# IDEG 2019 training day
# Advanced stream

# Survival,
# Multiple time scales and
# Competing risks

# Exercises & practicals

IDEG, Seoul
November 2019
http://BendixCarstensen.com/Epi/Courses/IDEG2015/
Version 2

Compiled Friday 22nd November, 2019, 15:32
from: `/home/bendix/teach/Epi/IDEG2019/pracs/pracs.tex`


Bendix Carstensen
Senior Statistician

Clinical Epidemiology
Steno Diabetes Center Copenhagen, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
b@bxc.dk
http://BendixCarstensen.com

# Contents

# Chapter 1

# Notes and practicals

This set of practicals will introduce you to the classical concepts of survival and mortality, and some important pitfalls.

The main example will be a dataset that resembles the Danish National Diabetes Register.

The text follows the structure in the lectures and contains a number of practicals. Basically, the exercises are practicals using R to run the analyses, so that you reproduce the results shown in the text on your own computer.

You are encouraged to use RStudio for running the code, it will enable you to keep track of what the resukts are and in praticular what results you get from modifying the code.

Not all (well, very few) R-commands are explained in detail in this document, so you are encouraged to read about the commands; the simplest way of getting more information an R function, say `ci.pred` is to write

```
> ?ci.pred
```

at the command prompt in R.

All the code shown in this document is available on the course website, bendixcarstensen.com/Epi/Courses/IDEG2019.

# Chapter 2

# Follow-up data in the `Epi` package

Follow-up data for a person basically consists of a time of entry, a time of exit and an indication of the status at exit (normally either "Alive" or "Dead"). Implicitly is also assumed a status *during* the follow-up (usually "Alive"). In multistate modeling, a persons can occupy several states during his follow-up.

In the `Epi`-package, follow-up data is in general represented by adding some extra variables to a data frame with information about dates of different events and other variables. The names of these extra variables are used to keep track of the follow-up. Such a data frame is called a `Lexis` object. The tools for handling follow-up data then use the structure of this data frame for special plots, tabulations etc.

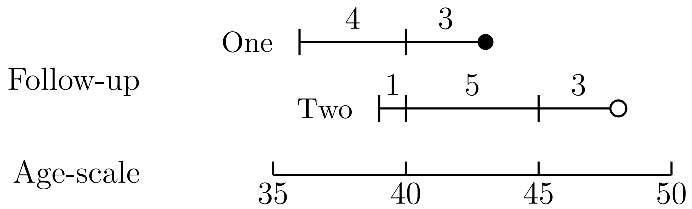


Figure 2.1: *Follow-up of two persons on the age-scale. Follow-up is allocated in small age-bands, whereby we can keep track of the persons'* current *age (a.k.a.* attained *age).*

## 2.1 Timescales

A timescale is a variable that varies deterministically *within* each person during follow-up, *e.g.*:

- Age
- Calendar time
- Time since diagnosis
- Time since treatment start

All timescales advance at the same pace, so the time a person is followed is the same on all timescales. Therefore, it suffices to use only the entry point on each of the time scale, for example:

- Age at entry

- Date of entry
- Time since diagnosis (*at* diagnosis this is 0)
- Time since treatment start (*at* treatment start this is 0)

In the `Epi` package, follow-up in a cohort is represented in a `Lexis` object. A `Lexis` object is a data frame with a bit of extra structure representing the follow-up.

## 2.2 The Danish diabetes data

In the `Epi` package is a dataset `DMlate`, which is a random sample of 10,000 persons from the Danish National Diabetes Register diagnosed between 1995-01-01 and 2009-12-31, followed till 2009-12-31. All dates are jittered by an amount of 7 days, so no set of dates in the dataset represent any real person.

1. First you should look at the documentation:

```
> library( Epi )
> data( DMlate )
> ?DMlate
> head( DMlate )
```

2. Using `DMlate` data, construct a `Lexis` data frame of follow-up of persons from diagnosis (`dodm`) to exit (`dox`), keeping track of whether persons were alive at exit or not (`is.na(dodth)`):

```
> Ldm <- Lexis( entry = list( per=dodm,
+                             age=dodm-dobth,
+                             tfd=0 ),
+                exit = list( per=dox ),
+         exit.status = factor( !is.na(dodth), labels=c("Alive","Dead") ),
+                data = DMlate )
NOTE: entry.status has been set to "Alive" for all.
NOTE: Dropping  4  rows with duration of follow up < tol
```

You can find a further description of the `Lexis` machinery in [1, 2].

The `entry` argument is a *named* list with the entry points on each of the timescales we want to use. It defines the names of the timescales and the entry points. The `exit` argument gives the exit time on *one* of the timescales, so the name of the element in this list must match one of the names of the `entry` list. This is sufficient, because the follow-up time on all time scales is the same, in this case `dox` − `dodm`. Now take a look at the result:

```
> options( digits=5 )
> print( head( Ldm ), digits=2 )
```

```
          per age tfd lex.dur lex.Cst lex.Xst lex.id sex dobth dodm dodth
50185    1999  59   0   11.08   Alive   Alive      1   F  1940 1999    NA
307563 2003  64   0    6.69   Alive   Alive      2   M  1939 2003    NA
294104 2005  86   0    5.45   Alive   Alive      3   F  1918 2005    NA
336439 2009  44   0    0.74   Alive   Alive      4   F  1965 2009    NA
245651 2009  76   0    1.34   Alive   Alive      5   M  1933 2009    NA
216824 2008  80   0    2.04   Alive    Dead      6   F  1928 2008  2010
       dooad doins  dox
50185      NA    NA 2010
307563   2007    NA 2010
294104     NA    NA 2010
336439     NA    NA 2010
245651     NA    NA 2010
216824     NA    NA 2010
```

```
> summary( Ldm )

Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive  7497 2499      9996     2499      54273      9996

> options( digits=7 )
```

The Lexis object Ldm has a variable for each timescale, the value of which is the entry point on this timescale. The follow-up time is in the variable lex.dur (**dur**ation).

3. We can show how the follow-up is distributed over calendar time and age by drawing a life line for each person (blue for men, red for women):

```
> plot( Ldm, col=c("blue","red")[Ldm$sex] )
```

## 2.3 Survival of diabetes patients

We will construct the survival curve for Danish diabetes patients, in the first instance ignoring the effects of sex and age. This is a curve that at any time shows the probability that a diabetes patients survive this far.

### 2.3.1 Kaplan-Meier estimator

The Kaplan-Meier estimator of survival is computed by the survfit function from the survival packages; overall survival as a function of diabetes duration. Since all patients start at duration (tfd), we can use lex.dur as the survival time.

4. Use the function Surv to define a the survival times and the survfit to calculate the Kaplan-Meier estimator:

```
> library( survival )
> km0 <- survfit( Surv(lex.dur,lex.Xst=="Dead") ~ 1, data=Ldm )
> km0
```

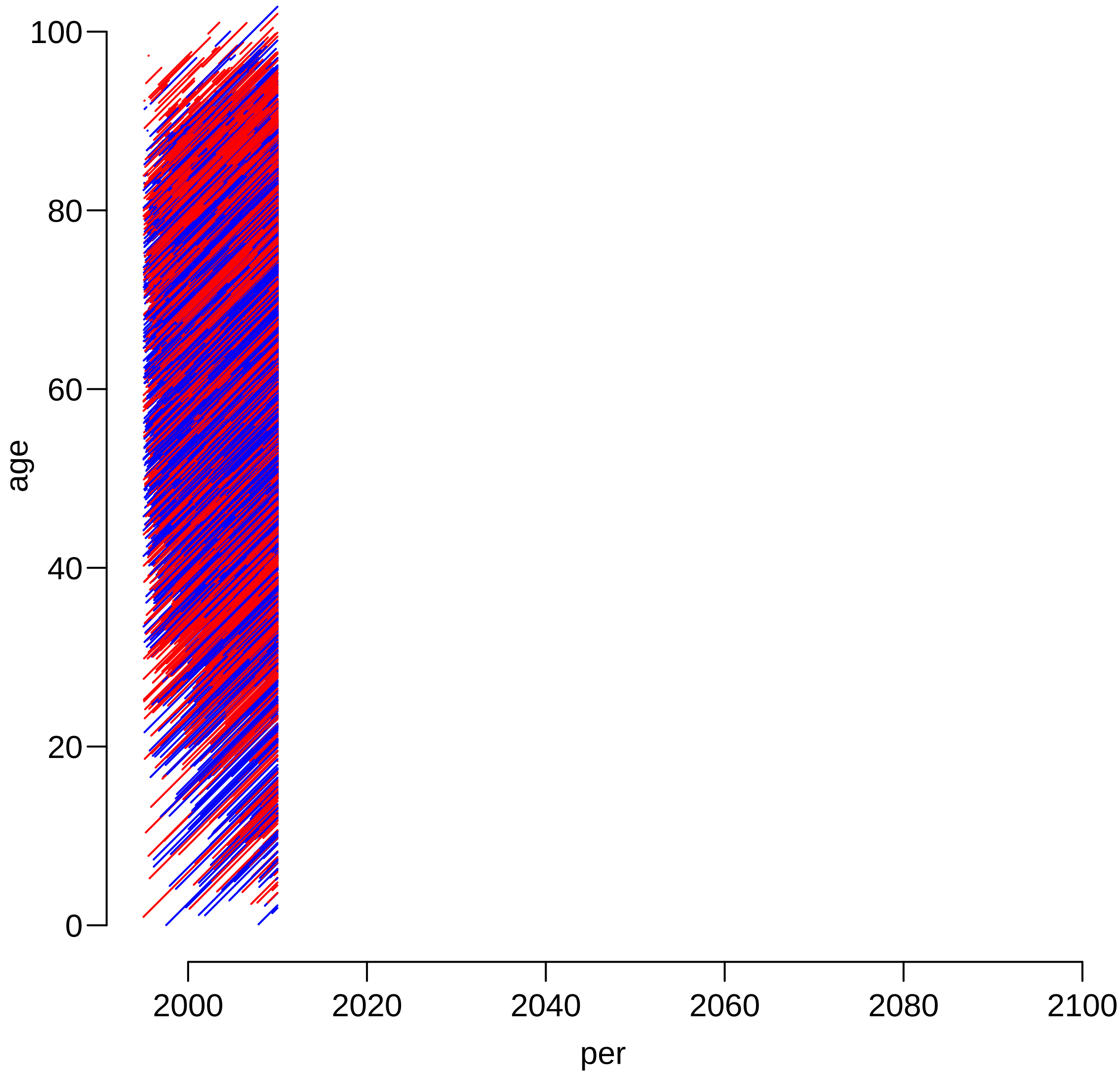


Figure 2.2: *Lexis diagram of the* **DMlate** *dataset.* ./flup-Lexis-dgm

```
Call: survfit(formula = Surv(lex.dur, lex.Xst == "Dead") ~ 1, data = Ldm)

      n  events  median 0.95LCL 0.95UCL
 9996.0  2499.0    14.5    14.2      NA

> plot( km0 )
```

5. We can also derive the survival function separately for men and women:

```
> library( survival )
> kms <- survfit( Surv(lex.dur,lex.Xst=="Dead") ~ sex, data=Ldm )
> kms

Call: survfit(formula = Surv(lex.dur, lex.Xst == "Dead") ~ sex, data = Ldm)

         n events median 0.95LCL 0.95UCL
sex=M 5183   1343   13.8    12.9      NA
sex=F 4813   1156   14.8    14.4      NA
```

```
> plot( kms, col=c("blue","red") )
```

### 2.3.2 Cox model for mortality

6. Now estimate the effect of sex by using the Cox model:

```
> cs <- coxph( Surv(lex.dur,lex.Xst=="Dead") ~ sex, data=Ldm )
> summary( cs )

Call:
coxph(formula = Surv(lex.dur, lex.Xst == "Dead") ~ sex, data = Ldm)

  n= 9996, number of events= 2499

         coef exp(coef) se(coef)     z Pr(>|z|)
sexF -0.11559   0.89084  0.04013 -2.88  0.00397

     exp(coef) exp(-coef) lower .95 upper .95
sexF    0.8908      1.123    0.8235    0.9637

Concordance= 0.516  (se = 0.005 )
Likelihood ratio test= 8.32  on 1 df,   p=0.004
Wald test            = 8.3  on 1 df,   p=0.004
Score (logrank) test = 8.31  on 1 df,   p=0.004
```

The Cox model is a model for hazard rates; the parameter we estimate is the *hazard ratio*. The associated survival function(s) are derived using the so-called *Breslow estimator*, implemented in `survfit.coxph`.

