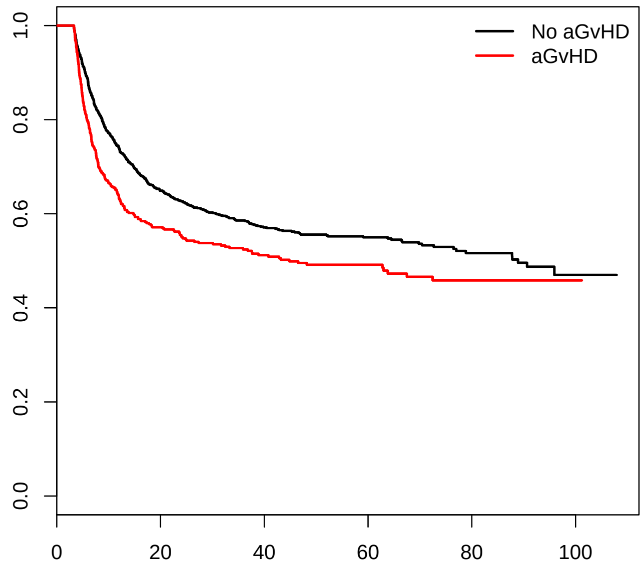

Figure 10: Survival curves for aGvHD and no aGvHD in the landmark data

say, judging from the figure, that the proportional hazards assumption is satisfied? Perform a formal test for the proportional hazards assumption.

Answer — The plot is in Figure 10. The test is performed first.

```
> cox.zph(cLM)
```

```
         rho chisq        p
group -0.135  13.8 0.000207
```

```
> sfLM <- survfit(Surv(rfs,rfsstat) ~ group, data=LMdata)
> plot(sfLM,col=1:2,lwd=2,mark.time=FALSE)
> legend("topright",c("No aGvHD","aGvHD"),lwd=2,col=1:2,bty="n")
```

Exercise 17 — 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.

Answer — Based on the results of Exercise 16 we will use aGvHD as a stratifying variable. The model becomes:

```
> c2LM <- coxph(Surv(rfs,rfsstat) ~ agecl4 + ditrcl4 + year + strata(group), data=LMdata)
> summary(c2LM)
```

```
Call:
coxph(formula = Surv(rfs, rfsstat) ~ agecl4 + ditrcl4 + year +
    strata(group), data = LMdata)

  n= 1887, number of events= 757

                         coef  exp(coef)   se(coef)      z Pr(>|z|)
agecl420-30 years  -0.1139501  0.8923025  0.2609945 -0.437   0.6624
agecl431-50 years   0.1227706  1.1306250  0.2488417  0.493   0.6218
agecl4> 50 years    0.4640400  1.5904867  0.2555391  1.816   0.0694 .
ditrcl43-6 months   0.1935968  1.2136069  0.2607356  0.743   0.4578
ditrcl46-12 months  0.2959674  1.3444263  0.2522098  1.173   0.2406
ditrcl4> 12 months  0.4201535  1.5221951  0.2489563  1.688   0.0915 .
year               -0.0002405  0.9997595  0.0189474 -0.013   0.9899
---
Signif. codes:  0 ï■¡***ï■¡ 0.001 ï■¡**ï■¡ 0.01 ï■¡*ï■¡ 0.05 ï■¡.ï■¡ 0.1 ï■¡ ï■¡ 1

                   exp(coef) exp(-coef) lower .95 upper .95
agecl420-30 years     0.8923     1.1207    0.5350     1.488
agecl431-50 years     1.1306     0.8845    0.6942     1.841
agecl4> 50 years      1.5905     0.6287    0.9639     2.624
ditrcl43-6 months     1.2136     0.8240    0.7280     2.023
ditrcl46-12 months    1.3444     0.7438    0.8201     2.204
ditrcl4> 12 months    1.5222     0.6569    0.9345     2.480
year                  0.9998     1.0002    0.9633     1.038

Concordance= 0.559  (se = 0.015 )
Rsquare= 0.021   (max possible= 0.994 )
Likelihood ratio test= 40.17  on 7 df,   p=1.17e-06
Wald test            = 40.74  on 7 df,   p=9.095e-07
Score (logrank) test = 41.47  on 7 df,   p=6.566e-07
```