Epidemiologic data analysis using R

Practicals 5

Faculty of Social Sciences, University of Tampere

–

Janne Pitkäniemi / Karri Seppä

2.3.2020

# Topics of practical 5

Learning objectives of this practical

* drawing life-lines of follow-up on members of a cohort using function Lexis.diagram() in Epi,
* splitting follow-up times simultaeously by ageband and calendar period, and expanding the data frame by the Lexis tools in Epi
* merging two dataframes into one data frame using function merge()
* computation of expected numbers of cases based on age- and period-specific rates, and comparison of the observed numbers of cases in a target cohort.

In this practical we consider time-splitting simultaneously in several time scales. Splitting the follow-up times of a special occupational or other cohort of interest jointly by age and calendar time is needed for estimation of the relative rate of the outcome event considered in the cohort as compared to a relevant reference population. Often the latter is a national general population, of which age-specific rates in different calendar periods are available from official statistics. Adjustment for age and calendar time in the relative rate estimation is then performed by the principle called indirect standardization, which provides the standardized incidence ratio SIR or standardized mortality ratio SMR as the main result. SIR equals the observed number of cases in the cohort divided by the expected number, the latter being estimated from the age-period specific reference rates and person-years in the cohort. We illustrate the use of the available tools in Epi by which these computations can be done in R.

**The Danish Thorotrast study**

In the period 1935-50 a contrast medium called Thorotrast was used for cerebral angiography (X-ray imaging of the brain). This contrast medium contained 232Th, thorium. It turns out that thorium is not excreted from the body, it is permanently deposited, some 60% in the liver, 20% in the spleen and some 10% in the bone marrow, and a very small fraction in other organs.

Thorium is an -emitting radionuclide, i.e. it emits -rays (He-nuclei) which is ionizing, but not particularly penetrating; it only penetrates 2-3 cell-layers. The half-life of 232Th is 1.4 1010 years, so the patients that have been injected with Thorotrast exposed are exposed to a constant,small -radiation for life.

The study cohort includes 999 Thorotrast patients who had a cerebral angiography in the period 1935-50. In addition there is a control group comprising 1480 reference subjects who had a cerebral angiography in a later but somewhat overlapping period, 1946-63, on similar indications as the Thorotrast patients, but with another contrast medium, not containg thorium. Persons undergoing cerebral angiography are in may cases seriously ill, they are suspected of cerebral malformations or tumours. Hence, both the Thorotrast group and the control group have very high mortality rates, and a pattern of causes of death that differs substantially from the general population. Especially during the first year after diagnosis there is a very high mortality among the patients, which is largely associated to the conditions that have lead to the cerebral angiography.

Therefore we start the follow-up of both Thorotrast patients and control patients one year after the angiography.

The cohort data are in the file thoro in the Epi package. Mortality rates (per 1000 personyears) for Denmark, by sex and 5 year ageband (0-4, 5-9, . . . , 90-95), for each of the 5-year calendar periods 1938-42, 1943-47, 1948-52, . . . 1988-92, are in the file gmortDK. Besides the overall mortality (rt), this file also contains the mortality rates for 15 different causes of death. Total mortality in the cohort by contrast medium In this exercise our outcome event will be death from any cause. The occurrence of death during the follow-up can be derived from value 1 of variable exitstat.

## 1. The data

Load the Epi package and the data frame thoro, and see its contents:

library(Epi)  
data( thoro )  
str( thoro )

'data.frame': 2470 obs. of 14 variables:  
 $ id : num 1 2 3 4 5 6 7 8 9 10 ...  
 $ sex : num 2 2 1 1 1 2 1 2 1 1 ...  
 $ birthdat: Date, format: "1916-08-11" "1927-11-05" ...  
 $ contrast: num 1 1 1 1 1 1 1 1 1 1 ...  
 $ injecdat: Date, format: "1938-10-17" "1943-11-28" ...  
 $ volume : num 22 80 10 10 10 20 10 40 34 10 ...  
 $ exitdat : Date, format: "1976-10-15" "1966-01-12" ...  
 $ exitstat: num 1 1 1 1 1 1 1 3 1 1 ...  
 $ cause : num 2 8 2 2 14 14 3 NA 2 2 ...  
 $ liverdat: Date, format: "1966-01-12" "1966-01-12" ...  
 $ liver : num 1 1 0 1 0 0 0 0 1 0 ...  
 $ hepcc : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ chola : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ hmang : num 1 1 0 1 0 0 0 0 1 0 ...