7. In this case, we derive two survival functions—one for men and one for women (using the `newdata` argument):

```
> plot( survfit( cs, newdata=data.frame(sex=c("M","F")) ),
+       col=c("blue","red"), lwd=2 )
```

Note that the two curves have jumps at the same points. The curves are not parallel, but the log of the curves are proportional—a consequence of the proportional hazards assumption in the Cox model.

8. We can compare the estimated survival functions for men and women as estimated separately by Kaplan-Meier estimators with those derived from the Cox-model:

```
> plot( survfit(cs, newdata=data.frame(sex=c("M","F")) ),
+       col=c("blue","red"), lwd=2 )
> lines( kms, col=c("blue","red") )
```

Even though the Cox model is really a model for the mortality rates, but we only ever see the rates as transformed to the survival scale. From the figure **??** it looks as if there is a peak in mortality in the first short period after diagnosis.

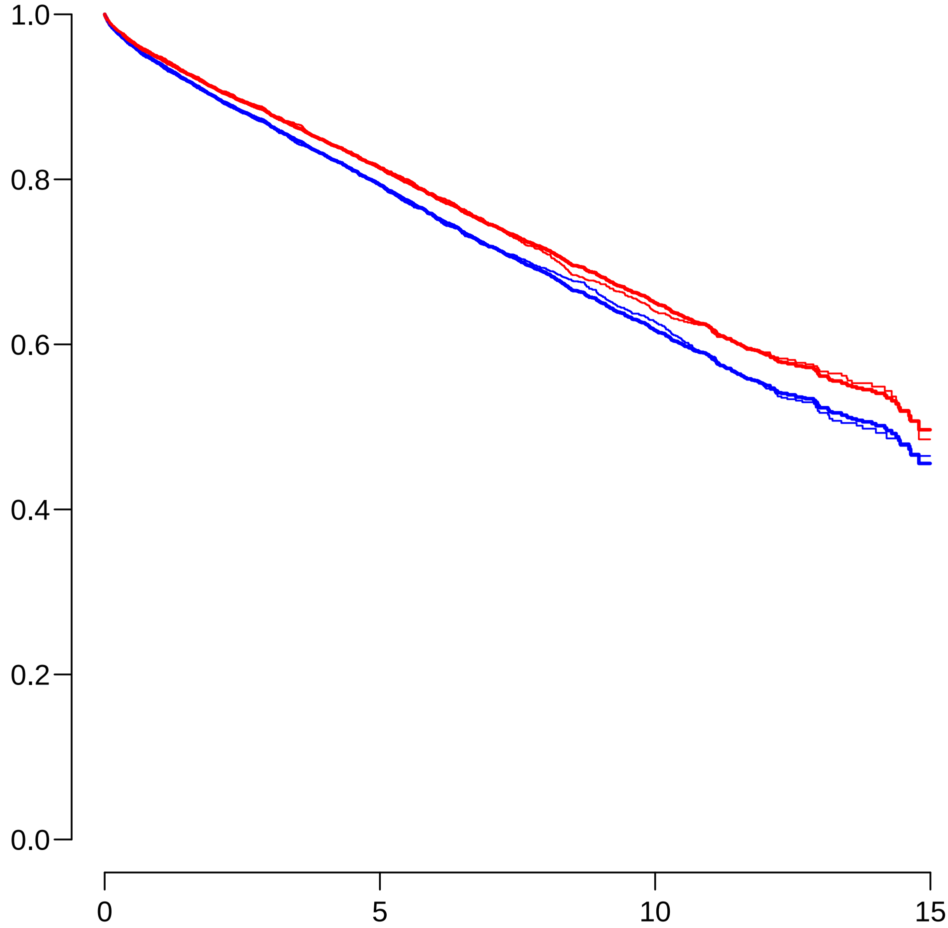


Figure 2.3: *Estimated survival curves for men and women, derived from a Cox-model with sex as the only covariate.* `./flup-cox-cmp`

## 2.4 Modeling mortality rates

If we want to get a handle on the underlying hazard $\lambda_0(t)$ we must resort to a `parametric` model for the hazard—an explicit model model for the baseline mortality rates.

### 2.4.1 Simple model for rate

A very simple model would be one where the rate were constant, estimated by the total number of deaths divided by the total risk time; `lex.Xst` is the variable that records in what state the person ends his follow-up — `Dead` is the state we are interested in and `lex.dur` is the variable that holds the risk time:

```
> sum(Ldm$lex.Xst=="Dead") / ( sum(Ldm$lex.dur)/1000 )
[1] 46.04477
```

9. We could do a hand-calculation of the confidence limits of this quantity, but it is easier to fit a Poisson model for rates, where we have the follow-up (events,person-years) as

outcome, and just `1` (the intercept) as covariate. Note that we are using the `poisreg` family available in the `Epi` package, where the response variable is a two column vector of events, resp. risk time:

```
> m0 <- glm( cbind(lex.Xst=="Dead",lex.dur/1000) ~ 1,
+            family = poisreg, data = Ldm )
> ci.exp( m0 )

            exp(Est.)     2.5%    97.5%
(Intercept)  46.04477 44.27447 47.88585
```

What is the parameter you estimate by the `(Intercept)`?

In most of the literature you will find phrases as: "…fitted a Poisson model with log-person-years as offset…". This refers to the common twist of Poisson models where you obtain the same result slightly differently:

```
> m1 <- glm( lex.Xst=="Dead" ~ 1, offset = log(lex.dur/1000),
+            family = poisson, data = Ldm )
> ci.exp( m1 )
            exp(Est.)     2.5%    97.5%
(Intercept)  46.04477 44.27453 47.88579
```

The result is exactly the same, but the `poisreg` approach is faster for larger datasets, for small datasets the difference in computing time is irrelevant. The main advantage is that it is more natural to have the response (events, risk time) on the l.h.s. of the tilde. We shall use the `poisreg` for the rest of this exercise.

10. In the `Epi` package is also a convenience wrapper for analysis of rates from data in a `Lexis` data frame:

```
> m2 <- glm.Lexis( Ldm, ~ 1, scale=1000 )

stats::glm Poisson analysis of Lexis object Ldm with log link:
Rates for the transition: Alive->Dead, lex.dur (person-time) scaled by 1000

> ci.exp( m2 )

            exp(Est.)     2.5%    97.5%
(Intercept)  46.04477 44.27447 47.88585
```

But this simple overall mortality estimate was not really what we are after, we want to know how mortality depends on time since diagnosis.

### 2.4.2 Splitting the follow-up time along a timescale

The follow-up time in a cohort can be subdivided by for example current age. This is achieved by the `splitLexis` (note that it is *not* called `split.Lexis`) or better, the `splitMulti` function from the `popEpi` package. This requires that the timescale and the breakpoints on this timescale are supplied.

11. Try splitting the time, where you use 1/20 years for the first year of follow-up and 1/2 year for the rest. Start out exploring how you do simple sequences with R:

```
> 0:19
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
> 0:19/20
 [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65
[15] 0.70 0.75 0.80 0.85 0.90 0.95
> c(0:19/20,seq(1,20,0.5))
 [1]  0.00  0.05  0.10  0.15  0.20  0.25  0.30  0.35  0.40  0.45  0.50
[12]  0.55  0.60  0.65  0.70  0.75  0.80  0.85  0.90  0.95  1.00  1.50
[23]  2.00  2.50  3.00  3.50  4.00  4.50  5.00  5.50  6.00  6.50  7.00
[34]  7.50  8.00  8.50  9.00  9.50 10.00 10.50 11.00 11.50 12.00 12.50
[45] 13.00 13.50 14.00 14.50 15.00 15.50 16.00 16.50 17.00 17.50 18.00
[56] 18.50 19.00 19.50 20.00
> library( popEpi )
> Sdm <- splitMulti( Ldm, tfd=c(0:19/20,seq(1,20,0.5)) )
> summary( Ldm )
Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive  7497 2499      9996      2499   54273.27      9996
> summary( Sdm, t=T )
Transitions:
     To
From     Alive Dead  Records:  Events: Risk time:  Persons:
  Alive 277890 2499    280389      2499   54273.27      9996

Timescales:
per age tfd
 ""  ""  ""
```

The split dataset now contains many records from each person, but the number of dead and the risk time is the same.

In the split `Lexis` object there are now three timescales that varies across the follow-up, `per`, `age` and `tfd`:

```
> Sdm[1:10,1:8]
   lex.id      per      age  tfd lex.dur lex.Cst lex.Xst sex
 1:      1 1998.917 58.66119 0.00    0.05   Alive   Alive   F
 2:      1 1998.967 58.71119 0.05    0.05   Alive   Alive   F
 3:      1 1999.017 58.76119 0.10    0.05   Alive   Alive   F
 4:      1 1999.067 58.81119 0.15    0.05   Alive   Alive   F
 5:      1 1999.117 58.86119 0.20    0.05   Alive   Alive   F
 6:      1 1999.167 58.91119 0.25    0.05   Alive   Alive   F
 7:      1 1999.217 58.96119 0.30    0.05   Alive   Alive   F
 8:      1 1999.267 59.01119 0.35    0.05   Alive   Alive   F
 9:      1 1999.317 59.06119 0.40    0.05   Alive   Alive   F
10:      1 1999.367 59.11119 0.45    0.05   Alive   Alive   F
```

This means that we can use any of these as explanatory variable for the rates, but for now we shall stick to `tfd`, `time from diagnosis`.

## 2.5 Modeling mortality rates

We are primarily interested in the effect of time since diagnosis, so we use `tfd` as quantitative covariate with a non-linear effect. This can for example be done by natural splines where we would need to choose various parameters (knots) for the spline.

We can instead use the `gam` (generalized additive models) machinery from the `mgcv` package, that fits a smooth function of a quantitative covariate, using penalized splines.

12. There is a wrapper for `gam` modeling data in `Lexis` objects—the `s()` specifies a smooth function:

```
> g0 <- gam.Lexis( Sdm, ~ s(tfd) )

mgcv::gam Poisson analysis of Lexis object Sdm with log link:
Rates for the transition: Alive->Dead

> summary( g0 )

Family: poisson
Link function: log

Formula:
cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(tfd)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.93227    0.02661  -110.2   <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(tfd) 7.714  8.501  115.9  <2e-16

R-sq.(adj) =  -9.48e-05   Deviance explained = 0.425%
UBRE = -0.90613  Scale est. = 1         n = 280389
```

13. The summary tells us very little about the shape of the effect; to this end we will need a `prediction` data frame where we indicate the points for which we want the predicted rates computed, which is done by the `ci.pred` function:

```
> nd <- data.frame( tfd=seq(0,15,0.1) )
> matshade( nd$tfd, ci.pred( g0, nd ), log="y" )
```

Try to make it a little nicer by scaling the rates to events per 100 years (%/year), and putting on proper axis annotation: We see that is a nadir at about 2 years after diagnosis, but also that the mortality increases only moderately after this.

### 2.5.1 Modeling survival

There is a 1-to-1 correspondence between mortality on one side and survival from a given point on the other hand, so the mortality curve can be transformed to a survival curve.

