Population-based cancer survival: measures and non-parametric estimation

Facilitators: B. Rachet and A. Belot

Acknowledgement: We adapted the contents of a practical which was originally written by Maja Pohar Perme.

The goal of this exercise is to get acquainted with the use of relative survival in R. Load the libraries 'survival' and 'relsurv'.

```
> library(survival)
> library(relsurv)
> data(colrec)
> source("frpop.r")
```

1 The data

We have two sets of data:

- The data on 5971 patients diagnosed with colon or rectum cancer between January 1st, 1994 and December 31st, 2000. Data set `colrec`
- The French population mortality tables. Data set *frpop*

1.1 Import and inspect the colrec dataset

In this session we will use the colrec dataset that contains anonymised data on 5971 patients diagnosed with colon or rectum cancer between January 1st, 1994 and December 31st, 2000. The original dataset is integrated in the relsurv package (object `colrec`).

Task 1: Check the data:

(i) Check the distribution of age and sex.
Hint: age is reported in days (use 365.241 for changes between days and years), sex: 1=male, 2=female

(ii) Find the first and last date of diagnosis.
Hint: use `as.Date` function to get dates

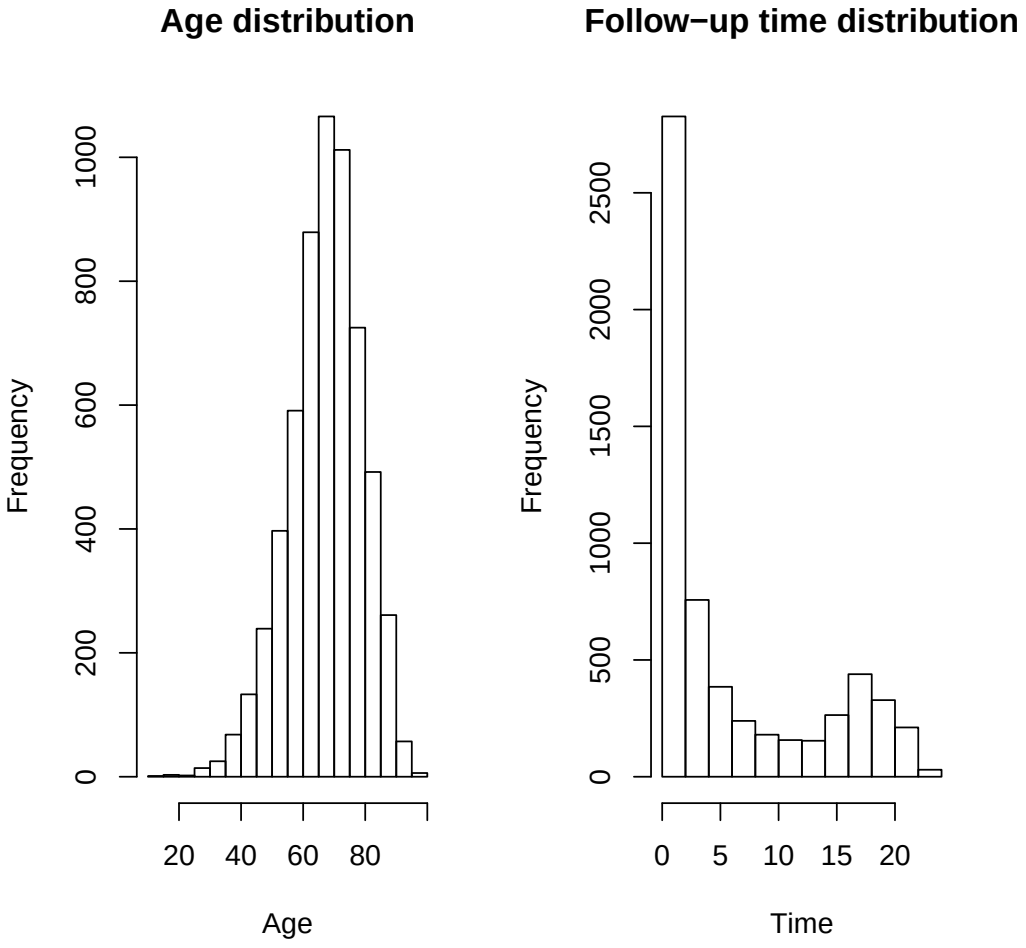
(iii) Check the distribution of follow-up times.
Hint: reported in days

(iv) How many patients have died during the follow-up time?
Hint: variable `stat`