The time variables are internally stored in the Dates format, i.e. days since 1/1/1970, but are formatted to print out as dates. To see this try:

head( thoro )

id sex birthdat contrast injecdat volume exitdat exitstat cause  
1 1 2 1916-08-11 1 1938-10-17 22 1976-10-15 1 2  
2 2 2 1927-11-05 1 1943-11-28 80 1966-01-12 1 8  
3 3 1 1902-10-12 1 1935-08-19 10 1959-09-21 1 2  
4 4 1 1918-05-12 1 1936-05-25 10 1977-04-23 1 2  
5 5 1 1902-12-07 1 1937-05-22 10 1945-05-22 1 14  
6 6 2 1903-09-19 1 1937-04-26 20 1944-09-27 1 14  
 liverdat liver hepcc chola hmang  
1 1966-01-12 1 0 0 1  
2 1966-01-12 1 0 0 1  
3 <NA> 0 0 0 0  
4 1977-04-23 1 0 0 1  
5 <NA> 0 0 0 0  
6 <NA> 0 0 0 0

thoro$birthdat[1:6]

[1] "1916-08-11" "1927-11-05" "1902-10-12" "1918-05-12" "1902-12-07"  
[6] "1903-09-19"

as.numeric(thoro$birthdat)[1:6]

[1] -19501 -15398 -24553 -18862 -24497 -24211

## 2. Date formatting

In the following it will be more convenient to have dates as fractional calendar years as produced by function cal.yr(). Likewise it is more convenient to have the outcome variable named as death and coded 1 for dead and 0 for censored:

thoro <- transform( thoro, dob = cal.yr( birthdat ),  
doi = cal.yr( injecdat )+1,  
dox = cal.yr( exitdat ),  
dol = cal.yr( liverdat ),  
death = as.numeric( exitstat==1 ) )

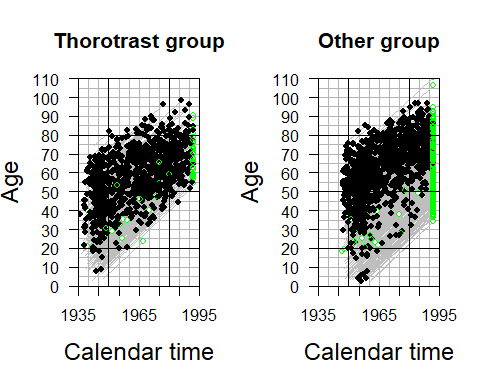
Assign individual person-times to variable y, compute the ages at entry and exit, and convert contrast variable into a factor changing the order of its levels, too. Look at univariate summaries of all variables:

thoro <- transform( thoro, y = dox - doi,  
agen = doi - dob, agex = dox - dob,  
contrast = factor(2-contrast,  
labels = c("Other", "Thoro") ) )  
summary( thoro )

id sex birthdat contrast   
 Min. : 1.0 Min. :1.000 Min. :1868-01-07 Other:1480   
 1st Qu.: 710.2 1st Qu.:1.000 1st Qu.:1897-08-26 Thoro: 990   
 Median :3245.5 Median :1.000 Median :1908-02-25   
 Mean :2474.6 Mean :1.477 Mean :1909-10-14   
 3rd Qu.:3867.8 3rd Qu.:2.000 3rd Qu.:1920-10-24   
 Max. :4494.0 Max. :2.000 Max. :1958-02-01   
   
 injecdat volume exitdat exitstat   
 Min. :1935-08-13 Min. : 0.000 Min. :1935-09-01 Min. :1.00   
 1st Qu.:1944-10-09 1st Qu.: 0.000 1st Qu.:1955-11-11 1st Qu.:1.00   
 Median :1948-11-22 Median : 0.000 Median :1970-01-11 Median :1.00   
 Mean :1949-02-01 Mean : 7.492 Mean :1970-02-11 Mean :1.22   
 3rd Qu.:1953-08-23 3rd Qu.:10.000 3rd Qu.:1986-09-23 3rd Qu.:1.00   
 Max. :1963-08-20 Max. :80.000 Max. :1992-02-20 Max. :3.00   
   