14. This is done in a similar way as predicting the mortality, using the `ci.surv` function:

```
> matshade( nd$tfd, ci.surv( g0, nd, int=0.1 ), lwd=2 )
> lines( survfit( Surv(lex.dur,lex.Xst=="Dead") ~1, data=Ldm) )
```

The second statement overlays the Kaplan-Meier with 95% CI in dotted lines, and they produce the same substantial conclusion, with a small tendency that the confidence intervals are slightly smaller.

15. The Cox model we fitted and used to predict survival separately for men and women is a *proportional hazards* model, meaning the the effect of time (`tfd`) is the same for men and women, and that the effect of sex is `multiplicative` — additive on the log-scale:

```
> gs <- gam.Lexis( Sdm, ~ s(tfd) + sex )

mgcv::gam Poisson analysis of Lexis object Sdm with log link:
Rates for the transition: Alive->Dead

> rbind( ci.exp( gs, subset="sex" ),
+        ci.exp( cs ) )

     exp(Est.)      2.5%     97.5%
sexF 0.8906416 0.8232732 0.9635228
sexF 0.8908372 0.8234534 0.9637351
```

So we see that the two estimated effects of age are the same in the Cox model as in the `gam` model.

16. We can show the hazard rates for women and men and the corresponding hazard rates and survival functions. To that end we need prediction data frames for men and women

```
> ndm <- data.frame( tfd=seq(0,15,0.1), sex="M" )
> ndf <- data.frame( tfd=seq(0,15,0.1), sex="F" )
> par( mfrow=c(1,2) )
> matshade( ndm$tfd, cbind( ci.pred(gs,ndm),
+                           ci.pred(gs,ndf) )*100, plot=TRUE,
+           col=c("blue","red"), log="y", lwd=2,
+           xlab="Time since diagnosis (years)",
+           ylab="Mortality per 100 PY" )
> matshade( ndm$tfd, cbind( ci.surv(gs,ndm,int=0.1),
+                           ci.surv(gs,ndf,int=0.1) ), plot=TRUE,
+           col=c("blue","red"), ylim=0:1,
+           xlab="Time since diagnosis (years)",
+           ylab="Survival probability" )
```

17. We can see to what extent males and women have the same mortality shape ("test of proportionality"), that is whether there is an *interaction* between time (`tfd`) and sex, by fitting the interaction model and comparing to the model without interaction:

```
> gi <- gam.Lexis( Sdm, ~ s(tfd, by=sex) + sex )

mgcv::gam Poisson analysis of Lexis object Sdm with log link:
Rates for the transition: Alive->Dead

> anova( gs, gi, test="Chisq" )
```

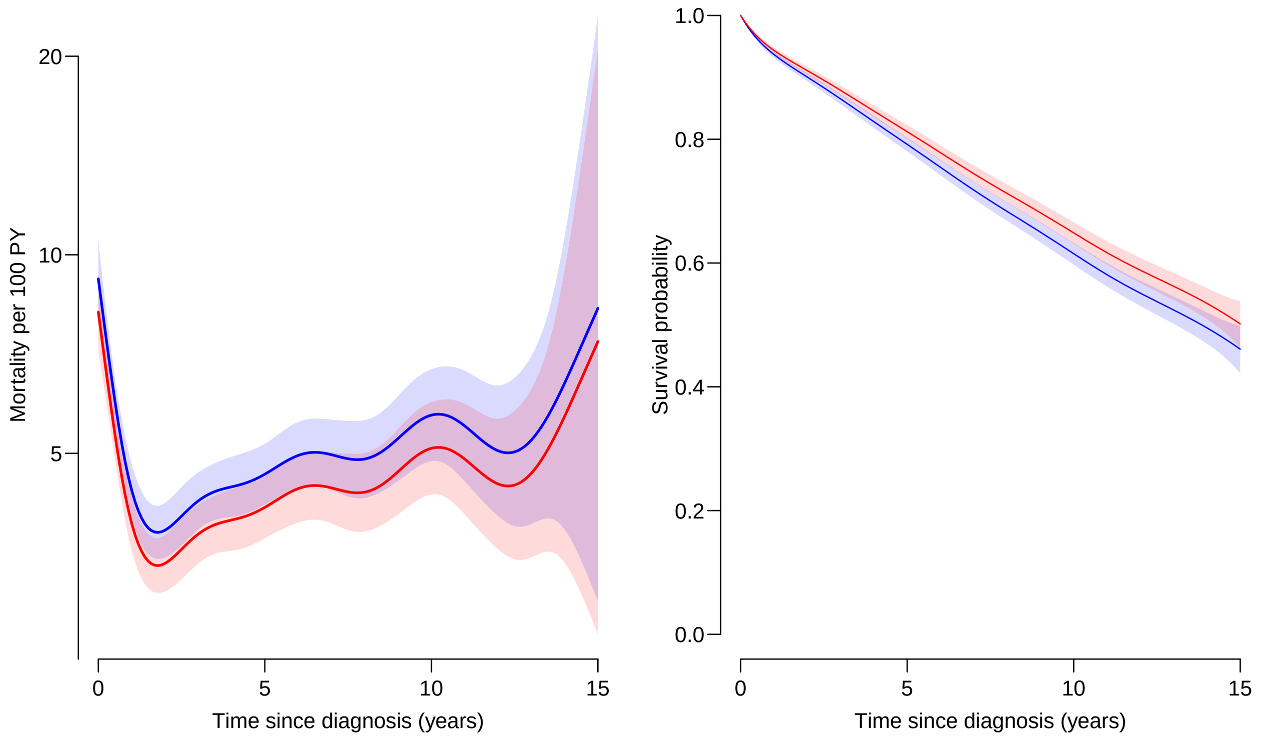


Figure 2.4: *Left: Mortality rates for men (blue) and women (red) with 95% CI (shades). Right: Survival curves for men and women from the main effects model.* ./flup-gam-reg

```
Analysis of Deviance Table

Model 1: cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(tfd) +
    sex
Model 2: cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(tfd,
    by = sex) + sex
  Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
1    280378      26294
2    280371      26280 7.7107   13.339  0.08908
```

Is there an interaction between sex and time since diagnosis?

Note that we keep the term `sex` in the model when we add the interaction using `by=sex`.

18. We can also plot the fitted rates for men and women and the rate-ratio as functions of time. The test for proportionality is not formally significant, but a deviance difference of 13 may easily hide an interesting structure, so we plot the two baseline rates and their ratio.

```
> par( mfrow=c(1,2) )
> matshade( ndm$tfd, cbind( ci.pred(gi,ndm)*100,
+                           ci.pred(gi,ndf)*100,
+                           ci.exp (gi,list(ndm,ndf)) ), plot=TRUE,
+           col=c("blue","red","black"), log="y", ylim=c(0.5,10),
+           xlab="Time since diagnosis (years)",
+           ylab="Mortality per 100 PY" )
> abline( h=c(1,exp(-coef(gs)["sexF"])), lty=3 )
```

```
> matshade( ndm$tfd, cbind( ci.surv(gi,ndm,int=0.1),
+                           ci.surv(gi,ndf,int=0.1) ), plot=TRUE,
+           col=c("blue","red"), ylim=0:1,
+           xlab="Time since diagnosis (years)",
+           ylab="Survival probability" )
```

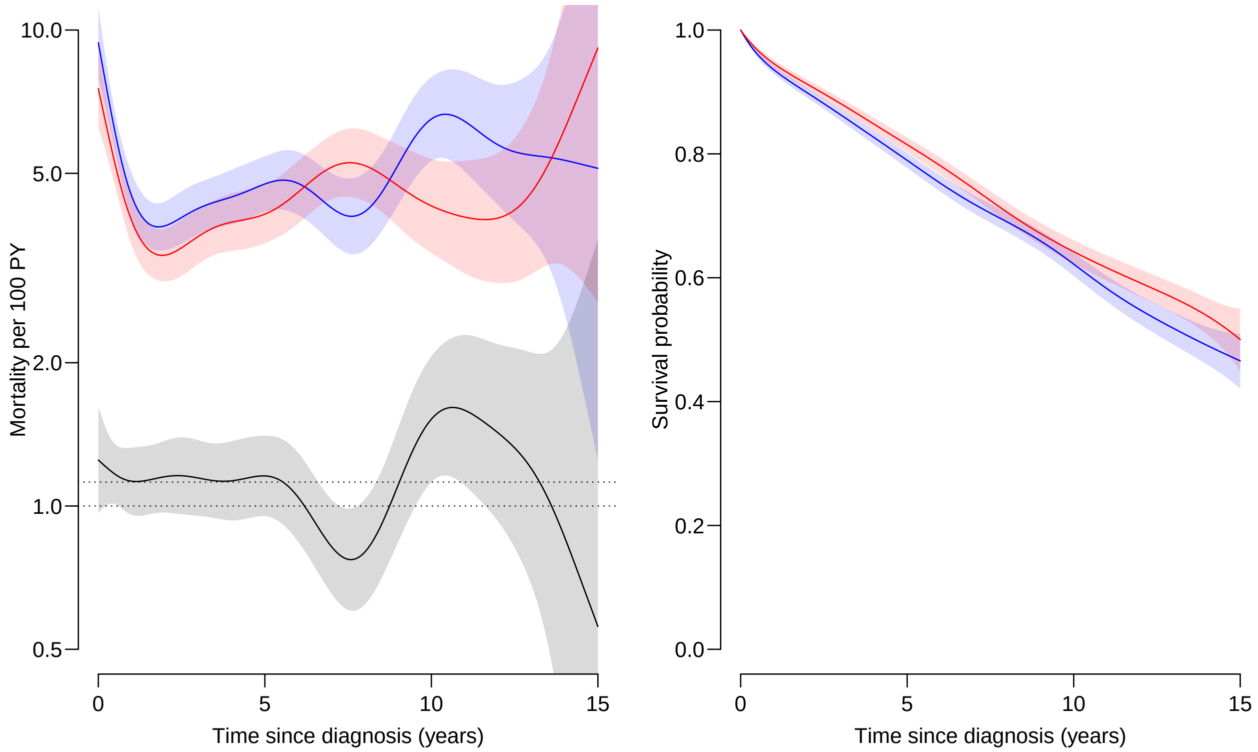


Figure 2.5: *Left: Mortality rates for men (blue) and women (red) and their rate-ratio (black) with 95% CI (sha.des). The dotted horizontal lines are at 1 and at the M/W rate-ratio from the main effects model. Right: Survival curves for men and women from the interaction model.* `./flup-gam-int`

It is clear from figure 2.5 that there are no noteworthy deviations from proportionality — the black curve has no clinically meaningful deviations from the horizontal dotted line at the overall RR.

## 2.6 Multiple time scales

It would be natural think that mortality depends not only on duration of diabetes but also on age, and perhaps even also on age *at* diagnosis.

### 2.6.1 Theory of multiple time scales

This is technical section that you may skip; it is not essential for doing the practical. But useful for understanding it.

Suppose we describe the mortality rates as a function of current age, $a$; duration of diabetes, $d$ and age at diagnosis, $e = a - d$ ("$e$" for **e**ntry into diabetes), then we have that

$a-d-e=0$. If we formally set up a model with only the effect of current age and age at diagnosis of diabetes:

$$\log\big(\mu(a,d)\big) = f(a) + h(e)$$

it is only superficially that this does not include duration; because since $d=a-e$, we may write:

$$\begin{aligned}
\log\big(\mu(a,d)\big) &= f(a) + h(e) + \beta d - \beta d \\
&= f(a) + h(e) + \beta(a-e) - \beta d \\
&= \big(f(a) + \beta a\big) + \big(h(e) - \beta e\big) - \beta d
\end{aligned}$$

Thus, even if duration is not formally included in the model we may claim that is has any linear effect we like, by simply asserting that the age and age at diagnosis effects are different. And there is no way to allocate a "correct" duration effect. One might of course on purely external grounds (*i.e.* unrelated to the data at hand) assert that there is no duration effect, for example. But this will never be founded in data.