1.2 Understand the mortality tables structure

In our analysis, we will assume that all patients come from France, thus we will use the French population tables.

```
> #put two graphs on the same figure:
> par(mfrow=c(1,2))
> #(i) age range:
> hist(colrec$age/365.241,main="Age distribution",xlab="Age")
> range(colrec$age/365.241)
[1] 12.48217 96.71696
> #sex distribution
> table(colrec$sex)
   1    2
3289 2682
> #(ii) year of diagnosis range
> range(colrec$diag)
[1] 1Jan94    30Dec2000
> hist(colrec$time/365.241, main="Follow-up time distribution",xlab="Time")
> #(iii) follow-up times range
> range(colrec$time/365.241)
[1]  0.002737918 22.308557911
> #number of deaths
> sum(colrec$stat)
[1] 4979
```



The population tables can be organized in various ways (e.g. number of demographical variables, cutpoints), the survival package object `ratetable` has been provided to help in working with the different formats.

The tables for this practical have already been put into the `ratetable` format, so you can source them into R.

For future reference, we describe how the tables have been obtained:

- For many purposes, the population tables can be downloaed from the web. A most useful source is the data from the Human Mortality Database (HMD, http://www.mortality.org), where one can freely download such data (only registration is required). This database contains population tables from many different countries in a uniform format suitable for 'relsurv' (choose period life tables 1x1 for men and women separately).
- The downloaded files can be merged into the `ratetable` object using the function `transrate.hmd` provided by the `relsurv` package.

```
> frpop <- transrate.hmd("mltper_1x1.txt","fltper_1x1.txt")
```

The obtained ratetable is a three-dimensional array that contains the daily hazards for every combination of a person's age, year and sex.

Task 2: Examine the population table object:

- According to which variables are the tables split?
  Hint: `attributes(frpop), dimid`
- Find the dimensions of the object.
  Hint: `attributes(frpop), dim`
- What is the span of calendar years covered?
  Hint: `attributes(frpop), cutpoints`
- Find the value of the daily hazard for individuals aged 80 in 2003
- Estimate the probability of surviving from 80 to 81 for a man aged 80 in 2003.

```
> #attributes(frpop)
> attributes(frpop)$dimid

[1] "age"  "year" "sex"

> attributes(frpop)$dim

[1] 111 201   2

> frpop["80", "2003",]

Rate table with dimension(s): sex
sex
        male       female
0.0001932516 0.0001049879

> exp(-frpop["80", "2003", "male"]*365)

[1] 0.9318934
```

2 Overall and expected survival

Task 3:

(i) Plot overall survival.

(ii) Estimate 5 and 10-year overall survival.

(iii) Plot expected survival on the same graph.

(iv) Estimate 10-year expected survival in the population. Interpret the results.

Note that the demographic varaibles have to be organized to match the ratetable object format and names. In the frpop, there are three dimensions:

- age is named 'age', reported in days
- calendar year is named 'year', reported in the date format
- sex is named 'sex', the first category is 'male', the second 'female'.

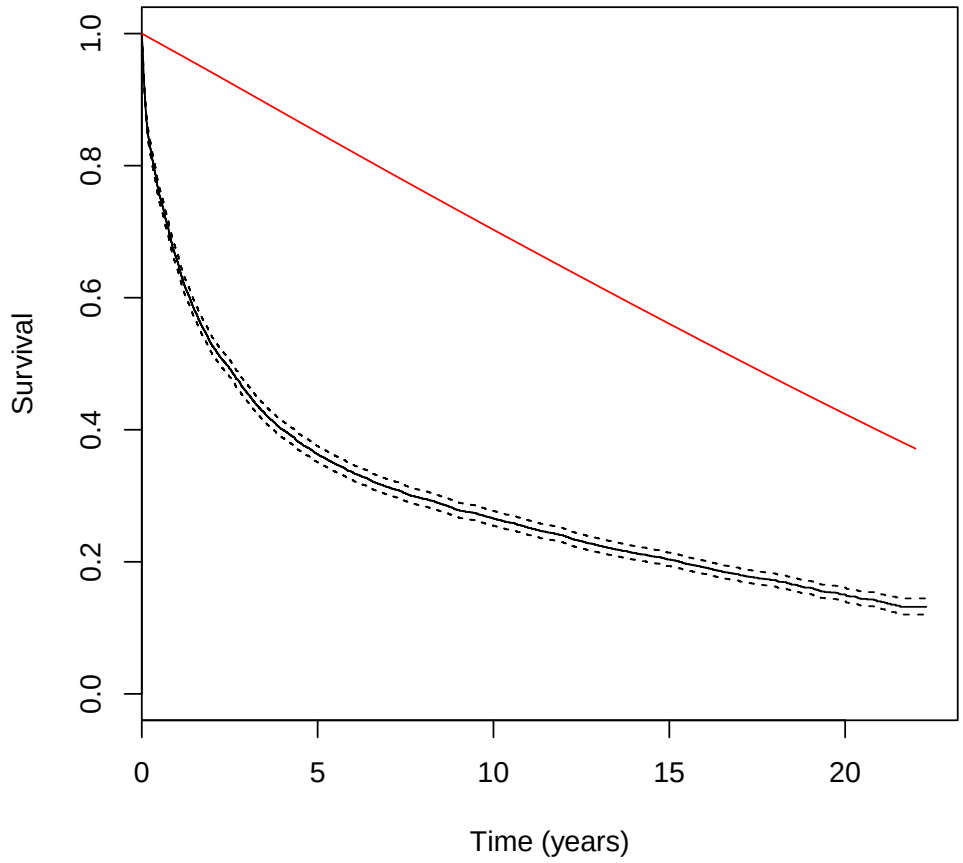
Hint: there are two options for the consolidation of the variables:

- *change the observed data set to match the ratetable*
- *use the `rmap` argument*

Hint: use the xscale-argument in plot.survfit to convert the time unit from days to years.

We see that the first couple of years are critical for patients (many deaths occur), but from the fifth year onwards the hazard decreases. The 10-year overall survival is 0.27.

```
> #(i)
> overall_surv <- survfit(Surv(time, stat)~1, data = colrec)
> plot(overall_surv, xlab = "Time (years)", ylab = "Survival", xscale = 365.241)
> #(ii)
> summary(overall_surv, times=c(5,10)*365.241)
Call: survfit(formula = Surv(time, stat) ~ 1, data = colrec)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1826   2163    3803    0.363 0.00622        0.351        0.375
 3652   1583     580    0.265 0.00572        0.254        0.277
> #(iii)
> exp.surv <- survexp( ~ 1, rmap=list(sex=sex, year=diag,age=age),times=(0:22)*365.241,
+      data=colrec, ratetable=frpop)
> lines(exp.surv, col=2)
> #(iv)
> summary(exp.surv,times=c(5,10)*365.241)
Call: survexp(formula = ~1, data = colrec, rmap = list(sex = sex, year = diag,
    age = age), times = (0:22) * 365.241, ratetable = frpop)

 time n.risk survival
 1826   5971    0.851
 3652   5971    0.703
```



3 Estimating crude and net survival

We are interested in 5-year survival, we will estimate 5-year net and crude survival.

Task 4:

(i) Limit the follow-up time to 5 years (censor all individuals after 5 years)