 cause liverdat liver hepcc   
 Min. : 1.000 Min. :1961-05-19 Min. :0.00000 Min. :0.00000   
 1st Qu.: 2.000 1st Qu.:1971-02-24 1st Qu.:0.00000 1st Qu.:0.00000   
 Median : 7.000 Median :1978-02-17 Median :0.00000 Median :0.00000   
 Mean : 6.412 Mean :1977-09-02 Mean :0.05263 Mean :0.01781   
 3rd Qu.: 8.000 3rd Qu.:1983-08-19 3rd Qu.:0.00000 3rd Qu.:0.00000   
 Max. :16.000 Max. :1990-07-18 Max. :1.00000 Max. :1.00000   
 NA's :504 NA's :2340   
 chola hmang dob doi   
 Min. :0.00000 Min. :0.00000 Min. :1868 Min. :1937   
 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:1898 1st Qu.:1946   
 Median :0.00000 Median :0.00000 Median :1908 Median :1950   
 Mean :0.01579 Mean :0.01336 Mean :1910 Mean :1950   
 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1921 3rd Qu.:1955   
 Max. :1.00000 Max. :1.00000 Max. :1958 Max. :1965   
   
 dox dol death y   
 Min. :1936 Min. :1961 Min. :0.000 Min. :-1.000   
 1st Qu.:1956 1st Qu.:1971 1st Qu.:1.000 1st Qu.: 2.962   
 Median :1970 Median :1978 Median :1.000 Median :21.031   
 Mean :1970 Mean :1978 Mean :0.796 Mean :20.026   
 3rd Qu.:1987 3rd Qu.:1984 3rd Qu.:1.000 3rd Qu.:34.480   
 Max. :1992 Max. :1991 Max. :1.000 Max. :52.988   
 NA's :2340   
 agen agex   
 Min. : 1.454 Min. : 2.174   
 1st Qu.:28.379 1st Qu.: 50.591   
 Median :41.355 Median : 61.431   
 Mean :40.302 Mean : 60.328   
 3rd Qu.:52.665 3rd Qu.: 71.437   
 Max. :80.179 Max. :110.637

1. Draw parallel Lexis-diagrams to describe the follow-up lifelines for both contrast groups separately, first for the Thoro group:

par(mfrow=c(1,2))  
Lexis.diagram( age = c(0,110), date = c(1935,1995),  
entry.date = doi, exit.date = dox, birth.date = dob,  
fail = 1\*(exitstat==1), cex.axis = 1.5, cex.lab = 1.5,  
lwd.life = 0.5, col.life = "gray",  
cex.fail = 0.8, pch.fail = c(1,16),  
col.fail = c("green","black"),  
main = "Thorotrast group",  
cex.main = 1.3,  
data = subset(thoro, contrast=="Thoro") )  
box()  
abline( v = c(1950, 1980) )  
abline( h = c(20, 50, 80) )  
  
Lexis.diagram( age = c(0,110), date = c(1935,1995),  
entry.date = doi, exit.date = dox, birth.date = dob,  
fail = 1\*(exitstat==1), cex.axis = 1.5, cex.lab = 1.5,  
lwd.life = 0.5, col.life = "gray",  
cex.fail = 0.8, pch.fail = c(1,16),  
col.fail = c("green","black"),  
main = "Other group",  
cex.main = 1.3,  
data = subset(thoro, contrast=="Other") )  
box()  
abline( v = c(1950, 1980) )  
abline( h = c(20, 50, 80) )



Draw a similar diagram for the Other group by appropriate subsetting.

## 3. Lexis object

Create a Lexis object from the data frame and check the content of its 20 first lines

thL <- Lexis( entry = list( per = doi ),  
 exit = list( per = dox,   
 age = dox - dob ),  
 exit.status = 1\*( exitstat<2 ),   
 data = thoro)