Therefore, it makes more sense to set up a model with non-linear effects of all three variables. But we still have the problem from the linear dependence:

$$\begin{aligned}
\log\big(\mu(a,d)\big) &= f(a) + g(d) + h(e) \\
&= f(a) + g(d) + h(e) + \gamma(a-d-e) \\
&= \big(f(a) + \gamma a\big) + \big(g(d) - \gamma d\big) + \big(h(e) - \gamma e\big) \\
&= \tilde{f}(a) + \tilde{g}(d) + \tilde{h}(e)
\end{aligned}$$

so we have two *different* sets of three effects that together produce the same mortality rates; this would be valid for *any* value of $\gamma$ we care to stick in the formula. This is essentially the age-period-cohort modeling problem once again, see [3].

However, even if we cannot separate the three effects in the model, we can still make perfectly valid predictions from the model, and certain contrasts will also be identifiable from the model. Notably we shall be able to estimate the mortality rate-ratio at a given age ($a$) between persons diagnosed at different ages, $e_1$ and $e_0$, and hence duration $a-e_1$ and $a-e_0$:

$$\begin{aligned}
\log(\mathrm{RR}) =& f(a) + g(a-e_1) + h(e_1) - \\
& f(a) - g(a-e_0) - h(e_0) \\
=& g(a-e_1) - g(a-e_0) + h(e_1) - h(e_0)
\end{aligned}$$

Since any other possible set of effects $\tilde{f}$, $\tilde{g}$ and $\tilde{h}$ are distinguished from these by a term $\gamma$ times the variable $(a-d-e)$, using these would yield:

$$\begin{aligned}
\log(\mathrm{RR}) =& \tilde{g}(a-e_1) - \tilde{g}(a-e_0) + \tilde{h}(e_1) - \tilde{h}(e_0) \\
=& \big(g(a-e_1) - \gamma(a-e_1)\big) - \\
& \big(g(a-e_0) - \gamma(a-e_0)\big) + \\
& \big(h(e_1) - \gamma e_1\big) - \\
& \big(h(e_0) - \gamma e_0\big) \\
=& g(a-e_1) - g(a-e_0) + h(e_1) - h(e_0) + \gamma(-a + e_1 + a - e_0 - e_1 + e_0) \\
=& g(a-e_1) - g(a-e_0) + h(e_1) - h(e_0)
\end{aligned}$$

showing that these contrasts are invariant under *any* reparametrization, and hence *are* identifiable from the model.

### 2.6.2 Practice of multiple time scales

You can find a published example of this type of analysis in [4], it is about mortality among Austraian diabetes patients.

Note that the `Lexis` data frame `Sdm` has three time scales defined; that is variable that vary *within* each person's records

```
> summary(Sdm,t=T)
Transitions:
     To
From     Alive Dead  Records:  Events: Risk time:  Persons:
  Alive 277890 2499    280389     2499   54273.27      9996

Timescales:
per age tfd
 ""  ""  ""
```

We are primarily interested in how mortality depend on age (current age, age at follow-up), duration of diabetes and age at diagnosis. Note that the latter is not a time scale it is a time-fixed variable that is constant within persons.

In terms of the variables in `Sdm` we are interested in current age (`age`), duration of diabetes (`tfd`) and age at diabetes diagnosis (`age` − `tfd`).

19. We can fit a model with these three variables:

```
> made <- gam.Lexis( transform( Sdm, ain=age-tfd ), ~ s(age) + s(tfd) + s(ain) )

mgcv::gam Poisson analysis of Lexis object transform(Sdm, ain = age - tfd) with log link
Rates for the transition: Alive->Dead

> summary( made )

Family: poisson
Link function: log

Formula:
cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(age) +
    s(tfd) + s(ain)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50860    0.03772  -93.01   <2e-16

Approximate significance of smooth terms:
           edf  Ref.df   Chi.sq p-value
s(age) 1.35858 1.64247 1329.054  <2e-16
s(tfd) 7.72148 8.50688  129.867  <2e-16
s(ain) 0.01354 0.02479    0.001   0.973

Rank: 27/28
R-sq.(adj) =  0.000915   Deviance explained = 9.14%
UBRE = -0.91433  Scale est. = 1         n = 280389
```

20. We see there is only a tiny effect of the `ain`, so we can omit this from the model:

```
> mad  <- gam.Lexis( transform( Sdm, ain=age-tfd ), ~ s(age) + s(tfd) )
mgcv::gam Poisson analysis of Lexis object transform(Sdm, ain = age - tfd) with log link
Rates for the transition: Alive->Dead

> summary( mad )

Family: poisson
Link function: log

Formula:
cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(age) +
    s(tfd)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50563    0.03844  -91.21   <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(age) 1.561  1.956   1665  <2e-16
s(tfd) 7.722  8.507    130  <2e-16

R-sq.(adj) =  0.000915   Deviance explained = 9.14%
UBRE = -0.91433  Scale est. = 1         n = 280389
```

21. We can the make a standard plot of the estimated effects:

```
> par( mfrow=c(1,2) )
> plot( mad )
```

The plots in figure 2.6 only show the *shapes* of the curves and their relative sizes (note the $x$-axes are different), not their *joint* effects. In order to show these we must make predictions of the rates where both time scales vary as in real life.

22. In order to report the mortality rates along several time scales we must allow the time to progress on all time scales at the same time, so we choose combinations of duration (`tfd`) and age at diagnosis (`ain`), and based on these compute the current age (`age`):

```
> nd <- data.frame( expand.grid( tfd=c(NA,seq(0,15,.1)),
+                                ain=c(3:7*10) ) )[-1,]
> nd$age = nd$ain + nd$tfd
```

We see that the `tfd` and the `age` varies at the the same pace, while `ain` in constant in each chunk. The `NA`s are inserted in order to separate the lines belonging to each date of diagnosis curve.

```
> matshade( nd$age, ci.pred( mad, nd )*1000, plot=TRUE,
+           lwd=3, lty=1, log="y", las=1,
+           xlab="Age at FU (years)",
+           ylab="Mortality rate per 1000 PY" )
> abline( v=3:7*10 )
```

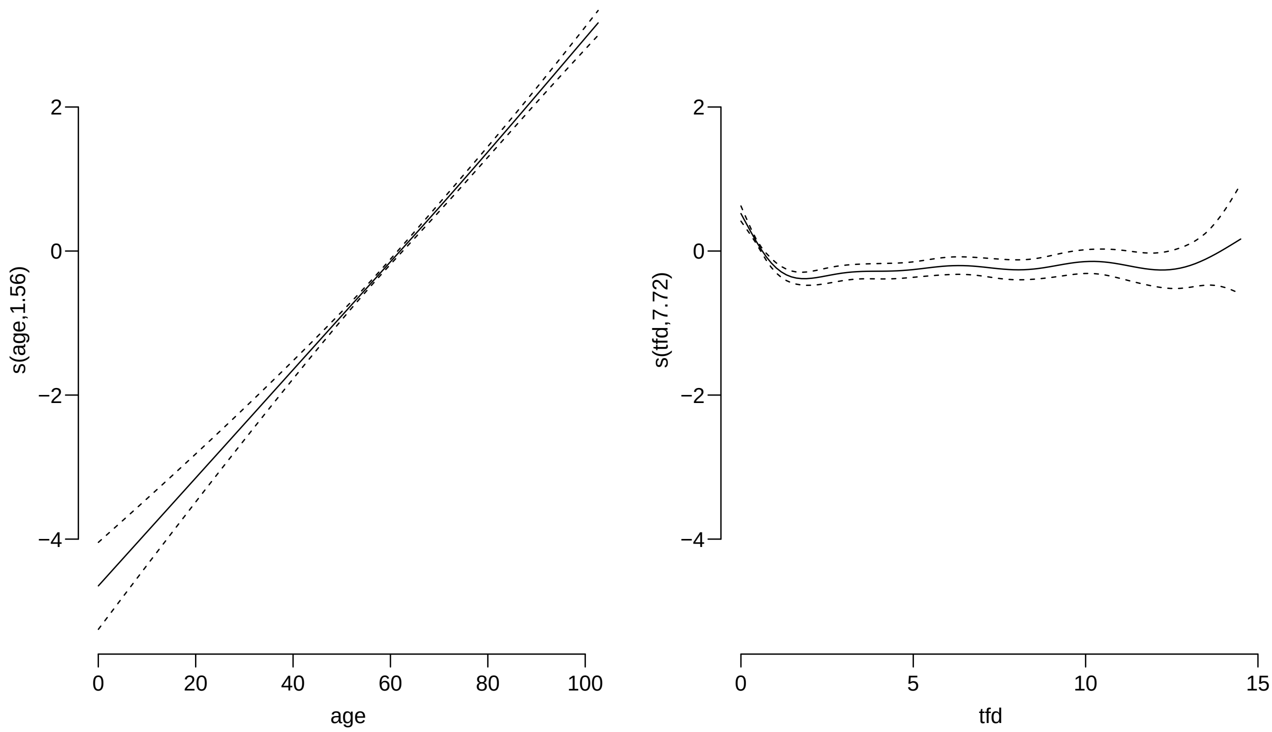


Figure 2.6: *Standard plots of the effects of age and duration from the mortality model.* `./flup-mort-std`

What is is your conclusion for the effect of duration / age at diagnosis on the mortality rates?

23. We can of course also make the analyses separately for men and women and show them together, also showing the M/W rate-ratio:

```
> mm <- gam.Lexis( subset( Sdm, sex=="M" ), ~ s(age) + s(tfd) )

mgcv::gam Poisson analysis of Lexis object subset(Sdm, sex == "M") with log link:
Rates for the transition: Alive->Dead

> mw <- gam.Lexis( subset( Sdm, sex=="F" ), ~ s(age) + s(tfd) )

mgcv::gam Poisson analysis of Lexis object subset(Sdm, sex == "F") with log link:
Rates for the transition: Alive->Dead

> summary(mm)

Family: poisson
Link function: log

Formula:
cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(age) +
    s(tfd)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.39565    0.05733  -59.23   <2e-16
```