(ii) Estimate 5-year crude mortality, interpret your results

(iii) Estimate 5-year net survival, interpret your results

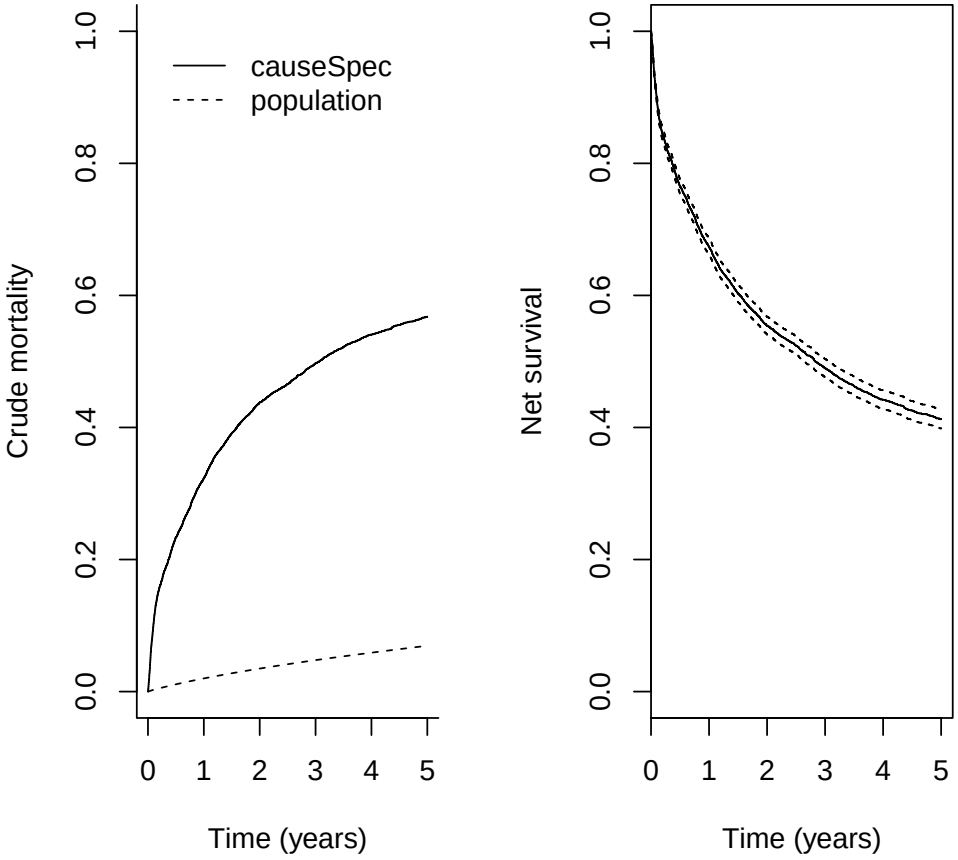
```
> #(i) Censor all individuals after 5 years of follow-up
> colrec$time5 <- pmin(colrec$time,5*365.241)     #limit to 5 years
> colrec$stat5 <- ifelse(colrec$time<5*365.241,colrec$stat,0) #censor those with longer follow
> #(ii)
> cru <- cmp.rel(Surv(time5, stat5) ~ 1, rmap=list(age = age, sex = sex, year = diag),
+                      data = colrec, ratetable = frpop)
> summary(cru,times=5*365.241)

$est
             1826.205
causeSpec  0.56797051
population 0.06941154

$var
               1826.205
causeSpec  4.671064e-05
population 4.910266e-07

> #(iii) Estimate 5-year net survival
> net_surv <- rs.surv(Surv(time5, stat5) ~ 1, rmap=list(age = age, sex = sex,
+                                                       year = diag),
+                      data = colrec, ratetable = frpop, method = "pohar-perme")
> summary(net_surv, times = 5*365.241)

Call: rs.surv(formula = Surv(time5, stat5) ~ 1, data = colrec, ratetable = frpop,
    method = "pohar-perme", rmap = list(age = age, sex = sex,
        year = diag))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1826   2163    3803    0.412 0.00728        0.398        0.427
```

Interpretation for crude mortality:
We already know that 0.64 patients died. 0.57 died due to cancer and 0.07 died due to other causes.

Two interpretations for net survival:

- On average, the ratio between the patient survival and that of their counterparts is 0.41
- In a hypothetical world, where patients could die of cancer only, 41 % of patients would still be alive after 5 years.

```
> #put two graphs on the same figure:
> par(mfrow=c(1,2))
> plot(cru, xlab = "Time (years)", ylab = "Crude mortality", xscale = 365.241)
> plot(net_surv, xlab = "Time (years)", ylab = "Net survival", xscale = 365.241)
```