NOTE: entry.status has been set to 0 for all.  
NOTE: Dropping 427 rows with duration of follow up < tol

thL[1:20, 1:10 ]

per age lex.dur lex.Cst lex.Xst lex.id id sex birthdat  
1 1939.791 23.18207 36.99589322 0 1 1 1 2 1916-08-11  
2 1944.906 17.06297 21.12457221 0 1 2 2 2 1927-11-05  
3 1936.629 33.85147 23.09034908 0 1 3 3 1 1902-10-12  
4 1937.396 19.03696 39.91170431 0 1 4 4 1 1918-05-12  
5 1938.387 35.45585 7.00000000 0 1 5 5 1 1902-12-07  
6 1938.316 34.60164 6.42231348 0 1 6 6 2 1903-09-19  
7 1938.261 28.68515 30.18412047 0 1 7 7 1 1909-07-30  
9 1939.690 28.47707 38.13757700 0 1 9 9 1 1911-03-20  
10 1938.781 39.23682 12.14715948 0 1 10 10 1 1899-07-18  
11 1939.408 14.09788 47.67077344 0 1 11 11 1 1925-04-24  
12 1939.350 20.92882 34.09650924 0 1 12 12 1 1918-06-04  
13 1939.172 27.25051 40.69472964 0 1 13 13 2 1911-12-04  
14 1938.934 31.28063 32.77412731 0 1 14 14 2 1907-08-28  
17 1940.139 56.42231 0.01300479 0 1 17 17 2 1883-09-19  
18 1940.169 37.66530 27.16153320 0 1 18 18 1 1902-07-04  
20 1940.413 39.50787 43.93086927 0 1 20 20 1 1900-11-27  
21 1940.528 36.25530 15.41067762 0 1 21 21 2 1904-04-10  
22 1940.563 35.29979 0.48117728 0 1 22 22 1 1905-04-07  
23 1940.648 28.67967 21.54893908 0 1 23 23 1 1911-12-21  
24 1940.640 27.36824 24.27036277 0 0 24 24 2 1913-04-10  
 contrast  
1 Thoro  
2 Thoro  
3 Thoro  
4 Thoro  
5 Thoro  
6 Thoro  
7 Thoro  
9 Thoro  
10 Thoro  
11 Thoro  
12 Thoro  
13 Thoro  
14 Thoro  
17 Thoro  
18 Thoro  
20 Thoro  
21 Thoro  
22 Thoro  
23 Thoro  
24 Thoro

The first descriptive task is to look at total mortality in the cohort overall and by contrast medium. Use function stat.table():

T1 <- stat.table( index = contrast,  
contents = list( N = count(), # group size  
D = sum(lex.Xst), # no. of deaths  
Y = sum(lex.dur), # person-times  
rate = ratio(lex.Xst, lex.dur, 1000) ), # rate/1000 y  
margin = T, data=thL )  
print(T1, digits = c(sum=0, ratio=1))

-------------------------------------------   
 contrast N D Y rate   
 -------------------------------------------   
 Other 1236 797 30518 26.1   
 Thoro 807 748 19244 38.9   
   
 Total 2043 1545 49761 31.0   
 -------------------------------------------

The follow-up of the two groups of patients takes place in very different time periods, and the two groups have slightly differing age-distributions, too. Therefore it is desirable to control for age and calendar time. This could be done by making an internal comparison of the two contrast groups controlled for age, sex, and calendar period. However, because of the different calendar periods of follow-up, a large portion of information would be lost. Instead, the comparison will be standardized for age, sex, and period, using SMRs.

The Danish mortality figures are in the file gmortDK. In order to be able to match the Danish population mortality rates to the follow-up data these must first be split by current age and calendar time. The names and coding of the age and period variables must be chosen to conform with that in gmortDK. 5. Load the population mortality data and have a look at its contents. You would notice that the calendar periods per are coded as yyyy 1900 where yyyy is the starting year of the period. Therefore, we have to form a new period variable pgr adding constant 1900 to the values of the existing variable:

data(gmortDK)  
gmortDK[c(1:10, 409:418), 1:6 ]

agr per sex risk dt rt  
1 0 38 1 996019 14079 14.135  
2 5 38 1 802334 726 0.905  
3 10 38 1 753017 600 0.797  
4 15 38 1 773393 1167 1.509  
5 20 38 1 813882 2031 2.495  
6 25 38 1 789990 1862 2.357  
7 30 38 1 800394 1973 2.465  
8 35 38 1 769731 2192 2.848  
9 40 38 1 694051 2639 3.802  
10 45 38 1 622257 3480 5.593  
409 45 88 2 600791 2091 3.480  
410 50 88 2 495108 2810 5.676  
411 55 88 2 453716 4239 9.343  
412 60 88 2 453914 6193 13.644  
413 65 88 2 456895 9366 20.499  
414 70 88 2 397399 12408 31.223  
415 75 88 2 348340 17484 50.192  
416 80 88 2 249119 21310 85.541  
417 85 88 2 133690 19870 148.630  
418 90 88 2 57038 16069 281.720