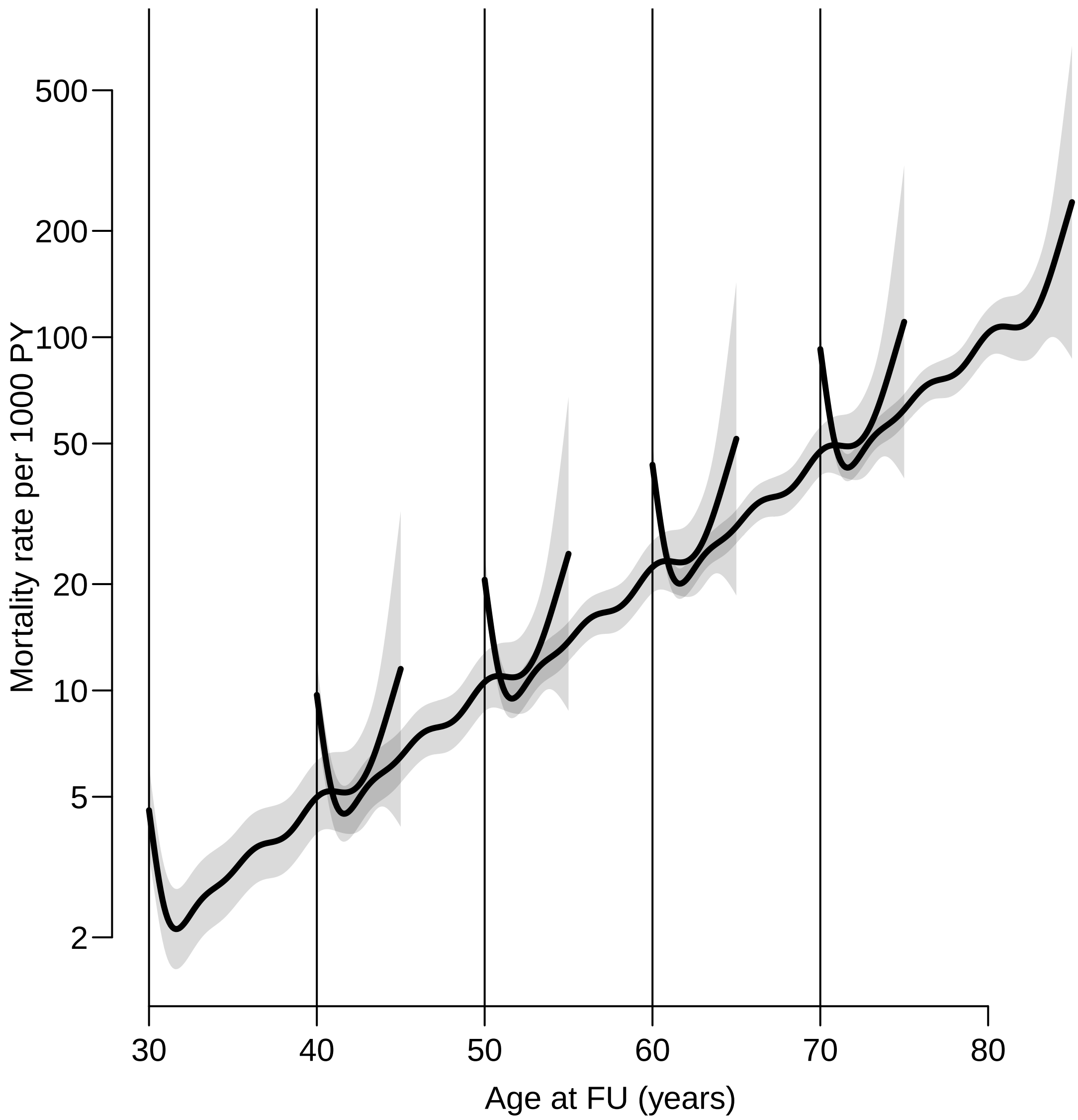


Figure 2.7: *Mortality rates for Danish diabetes patients by age and duration of diabetes for persons diagnosed at ages 30, 40 etc.* `./flup-mort`

```
Approximate significance of smooth terms:
          edf Ref.df  Chi.sq  p-value
s(age) 4.146  5.090 1015.90  < 2e-16
s(tfd) 7.461  8.323   84.92 1.35e-14

R-sq.(adj) =  0.000816   Deviance explained = 8.95%
UBRE = -0.91055  Scale est. = 1         n = 144190

> summary(mw)

Family: poisson
Link function: log

Formula:
cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(age) +
```

```
    s(tfd)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.65908    0.06489  -56.39   <2e-16

Approximate significance of smooth terms:
         edf Ref.df Chi.sq  p-value
s(age) 2.497  3.168 973.64  < 2e-16
s(tfd) 6.982  7.977  54.96 5.23e-09

R-sq.(adj) =  0.00112   Deviance explained = 10.3%
UBRE = -0.91904  Scale est. = 1         n = 136199
```

```
> matshade( nd$age, cbind( ci.pred( mm, nd )*1000,
+                          ci.pred( mw, nd )*1000,
+                ci.ratio( ci.pred( mm, nd ),
+                          ci.pred( mw, nd ) ) ), plot=TRUE,
+           lwd=3, lty=1, log="y", las=1, col=c("blue","red","black"),
+           xlab="Age at FU (years)",
+           ylab="Mortality rate per 1000 PY" )
> abline( h=1 )
```

From figure 2.8 we see approximately the same pattern as for overall mortality rates, a drop in mortality rates during the first about 2 years.

## 2.7 Cutting follow-up time at a specific date

If we have a recording of the date of a specific event as for example recovery, relapse or drug initiation, we can classify follow-up time as being before of after this intermediate event. This is what is usually termed a `time-dependent` covariate — it has different values during different parts of the follow-up for a single person.

This is achieved with the function `cutLexis`, which takes three arguments: the time point, the name of timescale that the time refers to, and the value of the (new) state following the date.

24. Now we define the date that a persons initiates drug treatment (the earlier of `dooad` and `doins`):

```
> Sdm$dodr <- pmin(Sdm$dooad,Sdm$doins,na.rm=TRUE)
```

Then use this to subdivide records of follow-up into records that concerns time before and time after drug initiation:

```
> S3 <- cutLexis( data = Sdm,
+                  cut = Sdm$dodr,
+            timescale = "per",
+            new.state = "Drug",
+     precursor.states = "Alive" )
> summary( Sdm )
```

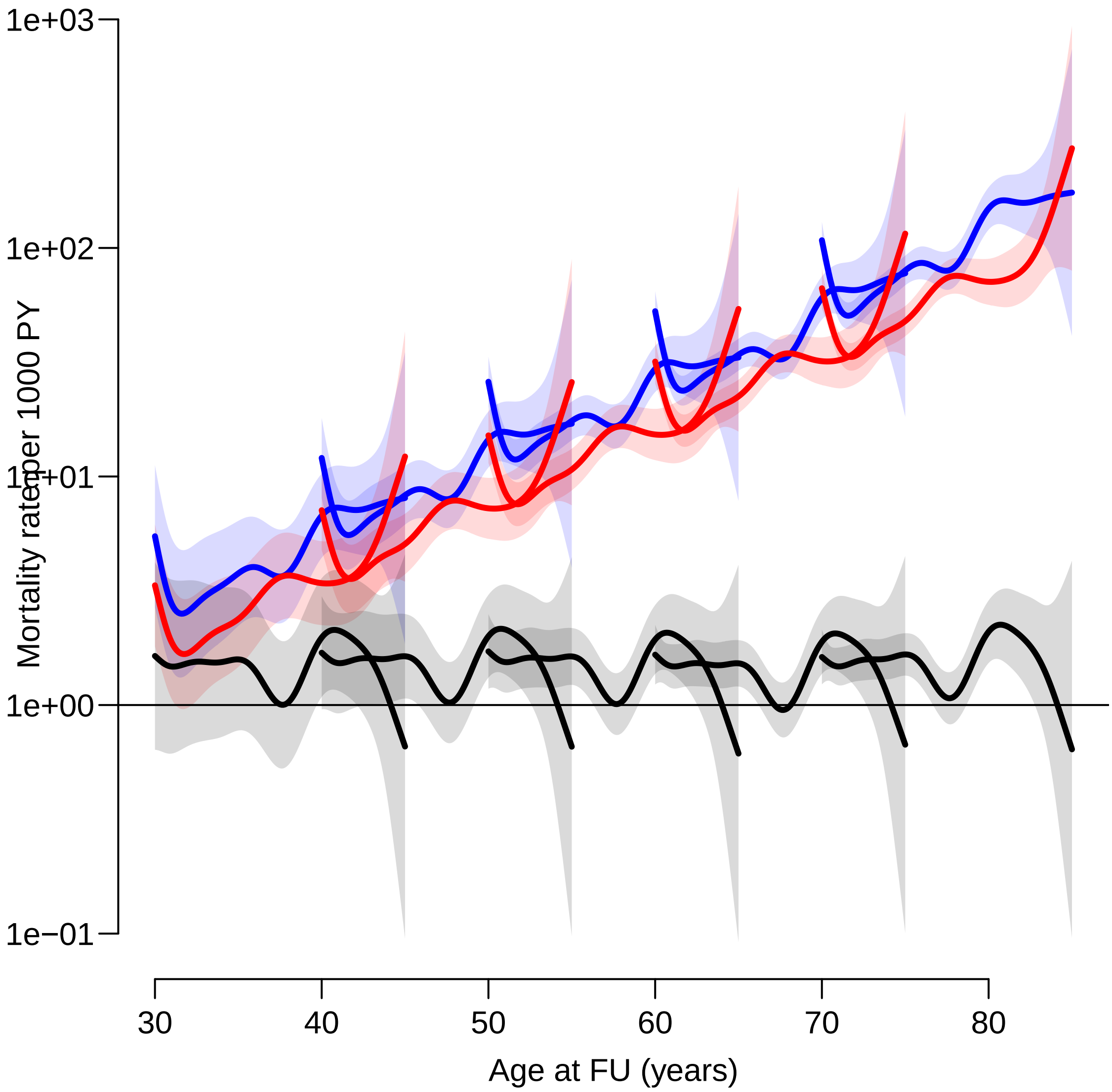


Figure 2.8: *Mortality rates for Danish diabetes patients by sex, age and duration of diabetes for persons diagnosed at ages 30, 40 etc. Men i blue, women in red and the M/W rate ratio in black.* ./flup-mort-sex

```
Transitions:
     To
From      Alive Dead  Records:  Events: Risk time:  Persons:
  Alive 277890 2499    280389     2499   54273.27      9996
```

```
> summary( S3 )
```

```
Transitions:
     To
From      Alive   Drug Dead  Records:  Events: Risk time:  Persons:
  Alive 140147   3646 1056    144849     4702   22920.27      7532
  Drug       0 137743 1443    139186     1443   31353.00      6110
  Sum   140147 141389 2499    284035     6145   54273.27      9996
```

The `precursor.states=` argument is naming those states that will be over-written by the new state. For example, person 12 exits as `Alive` (a precursor state), and thus in the new data frame the person ends in state `Drug`, but person 75 exits `Dead` and so even if he goes to `Drug`, he still exits `Dead`, `Dead` is *not* a precursor state.