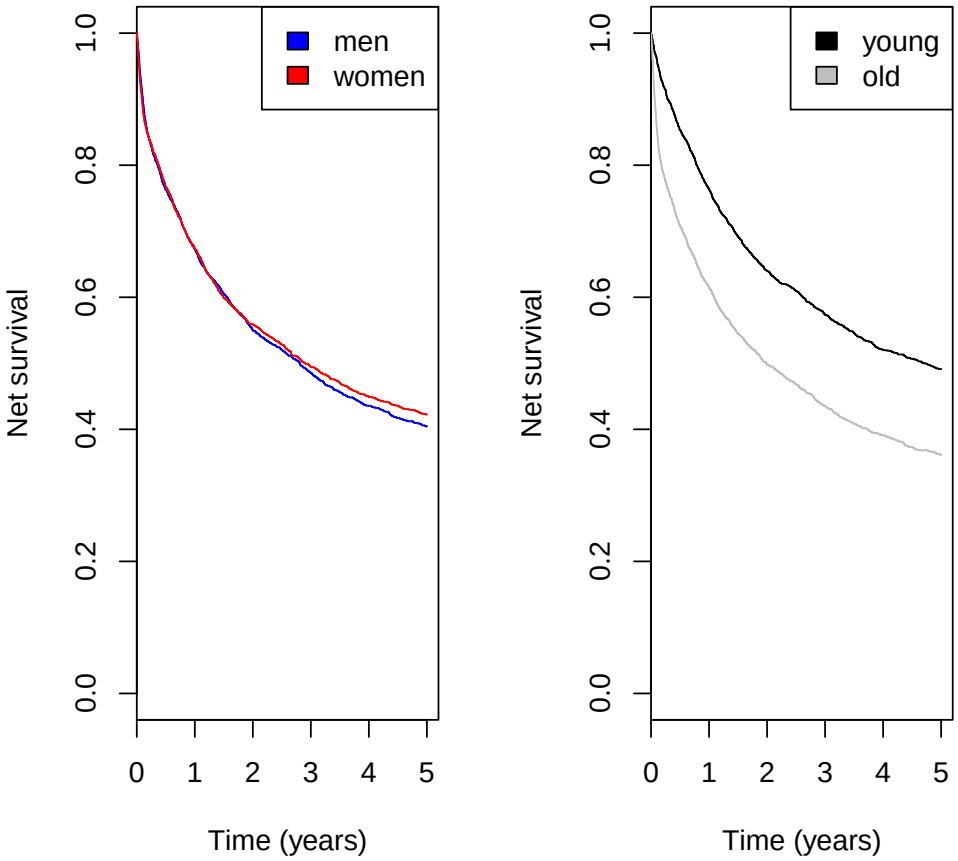


4 Net survival with respect to covariates

Task 5: Compare the 5-year net survival with respect to sex and age. To this end, split individuals to below and above 65 years.
Hint: use the function `rs.diff`.

```
> colrec$age_over65 <- ifelse(colrec$age<=65*365.241,0,1)
> net_surv_sex <- rs.surv(Surv(time5, stat5) ~ sex, rmap=list(age = age, sex = sex,
+                 year = diag),data = colrec, ratetable = frpop, method = "pohar-perme")
> summary(net_surv_sex,times=5*365.241)

Call: rs.surv(formula = Surv(time5, stat5) ~ sex, data = colrec, ratetable = frpop,
    method = "pohar-perme", rmap = list(age = age, sex = sex,
        year = diag))

               sex=1
        time       n.risk      n.event     survival      std.err lower 95% CI
    1.83e+03     1.14e+03     2.15e+03     4.04e-01     9.98e-03     3.85e-01
upper 95% CI
    4.24e-01

               sex=2
        time       n.risk      n.event     survival      std.err lower 95% CI
    1.83e+03     1.02e+03     1.66e+03     4.22e-01     1.06e-02     4.02e-01
upper 95% CI
    4.44e-01

> rs.diff(Surv(time5, stat5) ~ sex, rmap=list(age = age, sex = sex,
+                     year = diag),data = colrec, ratetable = frpop)

Value of test statistic: 0.8642386
Degrees of freedom: 1
P value: 0.3525553

> net_surv_age <- rs.surv(Surv(time5, stat5) ~ age_over65, rmap=list(age = age,
+                     sex = sex, year = diag),
+                     data = colrec, ratetable = frpop, method = "pohar-perme")
> summary(net_surv_age,times=5*365.241)

Call: rs.surv(formula = Surv(time5, stat5) ~ age_over65, data = colrec,
    ratetable = frpop, method = "pohar-perme", rmap = list(age = age,
        sex = sex, year = diag))

               age_over65=0
        time       n.risk      n.event     survival      std.err lower 95% CI
    1.83e+03     1.10e+03     1.24e+03     4.91e-01     1.08e-02     4.70e-01
upper 95% CI
    5.13e-01

               age_over65=1
        time       n.risk      n.event     survival      std.err lower 95% CI
    1.83e+03     1.06e+03     2.56e+03     3.61e-01     9.61e-03     3.43e-01
upper 95% CI
    3.80e-01

> rs.diff(Surv(time5, stat5) ~ age_over65, rmap=list(age = age, sex = sex,
+                                                    year = diag),
+                     data = colrec, ratetable = frpop)
```

```
> #put two graphs on the same figure:
> par(mfrow=c(1,2))
> plot(net_surv_sex, xlab = "Time (years)", ylab = "Net survival",
+      xscale = 365.241,col=c(4,2))
> legend("topright",fill=c(4,2),legend=c("men","women"))
> plot(net_surv_age, xlab = "Time (years)", ylab = "Net survival",
+          xscale = 365.241,col=c(1,"grey"))
> legend("topright",fill=c(1,"grey"),legend=c("young","old"))
```




```
Value of test statistic: 125.5004
Degrees of freedom: 1
P value: 0
```

No difference can be observed between men and women. On the other hand, age seems to be a very important factor, with younger patients having a lower excess hazard and hence a better net survival.