gmortDK$pgr <- gmortDK$per + 1900

## 5. Split the follow-up data

Using the thoro by two timescales: age in 5-year agebands named by their lower cutpoints 0, 5, 10,. . . , 90, and calendar time in 5-year periods divided by cutpoints 1938, 1943,. . . , 1993 and named according to their lower limit years (“left”).

NB! You must make sure that the age-group and the period variables in the new dataset have the same names as in gmortDK, namely agr and per, respectively.

thL.a <- splitLexis( thL, "age", breaks = seq(5, 95, 5) )  
thL.ap <- splitLexis( thL.a, "per", breaks = seq(1938, 1993,5) )  
thL.ap$agr <- timeBand( thL.ap, "age", "left")  
thL.ap$pgr <- timeBand( thL.ap, "per", "left")

How many observations there are in the new data frame? Use function dim(). Take a look at the content of the rows pertaining to id-number 1:

thL.ap[thL.ap$id==1, ]

lex.id per age lex.dur lex.Cst lex.Xst id sex birthdat  
1 1 1939.791 23.18207 1.8179329 0 0 1 2 1916-08-11  
2 1 1941.609 25.00000 1.3908282 0 0 1 2 1916-08-11  
3 1 1943.000 26.39083 3.6091718 0 0 1 2 1916-08-11  
4 1 1946.609 30.00000 1.3908282 0 0 1 2 1916-08-11  
5 1 1948.000 31.39083 3.6091718 0 0 1 2 1916-08-11  
6 1 1951.609 35.00000 1.3908282 0 0 1 2 1916-08-11  
7 1 1953.000 36.39083 3.6091718 0 0 1 2 1916-08-11  
8 1 1956.609 40.00000 1.3908282 0 0 1 2 1916-08-11  
9 1 1958.000 41.39083 3.6091718 0 0 1 2 1916-08-11  
10 1 1961.609 45.00000 1.3908282 0 0 1 2 1916-08-11  
11 1 1963.000 46.39083 3.6091718 0 0 1 2 1916-08-11  
12 1 1966.609 50.00000 1.3908282 0 0 1 2 1916-08-11  
13 1 1968.000 51.39083 3.6091718 0 0 1 2 1916-08-11  
14 1 1971.609 55.00000 1.3908282 0 0 1 2 1916-08-11  
15 1 1973.000 56.39083 3.6091718 0 0 1 2 1916-08-11  
16 1 1976.609 60.00000 0.1779603 0 1 1 2 1916-08-11  
 contrast injecdat volume exitdat exitstat cause liverdat liver hepcc  
1 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
2 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
3 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
4 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
5 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
6 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
7 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
8 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
9 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
10 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
11 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
12 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
13 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
14 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
15 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
16 Thoro 1938-10-17 22 1976-10-15 1 2 1966-01-12 1 0  
 chola hmang dob doi dox dol death y agen  
1 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
2 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
3 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
4 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
5 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
6 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
7 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
8 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
9 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
10 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
11 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
12 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
13 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
14 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
15 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
16 0 1 1916.609 1939.791 1976.787 1966.03 1 36.99589 23.18207  
 agex agr pgr  
1 60.17796 20 1938  
2 60.17796 25 1938  
3 60.17796 25 1943  
4 60.17796 30 1943  
5 60.17796 30 1948  
6 60.17796 35 1948  
7 60.17796 35 1953  
8 60.17796 40 1953  
9 60.17796 40 1958  
10 60.17796 45 1958  
11 60.17796 45 1963  
12 60.17796 50 1963  
13 60.17796 50 1968  
14 60.17796 55 1968  
15 60.17796 55 1973  
16 60.17796 60 1973

Make sure that you understand how these records relate to the original records in thoro.