25. You can inspect how the follow-up records look for two select persons:

```
> subset( Sdm, lex.id %in% c(12,75) )[,1:8]
```

```
    lex.id      per      age  tfd    lex.dur lex.Cst lex.Xst sex
 1:     12 2008.114 56.90075 0.00 0.05000000   Alive   Alive   M
 2:     12 2008.164 56.95075 0.05 0.05000000   Alive   Alive   M
 3:     12 2008.214 57.00075 0.10 0.05000000   Alive   Alive   M
 4:     12 2008.264 57.05075 0.15 0.05000000   Alive   Alive   M
 5:     12 2008.314 57.10075 0.20 0.05000000   Alive   Alive   M
 6:     12 2008.364 57.15075 0.25 0.05000000   Alive   Alive   M
 7:     12 2008.414 57.20075 0.30 0.05000000   Alive   Alive   M
 8:     12 2008.464 57.25075 0.35 0.05000000   Alive   Alive   M
 9:     12 2008.514 57.30075 0.40 0.05000000   Alive   Alive   M
10:     12 2008.564 57.35075 0.45 0.05000000   Alive   Alive   M
11:     12 2008.614 57.40075 0.50 0.05000000   Alive   Alive   M
12:     12 2008.664 57.45075 0.55 0.05000000   Alive   Alive   M
13:     12 2008.714 57.50075 0.60 0.05000000   Alive   Alive   M
14:     12 2008.764 57.55075 0.65 0.05000000   Alive   Alive   M
15:     12 2008.814 57.60075 0.70 0.05000000   Alive   Alive   M
16:     12 2008.864 57.65075 0.75 0.05000000   Alive   Alive   M
17:     12 2008.914 57.70075 0.80 0.05000000   Alive   Alive   M
18:     12 2008.964 57.75075 0.85 0.05000000   Alive   Alive   M
19:     12 2009.014 57.80075 0.90 0.05000000   Alive   Alive   M
20:     12 2009.064 57.85075 0.95 0.05000000   Alive   Alive   M
21:     12 2009.114 57.90075 1.00 0.50000000   Alive   Alive   M
22:     12 2009.614 58.40075 1.50 0.38364134   Alive   Alive   M
23:     75 2003.240 56.30390 0.00 0.05000000   Alive   Alive   M
24:     75 2003.290 56.35390 0.05 0.05000000   Alive   Alive   M
25:     75 2003.340 56.40390 0.10 0.05000000   Alive   Alive   M
26:     75 2003.390 56.45390 0.15 0.05000000   Alive   Alive   M
27:     75 2003.440 56.50390 0.20 0.05000000   Alive   Alive   M
28:     75 2003.490 56.55390 0.25 0.05000000   Alive   Alive   M
29:     75 2003.540 56.60390 0.30 0.05000000   Alive   Alive   M
30:     75 2003.590 56.65390 0.35 0.05000000   Alive   Alive   M
31:     75 2003.640 56.70390 0.40 0.05000000   Alive   Alive   M
32:     75 2003.690 56.75390 0.45 0.05000000   Alive   Alive   M
33:     75 2003.740 56.80390 0.50 0.05000000   Alive   Alive   M
34:     75 2003.790 56.85390 0.55 0.05000000   Alive   Alive   M
35:     75 2003.840 56.90390 0.60 0.05000000   Alive   Alive   M
36:     75 2003.890 56.95390 0.65 0.05000000   Alive   Alive   M
37:     75 2003.940 57.00390 0.70 0.05000000   Alive   Alive   M
38:     75 2003.990 57.05390 0.75 0.05000000   Alive   Alive   M
39:     75 2004.040 57.10390 0.80 0.05000000   Alive   Alive   M
40:     75 2004.090 57.15390 0.85 0.05000000   Alive   Alive   M
41:     75 2004.140 57.20390 0.90 0.05000000   Alive   Alive   M
42:     75 2004.190 57.25390 0.95 0.05000000   Alive   Alive   M
43:     75 2004.240 57.30390 1.00 0.02669405   Alive    Dead   M
    lex.id      per      age  tfd    lex.dur lex.Cst lex.Xst sex
```

```
> subset( S3 , lex.id %in% c(12,75) )[,1:8]
```

```
    lex.id      per      age      tfd    lex.dur lex.Cst lex.Xst sex
 1:     12 2008.114 56.90075 0.000000 0.05000000   Alive   Alive   M
```

```
 2:      12 2008.164 56.95075 0.050000 0.05000000   Alive   Alive   M
 3:      12 2008.214 57.00075 0.100000 0.05000000   Alive   Alive   M
 4:      12 2008.264 57.05075 0.150000 0.05000000   Alive   Alive   M
 5:      12 2008.314 57.10075 0.200000 0.05000000   Alive   Alive   M
 6:      12 2008.364 57.15075 0.250000 0.05000000   Alive   Alive   M
 7:      12 2008.414 57.20075 0.300000 0.05000000   Alive   Alive   M
 8:      12 2008.464 57.25075 0.350000 0.05000000   Alive   Alive   M
 9:      12 2008.514 57.30075 0.400000 0.05000000   Alive   Alive   M
10:      12 2008.564 57.35075 0.450000 0.05000000   Alive   Alive   M
11:      12 2008.614 57.40075 0.500000 0.05000000   Alive   Alive   M
12:      12 2008.664 57.45075 0.550000 0.05000000   Alive   Alive   M
13:      12 2008.714 57.50075 0.600000 0.05000000   Alive   Alive   M
14:      12 2008.764 57.55075 0.650000 0.05000000   Alive   Alive   M
15:      12 2008.814 57.60075 0.700000 0.05000000   Alive   Alive   M
16:      12 2008.864 57.65075 0.750000 0.05000000   Alive   Alive   M
17:      12 2008.914 57.70075 0.800000 0.05000000   Alive   Alive   M
18:      12 2008.964 57.75075 0.850000 0.05000000   Alive   Alive   M
19:      12 2009.014 57.80075 0.900000 0.05000000   Alive   Alive   M
20:      12 2009.064 57.85075 0.950000 0.05000000   Alive   Alive   M
21:      12 2009.114 57.90075 1.000000 0.50000000   Alive   Alive   M
22:      12 2009.614 58.40075 1.500000 0.12902122   Alive    Drug   M
23:      12 2009.743 58.52977 1.629021 0.25462012    Drug    Drug   M
24:      75 2003.240 56.30390 0.000000 0.05000000   Alive   Alive   M
25:      75 2003.290 56.35390 0.050000 0.05000000   Alive   Alive   M
26:      75 2003.340 56.40390 0.100000 0.05000000   Alive   Alive   M
27:      75 2003.390 56.45390 0.150000 0.05000000   Alive   Alive   M
28:      75 2003.440 56.50390 0.200000 0.05000000   Alive   Alive   M
29:      75 2003.490 56.55390 0.250000 0.05000000   Alive   Alive   M
30:      75 2003.540 56.60390 0.300000 0.05000000   Alive   Alive   M
31:      75 2003.590 56.65390 0.350000 0.05000000   Alive   Alive   M
32:      75 2003.640 56.70390 0.400000 0.05000000   Alive   Alive   M
33:      75 2003.690 56.75390 0.450000 0.05000000   Alive   Alive   M
34:      75 2003.740 56.80390 0.500000 0.05000000   Alive   Alive   M
35:      75 2003.790 56.85390 0.550000 0.05000000   Alive   Alive   M
36:      75 2003.840 56.90390 0.600000 0.05000000   Alive   Alive   M
37:      75 2003.890 56.95390 0.650000 0.00982204   Alive    Drug   M
38:      75 2003.900 56.96372 0.659822 0.04017796    Drug    Drug   M
39:      75 2003.940 57.00390 0.700000 0.05000000    Drug    Drug   M
40:      75 2003.990 57.05390 0.750000 0.05000000    Drug    Drug   M
41:      75 2004.040 57.10390 0.800000 0.05000000    Drug    Drug   M
42:      75 2004.090 57.15390 0.850000 0.05000000    Drug    Drug   M
43:      75 2004.140 57.20390 0.900000 0.05000000    Drug    Drug   M
44:      75 2004.190 57.25390 0.950000 0.05000000    Drug    Drug   M
45:      75 2004.240 57.30390 1.000000 0.02669405    Drug    Dead   M
    lex.id      per      age      tfd    lex.dur lex.Cst lex.Xst sex
```

The dataset now has a record for each small ($< 0.5 year$) interval of follow-up. For each interval we have the indicator, `lex.Cst` of the state in which the time (`lex.dur`) is spent, and and indicator, `lex.Xst` of the state the person moves to at the end of the follow-up.

26. We can show how the persons move between states:

```
> boxes( S3, boxpos=TRUE, scale.R=100, show.BE=TRUE )
```

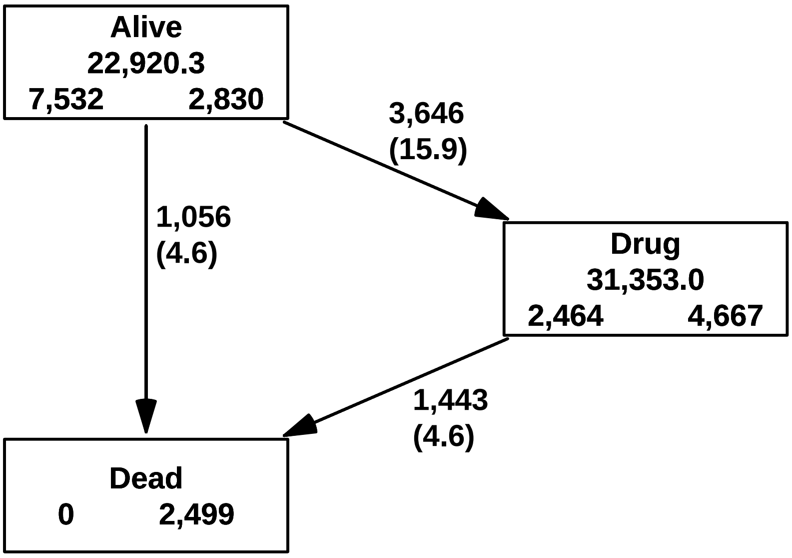


Figure 2.9: *Movement of diabetes patients between states* `./flup-box3`

`boxpos=TRUE` lets `boxes` decide where to put the boxes, `scale.R=100` scales the rates on the arrows to events per 100 PY and `show.BE=TRUE` puts the number of persons beginning, resp. ending their follow-up in each state in the boxes. The number in the middle of the boxes is the total amount of risk time spent in each state.

## 2.8 Competing risks — multiple types of events

If we want assess how long newly diagnosed diabetes patients remain without pharmaceutical treatment, we must take into account of those who die too. More precisely speaking we want to know what the probability is of being in each of the states: 1) remain alive without treatment (`Alive`) 2) being dead without any treatment 3) having initiated treatment, the latter regardless of subsequent death or not.

It is commonly seen that a traditional survival analyses are conducted where transition to `Drug` is taken as event and deaths just counted as censorings. This is wrong; it will overestimate the probability of going on drugs. The standard reference for an overview of this is [5].

There is nothing wrong with the estimate of the *rate* of initiating drugs. It is only the calculation of the survival *probability* that is wrong — the probability of having initiated a drug depends on `both` the rate of drug initiation *and* the mortality rate. That is on the rates *out* of the `Alive` box in figure 2.9.

### 2.8.1 Simple approach

The `survfit` function has a facility that correctly estimates the probabilities of being in each state. It looks like a normal survival analysis but if the last argument to the `Surv` function is a factor, a proper estimation of the probabilities of being in each state will result. It is assumed that the first level corresponding to censoring, and the remaining levels to the possible types of events.

27. In this case we restrict the follow-up to that in the `Alive` state, and the outcome factor is therefore `lex.Xst` with levels `Alive` (censoring without transition) and `Drug` and `Dead`:

```
> boxes( subset(S3,lex.Cst=="Alive"), boxpos=TRUE, scale.R=100, show.BE=TRUE )
```

28. The question addressed by this competing risk analysis the probability of starting drug treatment, and the `Drug` state here means "having been on pharmaceutical treatment, disregarding subsequent death". The other event considered is `Dead` which means "dead without initiating pharmaceutical treatment".

```
> levels(S3$lex.Xst)

[1] "Alive" "Drug"  "Dead"

> m3 <- survfit( Surv(tfd,tfd+lex.dur,lex.Xst) ~ 1,
+                data=subset(S3,lex.Cst=="Alive"), id=lex.id )
> matplot( m3$time, m3$pstate,
+          type="s", lty=1, lwd=4,
+          col=c("ForestGreen","red","black") )
```

The `m3` contains the correct probabilities of being in the `Alive` state (green), having left to the `Drug` state (red), resp. `Dead` (black). These three curves have sum 1, so basically a way of distributing the persons according to their state at each time.