5 Longer follow-up times

Task 6:

(i) Check the amount of information in your data for individuals aged over 80 after 10 or 15 years of follow-up.
Hint: use the function `nessie`

(ii) Plot 15-year net survival for a sensible subgroup of patients.

```
> #(i)
> breaks <- c(0, seq(from = 45, to = 90, by = 5), Inf)
> colrec$agegr <- cut(colrec$age / 365.241, breaks)
> nessie(Surv(time, stat) ~ agegr+sex, data = colrec, ratetable = frpop,
+         times = c(0,2,5,10,15), rmap = list(age = age, sex = sex, year = diag))

                      0     2     5    10    15 c.exp.surv
agegr(0,45],sex=1   135 134.2 133.0 130.2 126.7       41.3
agegr(0,45],sex=2   111 110.7 110.2 109.1 107.7       46.5
agegr(45,50],sex=1  141 139.5 137.0 131.8 125.2       33.0
agegr(45,50],sex=2   98  97.6  96.8  95.2  93.1       38.4
agegr(50,55],sex=1  239 235.5 229.4 217.4 202.7       28.7
agegr(50,55],sex=2  158 157.0 155.3 151.9 147.3       34.1
agegr(55,60],sex=1  368 360.2 347.1 321.6 291.2       24.5
agegr(55,60],sex=2  223 221.1 217.8 211.0 201.7       29.6
agegr(60,65],sex=1  547 529.4 500.6 446.2 381.4       20.2
agegr(60,65],sex=2  332 327.8 320.4 304.4 281.5       24.8
agegr(65,70],sex=1  635 604.7 555.2 463.0 355.2       16.2
agegr(65,70],sex=2  431 422.3 406.9 372.4 320.6       20.2
agegr(70,75],sex=1  543 503.8 440.5 324.1 200.9       12.4
agegr(70,75],sex=2  469 453.3 424.7 359.1 265.5       15.9
agegr(75,80],sex=1  340 302.4 242.9 144.8  64.0        9.3
agegr(75,80],sex=2  385 362.2 320.3 232.1 131.0       12.0
agegr(80,85],sex=1  207 167.5 112.3  43.8  10.6        6.4
agegr(80,85],sex=2  285 251.1 194.3 101.0  34.4        8.3
agegr(85,90],sex=1  110  78.0  40.9   9.5   1.2        4.5
agegr(85,90],sex=2  151 120.6  77.3  26.2   4.7        5.9
agegr(90,Inf],sex=1  24  13.6   4.9   0.5   0.0        3.1
agegr(90,Inf],sex=2  39  25.1  10.9   1.7   0.1        3.7
```

```
> net_surv_sex <- nessie(Surv(time5, stat5) ~ sex, rmap=list(age = age, sex = sex,
+                     year = diag), data = colrec, ratetable = frpop)

         0      1      2      3      4      5 c.exp.surv
sex=1 3289 3178.7 3068.8 2959.4 2850.8 2743.7       18.0
sex=2 2682 2616.5 2548.9 2479.0 2407.4 2334.9       19.9
```

We can observe that:

- For individuals above 90 (or the total sample containing them) even the 5-year net survival is a stretch.
- If 10-year net survival is of interest, one should limit to individuals under 90
- If 15-year net survival is of interest, one should limit to individuals under 85.

Note that: if a very small proportion of individuals from a certain subgroup remains, they get a very large weight. If no individuals from a certain subgroup remain at risk, their values do not get represented in the average (they are taken to be equal to average)

```
> #(ii)
> colrec$time15 <- pmin(colrec$time,15*365.241)     #limit to 15 years
> #censor those with longer follow-up times
> colrec$stat15 <- ifelse(colrec$time<15*365.241,colrec$stat,0)
> net_surv <- rs.surv(Surv(time15, stat15) ~ 1 , rmap=list(age = age, sex = sex,
+                       year = diag), data = colrec[colrec$age <= 80*365.241,],
+                       ratetable = frpop,method = "pohar-perme")
> plot(net_surv, xlab = "Time (years)", ylab = "Net survival", xscale = 365.241)
```