## 6. SMR

In order to conduct the SMR-calculations, we now match the rate file gmortDK to the data frame expanded from thoro (we only take 4 columns of gmortDK, and skip all the cause-specific rates):

thL.apd <- merge( thL.ap, gmortDK[, c("agr", "pgr", "sex", "rt")],  
by = c("agr", "pgr", "sex") )

## 7. Expected deaths

We add two variables to this expanded data frame: y containing the person-time, and e with the expected number of deaths = the person-years multiplied by the mortality rate from the reference population, and put d to denote the observation of outcome event.

thL.apd <- transform( thL.apd, d = lex.Xst,  
y = lex.dur, e = lex.dur\*(rt/1000) )  
str( thL.apd )

Classes 'Lexis' and 'data.frame': 21748 obs. of 33 variables:  
 $ sex : num 1 1 1 1 1 1 1 1 1 1 ...  
 $ agr : num 10 10 10 10 10 10 10 10 10 10 ...  
 $ pgr : num 1938 1938 1938 1938 1938 ...  
 $ lex.id : int 236 98 11 273 239 97 757 930 549 864 ...  
 $ per : num 1940 1939 1939 1941 1941 ...  
 $ age : num 13.4 11.9 14.1 14.6 10 ...  
 $ lex.dur : num 1.608 3.084 0.902 0.431 1.984 ...  
 $ lex.Cst : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ lex.Xst : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ id : num 262 98 11 301 265 ...  
 $ birthdat: Date, format: "1926-11-11" "1927-07-16" ...  
 $ contrast: Factor w/ 2 levels "Other","Thoro": 2 2 2 2 2 2 2 2 2 2 ...  
 $ injecdat: Date, format: "1939-04-03" "1938-06-15" ...  
 $ volume : num 24 20 10 12 40 18 20 20 10 10 ...  
 $ exitdat : Date, format: "1990-10-05" "1979-01-23" ...  
 $ exitstat: num 1 1 1 3 1 1 2 1 1 2 ...  
 $ cause : num 2 8 2 NA 2 2 NA 2 8 NA ...  
 $ liverdat: Date, format: NA NA ...  
 $ liver : num 0 0 0 0 1 0 0 1 0 0 ...  
 $ hepcc : num 0 0 0 0 0 0 0 1 0 0 ...  
 $ chola : num 0 0 0 0 1 0 0 0 0 0 ...  
 $ hmang : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ dob : num 1927 1928 1925 1926 1931 ...  
 $ doi : num 1940 1939 1939 1941 1940 ...  
 $ dox : num 1991 1979 1987 1974 1979 ...  
 $ dol : num NA NA NA NA 1979 ...  
 $ death : num 1 1 1 0 1 1 0 1 1 0 ...  
 $ y : num 1.608 3.084 0.902 0.431 1.984 ...  
 $ agen : num 13.39 11.92 14.1 14.57 9.41 ...  
 $ agex : num 63.9 51.5 61.8 47.7 47.5 ...  
 $ rt : num 0.797 0.797 0.797 0.797 0.797 0.797 0.797 0.797 0.797 0.797 ...  
 $ d : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ e : num 0.001282 0.002458 0.000719 0.000344 0.001581 ...  
 - attr(\*, "breaks")=List of 2  
 ..$ per: num 1938 1943 1948 1953 1958 ...  
 ..$ age: num 5 10 15 20 25 30 35 40 45 50 ...  
 - attr(\*, "time.scales")= chr "per" "age"  
 - attr(\*, "time.since")= chr "" ""

## 8. Tabulate

Use stat.table() to make a table of SMR (observed/expected) classified by contrast

T2 <- stat.table( contrast,  
list( Y = sum(y), D = sum(d), E = sum(e),  
SMR = ratio(d,e), "rate /1000y" = ratio(d,y,1000) ),  
margin = T, data = thL.apd )  
print(T2, dig = c(sum=1, ratio=2))

---------------------------------------------------   
 contrast Y D E SMR rate   
 /1000y   
 ---------------------------------------------------   
 Other 30459.2 792.0 456.8 1.73 26.00   
 Thoro 19230.0 745.0 216.1 3.45 38.74   
   
 Total 49689.2 1537.0 672.9 2.28 30.93   
 ---------------------------------------------------

In order to assess if the SMR’s between thorotrust and other groups are different fit poisson regression model with offset term as log of the expected number of events calculated earlier offset(log(e)).