29. It is therefore natural to stack the probabilities, which can be done by `stackedCIF`:

```
> par( mfrow=c(1,2) )
> plot( m3, col=c("red","black","ForestGreen"), lwd=4 )
> stackedCIF( m3, col=c("red","black") )
```

### 2.8.2 What not to do

A very common error is to use a *partial* outcome such as `Drug`, in this case there is a competing type of event, `Dead`. If that is ignored and a traditional survival analysis is made *as if* `Drug` was the only possible event, we will have a substantial *over*estimate of the cumulative probability of going on drug as illustrated in this analysis:

```
> par( mfrow=c(1,2) )
> par( mfrow=c(1,2) )
> rbt <- c("red","black","transparent")
> mat2pol( m3$pstate, c(2,3,1), x=m3$time, col=rbt )
> mat2pol( m3$pstate, c(3,2,1), x=m3$time, col=rbt[c(2,1,3)] )
```

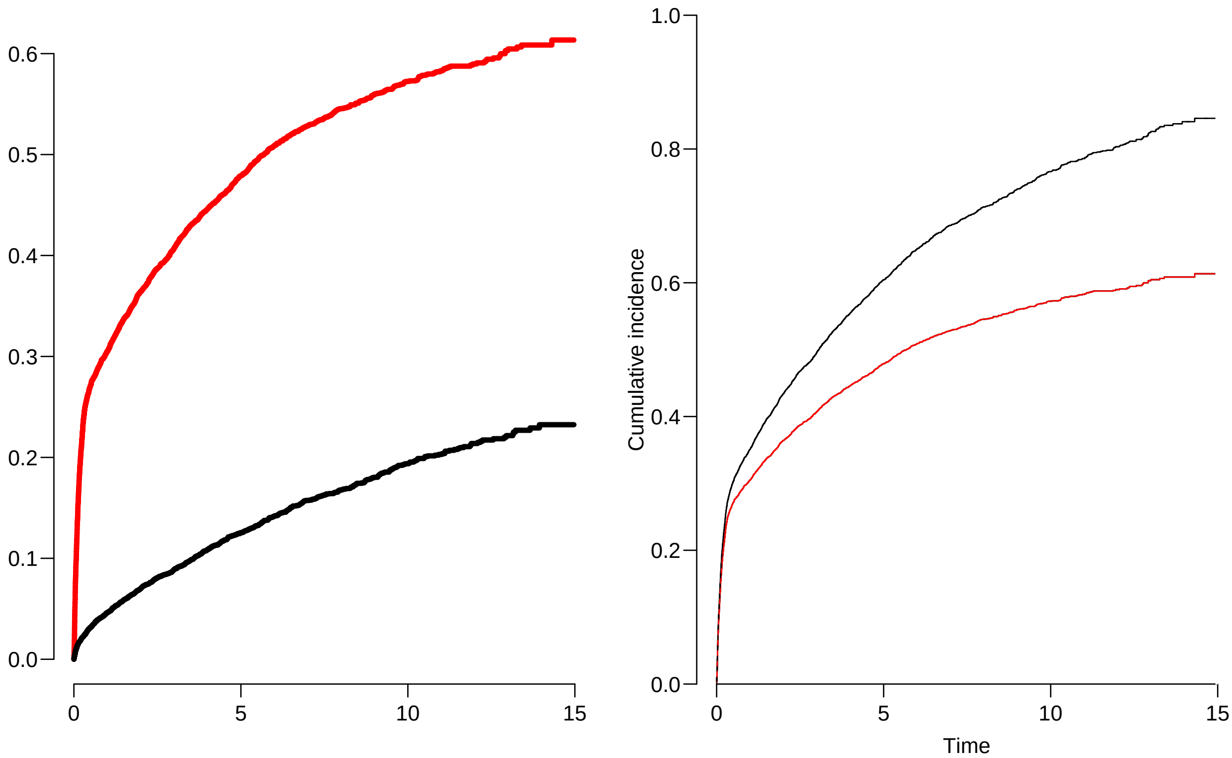


Figure 2.10: *Separate state probabilities (left) and stacked state probabilities (right).* `Alive` *is green (white in right plot),* `Drug` *is red and* `Dead` *is black.* ./flup-surv2

```
> par( mfrow=c(1,2) )
> # Wrong analysis of Drug start:
> m2 <- survfit( Surv(tfd,tfd+lex.dur,lex.Xst=="Drug") ~ 1,
+                data=subset(S3,lex.Cst=="Alive") )
> # Wrong analysis of Death prob:
> M2 <- survfit( Surv(tfd,tfd+lex.dur,lex.Xst=="Dead") ~ 1,
+                data=subset(S3,lex.Cst=="Alive") )
> # Compare the wrong analyses with the correct one:
> par( mfrow=c(1,2) )
> mat2pol( m3$pstate, c(2,3,1), x=m3$time, col=c("red","black","transparent") )
>   lines( m2$time, 1-m2$surv, lwd=3, col="red" )
> mat2pol( m3$pstate, c(3,2,1), x=m3$time, col=c("black","red","transparent") )
>   lines( M2$time, 1-M2$surv, lwd=3, col="black" )
```

From figure 2.11 we see that by ignoring the possibility of dying we will be substantially overestimating the probability of going on pharmaceutical treatment. The red curve is based on what is often termed the 'net' probability—it refers to the probability of `Drug` in a setting where 1) death does not occur and 2) the rate of pharmaceutical treatment is still the same.

The question 1) is often asked in competing risk situations, and its is believed to be answered by the cyan curve. But the highly unrealistic assumption 2) is often forgotten. How the rate of pharmaceutical treatment would look in the absence of death (or vice versa for that matter) cannot be deduced from data where both types of risk is present. It is essentially a theological question.

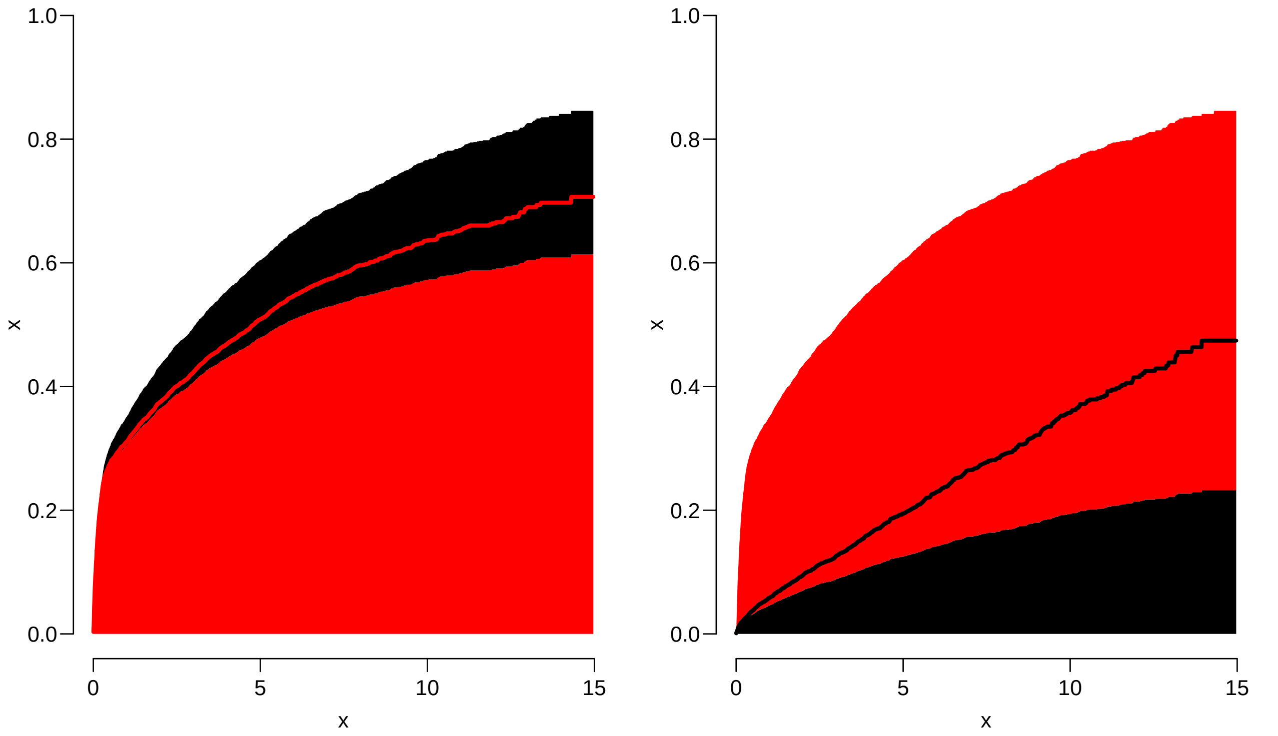


Figure 2.11: *Stacked state probabilities;* `Alive` *is white ,* `Drug` *is red and* `Dead` *is black. The red line in the left panel is the wrong (but often computed) "cumulative risk" of* `Drug`*, and the black line in the right panel is the wrong (but often computed) "cumulative risk" of* `Death`*. The black and the red areas in the two plots are of identical size at any time, only they are stacked differently.* `./flup-surv3`

### 2.8.3 A mathematical explanation

Suppose the rate of drug initiation (`Alive`→`Drug`) is $\lambda(t)$ and the mortality before drug initiation (`Alive`→`Dead`) is $\mu(t)$, then the probability of being alive without drug treatment at time $t$ is:

$$S(t) = \exp\left(-\int_0^t \lambda(s) + \mu(s)\,\mathrm{d}s\right)$$

and the cumulative risk of initiating drug before time $t$ is:

$$R_{\texttt{Drug}}(t) = \int_0^t \lambda(u) S(u)\,\mathrm{d}u = \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s) + \mu(s)\,\mathrm{d}s\right)\mathrm{d}u \qquad (2.1)$$

—and similarly for cumulative risk of death.

The error committed in the analysis pretending that only the event `Drug` is present is *not* in the calculations of the cause-specific rates, it is *only* in the calculations of the cumulative risk (probability of transition to `Drug`). The red line in figure 2.11 comes from omitting the green term $\mu(s)$ from formula 2.1. The temptation is apparently that if you do that the mathematics becomes nicer:

$$R_{\texttt{Drug}}(t) = \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s)\,\mathrm{d}s\right)\mathrm{d}u = 1 - \exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right) \qquad (2.2)$$

and this is precisely what comes out of standard programs when regarding `Drug` as the only type of event.

So there is no such thing as a competing risks analysis of event *rates*; the competing risks aspect comes about only when you want to address cumulative risk of a particular event. In which case you probably want to look at cumulative risks of all types of events.

## 2.9 Modeling cause specific events

As we just saw, there is nothing wrong with modeling the cause-specific event-rates, the problem lies in transforming them into probabilities.

As above we can model the two sets of rates by a parametric model:

```
> gl <- gam.Lexis( S3, ~ s(tfd), from="Alive", to="Drug" )
mgcv::gam Poisson analysis of Lexis object S3 with log link:
Rates for the transition: Alive->Drug
> gm <- gam.Lexis( S3, ~ s(tfd), from="Alive", to="Dead" )

mgcv::gam Poisson analysis of Lexis object S3 with log link:
Rates for the transition: Alive->Dead
```

We can derive the estimated rates by time by using prediction frames:

```
> int <- 0.01
> nd <- data.frame( tfd=seq(int,15,int)-int/2 )
> mrt <- ci.pred( gm, nd )[,1]
> lam <- ci.pred( gl, nd )[,1]
```

The vectors `mrt` and `lam` now contain the two rates evaluated at the midpoint of intervals of length `int`=0.01 years. Since the variable `lex.dur` is in units of years, the rates are in units of events per 1 person-year.

We can the translate the integrals above directly into computer code, using the fact that an integral is just a sum, and since we want the integrals as functions of the upper limits we use `cumsum`, remembering t multiply by the interval length:

```
> Lam <- cumsum( lam*int )          # cumulative incidence of Drug
> Mrt <- cumsum( mrt*int )          # cumulative mortality
> Srv <- exp( -(Mrt+Lam) )            # survival in ALive
> Rdr <- cumsum( lam * Srv * int ) # cumulative risk of Drug
> Rdt <- cumsum( mrt * Srv * int ) # cumulative risk of Death
```

Now we have the survival (in `Alive`) in `Srv` and the cumulative risks in `Rdr` and `Rdt`.

```
> plot( m3, col=c("red","black","ForestGreen"), lwd=1 )
> matlines( nd$tfd, cbind( Srv, Rdr, Rdt ),
+           type="l", lty="21", lwd=3, lend="butt",
+           col=c("ForestGreen","red","black") )
```

From figure 2.12 we see that the results from the two approaches are pretty much the same, the smoothed curve gives a bit higher value for the probability of being on drug around 1 year.