m1<-glm(d~offset(log(e)),  
 family=poisson,  
 data=thL.apd[thL.apd$contrast=="Thoro",])  
  
m2<-glm(d~offset(log(e)),  
 family=poisson,  
 data=thL.apd[thL.apd$contrast=="Other",])  
  
ci.lin(m1,Exp=TRUE)[,c(5:7)]

exp(Est.) 2.5% 97.5%   
 3.446775 3.207948 3.703383

ci.lin(m2,Exp=TRUE)[,c(5:7)]

exp(Est.) 2.5% 97.5%   
 1.733781 1.617142 1.858833

## 9. sir() function in popEpi package

library(popEpi)  
gmortDK$ref.rate <- gmortDK$rt/1000  
m <- sir( coh.data = thL.apd, coh.obs = 'd', coh.pyrs = 'y',   
 ref.data = gmortDK, ref.rate = 'ref.rate',   
 adjust = c('agr', 'pgr' ,'sex'),   
 print = 'contrast', conf.type = "wald")  
m

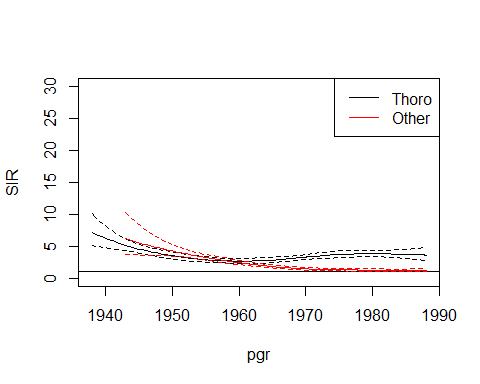
SIR (adjusted by agr, pgr, sex) with 95% confidence intervals (wald)   
Test for homogeneity: p < 0.001   
  
 Total sir: 2.28 (2.17-2.4)  
 Total observed: 1537  
 Total expected: 672.95  
 Total person-years: 49689   
  
  
 contrast observed expected pyrs sir sir.lo sir.hi p\_value  
1: Other 792 456.81 30459.24 1.73 1.62 1.86 0  
2: Thoro 745 216.14 19229.97 3.45 3.21 3.70 0

Estimate SMR over calendar time using splines

mT <- sirspline(   
 coh.data = thL.apd[thL.apd$contrast=="Thoro",], coh.obs = 'd', coh.pyrs = 'y',   
 ref.data = gmortDK, ref.rate = 'ref.rate',   
 adjust = c('agr', 'pgr' ,'sex'), spline = 'pgr', knots=5)  
  
mO <- sirspline(   
 coh.data = thL.apd[thL.apd$contrast=="Other",], coh.obs = 'd', coh.pyrs = 'y',   
 ref.data = gmortDK, ref.rate = 'ref.rate',   
 adjust = c('agr', 'pgr' ,'sex'), spline = 'pgr', knots=5)

Point estimates with 95% confidence intervals for the two groups

plot(mT, conf.int = F, ylim=c(0,30))  
lines(mT, conf.int = T, lty=2)  
lines(mO, conf.int = F,col="red")  
lines(mO, conf.int = T,col="red", lty=2)  
legend("topright",c("Thoro","Other"),col=c("black","red"),lty=1)



## 10. Use data.table to obtain a table of observed and expected deaths cross-classified by age and period:

#stat.table( list( agr, pgr ), list( D=sum(d), E=sum(e) ),  
#margin=T, data=thL.apd )  
  
library(data.table)  
thLnew<-data.table(thL.apd)  
thLnew[order(agr,pgr),  
 .(D=sum(d,na.rm=F),E=sum(e,na.rm=F)),  
 by=list(agr,pgr)]

agr pgr D E  
 1: 5 1938 0 0.003109836  
 2: 5 1943 0 0.013878131  
 3: 5 1948 0 0.019573021  
 4: 5 1953 1 0.036377945  
 5: 5 1958 1 0.036838865  
 ---   
172: 90 1968 1 1.140196708  
173: 90 1973 1 4.311583107  
174: 90 1978 1 5.953333751  
175: 90 1983 8 6.836801157  
176: 90 1988 3 4.748360712

Then, create a similar table restricted to the Thoro patient group using

data= subset( thL.apd, contrast=="Thoro")

(which chooses from the data frame thL.apd only those subjects, whose contrast value was Thoro) as argument to stat.table(). Finally, display a similar table for the Other patient group.