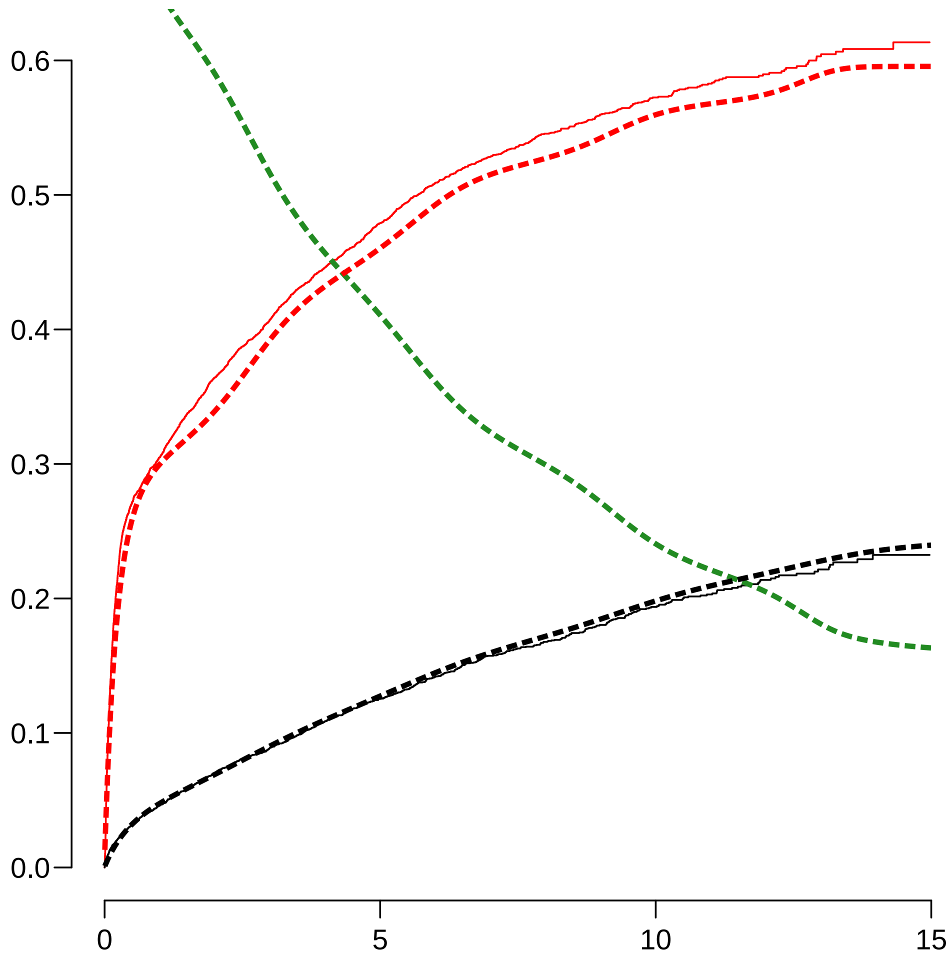


Figure 2.12: *Comparison of the non-parametric estimates of the state probabilities (thin lines) with the probabilities based on* `gam`*-smoothed rates (thick broken lines).* `./flup-comp`

### 2.9.1 Limitations

The parametric approach produces more credible estimates then does the non-parametric approach, but as it is seen here, the approach relies on the availability of explicit formulae. In practice this means that competing risk models can be treated, but that more complicated multistate models cannot be reported using this approach.

Neither of the approaches here will deal with models where rates depend on multiple time scales, particularly on timescales defined as time since entry to an intermediate state (such as mortality depending on the time since drug initiation or on the time between diagnosis and drug initiation). This type of results can essentially only be derived using simulation, you may want to consult the so-called vignette on this in the `Epi` package:

```
> vignette("simLexis","Epi")
```

### 2.9.2 Further material

On my website is a page called “Modern demographic methods in epidemiology”, see `http://bendixcarstensen.com/AdvCoh/`. It contains a number of links to other course material and to a number of reports and background papers that may have your interest.

# Chapter 3

# Basic concepts in survival and demography

The following is a condensed overview of concepts central to handling follow-up and survival data; the target audience for this section is

- epidemiologists who wants a handy overview of the mathematical relationships between the theoretical concepts
- statisticians (and probabilists, mathematicians) who want to get an overview of how the various concepts in probability translates to epidemiological concepts

It is not an integra part of the ciurse stream; it is meant a help for future reference. The following is a summary of relations between various quantities used in analysis of follow-up studies. They are ubiquitous in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

## 3.1 Probability

**Survival function:**

$$\begin{aligned} S(t) &= \mathrm{P}\{\text{survival at least till } t\} \\ &= \mathrm{P}\{T > t\} = 1 - \mathrm{P}\{T \leq t\} = 1 - F(t) \end{aligned}$$

where $T$ is the variable "time of death"

**Conditional survival function:**

$$\begin{aligned} S(t|t_{\text{entry}}) &= \mathrm{P}\{\text{survival at least till } t| \text{ alive at } t_{\text{entry}}\} \\ &= S(t)/S(t_{\text{entry}}) \end{aligned}$$

**Cumulative distribution function** of death times (cumulative risk):

$$\begin{aligned} F(t) &= \mathrm{P}\{\text{death before } t\} \\ &= \mathrm{P}\{T \leq t\} = 1 - S(t) \end{aligned}$$

**Density function** of death times:

$$f(t) = \lim_{h\to 0} \mathrm{P}\{\text{death in } (t, t+h)\}/h = \lim_{h\to 0} \frac{F(t+h)-F(t)}{h} = F'(t)$$

**Intensity:**

$$\begin{aligned}
\lambda(t) &= \lim_{h\to 0} \mathrm{P}\{\text{event in } (t, t+h] \mid \text{alive at } t\}/h \\
&= \lim_{h\to 0} \frac{F(t+h)-F(t)}{S(t)h} = \frac{f(t)}{S(t)} \\
&= \lim_{h\to 0} -\frac{S(t+h)-S(t)}{S(t)h} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}
\end{aligned}$$

The intensity is also known as the hazard function, hazard rate, mortality/morbidity rate or simply "rate".

Note that $f$ and $\lambda$ are *scaled* quantities, they have dimension time$^{-1}$.

**Relationships** between terms:

$$\begin{aligned}
-\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\
&\Updownarrow \\
S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\big(-\Lambda(t)\big)
\end{aligned}$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the *integrated intensity* or the **cumulative rate**. It is *not* an intensity (rate), it is dimensionless, despite its name.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

**The cumulative *risk*** of an event (to time $t$) is:

$$F(t) = \mathrm{P}\{\text{Event before time } t\} = \int_0^t \lambda(u)S(u)\,\mathrm{d}u = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ ($< 0.05$), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

## 3.2 Statistics

**Likelihood** contribution from follow up of one person:

The likelihood from a number of small pieces of follow-up from one individual is a product of conditional probabilities:

$$
\begin{aligned}
\mathrm{P}\{\text{event at } t_4|\text{entry at } t_0\} &= \mathrm{P}\{\text{survive } (t_0,t_1)|\text{ alive at } t_0\}\times\\
&\quad \mathrm{P}\{\text{survive } (t_1,t_2)|\text{ alive at } t_1\}\times\\
&\quad \mathrm{P}\{\text{survive } (t_2,t_3)|\text{ alive at } t_2\}\times\\
&\quad \mathrm{P}\{\text{event at } t_4|\text{ alive at } t_3\}
\end{aligned}
$$

Each term in this expression corresponds to one *empirical rate*[1] $(d, y) = (\#\text{deaths}, \#\text{risk time})$, i.e. the data obtained from the follow-up of one person in the interval of length $y$. Each person can contribute many empirical rates, most with $d = 0$; $d$ can only be 1 for the *last* empirical rate for a person.

**Log-likelihood** for one empirical rate $(d, y)$:

$$\ell(\lambda) = \log\big(\mathrm{P}\{d \text{ events in } y \text{ follow-up time}\}\big) = d\log(\lambda) - \lambda y$$

This is under the assumption that the rate ($\lambda$) is constant over the interval that the empirical rate refers to.

**Log-likelihood for several persons.** Adding log-likelihoods from a group of persons (only contributions with identical rates) gives:

$$D\log(\lambda) - \lambda Y,$$

where $Y$ is the total follow-up time ($Y = \sum_i y_i$), and $D$ is the total number of failures ($D = \sum_i d_i$), where the sums are over individuals' contributions with the *same* rate, $\lambda$, for example from the same age-class fro all individuals.

Note: The Poisson log-likelihood for an observation $D$ with mean $\lambda Y$ is:

$$D\log(\lambda Y) - \lambda Y = D\log(\lambda) + D\log(Y) - \lambda Y$$

The term $D\log(Y)$ does not involve the parameter $\lambda$, so the likelihood for an observed rate $(D, Y)$ can be maximized by pretending that the no. of cases $D$ is Poisson with mean $\lambda Y$. But this does *not* imply that $D$ follows a Poisson-distribution. It is entirely a likelihood based computational convenience. Anything that is not likelihood based is not justified.

**A linear model** for the log-rate, $\log(\lambda) = X\beta$ implies that

$$\lambda Y = \exp\big(\log(\lambda) + \log(Y)\big) = \exp\big(X\beta + \log(Y)\big)$$

Therefore, in order to get a linear model for $\log(\lambda)$ we must require that $\log(Y)$ appear as a variable in the model for $D \sim (\lambda Y)$ with the regression coefficient fixed to 1, a so-called *offset*-term in the linear predictor.

[1] This is a concept coined by BxC, and so is not necessarily generally recognized.

## 3.3 Competing risks

**Competing risks:** If there are more than one, say 3, causes of death, occurring with (cause-specific) rates $\lambda_1$, $\lambda_2$, $\lambda_3$, that is:

$$\lambda_c(a) = \lim_{h \to 0} \mathrm{P}\{\text{death from cause } c \text{ in } (a, a+h] \mid \text{alive at } a\} / h, \quad c = 1, 2, 3$$

The survival function is then:

$$S(a) = \exp\left(-\int_0^a \lambda_1(u) + \lambda_2(u) + \lambda_3(u)\,\mathrm{d}u\right)$$

because you have to escape all 3 causes of death. The probability of dying from cause 1 before age $a$ (the cause-specific cumulative risk) is:

$$F_1(a) = \mathrm{P}\{\text{dead from cause 1 at } a\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u \neq 1 - \exp\left(-\int_0^a \lambda_1(u)\,\mathrm{d}u\right)$$

The term $\exp(-\int_0^a \lambda_1(u)\,\mathrm{d}u)$ is sometimes referred to as the "cause-specific survival", but it does not have any probabilistic interpretation in the real world. It is the survival under the assumption that only cause 1 existed and that the mortality rate from this cause was the same as when the other causes were present too.

Together with the survival function, the cause-specific cumulative risks represent a classification of the population at any time in those alive and those dead from causes 1, 2 and 3 respectively:

$$1 = S(a) + \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_2(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_3(u)S(u)\,\mathrm{d}u, \quad \forall a$$

**Subdistribution hazard** Fine and Gray defined models for the so-called subdistribution hazard, $\tilde{\lambda}_i(a)$. Recall the relationship between between the hazard ($\lambda$) and the cumulative risk ($F$):

$$\lambda(a) = -\frac{\mathrm{d}\log\big(S(a)\big)}{\mathrm{d}a} = -\frac{\mathrm{d}\log\big(1 - F(a)\big)}{\mathrm{d}a}$$

When more competing causes of death are present the Fine and Gray idea is to use this transformation to the cause-specific cumulative risk for cause 1, say:

$$\tilde{\lambda}_1(a) = -\frac{\mathrm{d}\log\big(1 - F_1(a)\big)}{\mathrm{d}a}$$

Here, $\tilde{\lambda}_1$ is called the subdistribution hazard; as a function of $F_1(a)$ it depends on the survival function $S$, which depends on *all* the cause-specific hazards:

$$F_1(a) = \mathrm{P}\{\text{dead from cause 1 at } a\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u$$

The subdistribution hazard is merely a transformation of the cause-specific cumulative risk. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard. It is a mathematical construct that is not interpretable as a hazard despite its name.

## 3.4 Demography

**Expected residual lifetime:** The expected lifetime (at birth) is simply the variable age ($a$) integrated with respect to the distribution of age at death:

$$\mathrm{EL} = \int_0^\infty a f(a)\,\mathrm{d}a$$

where $f$ is the density of the distribution of lifetime (age at death).

The relation between the density $f$ and the survival function $S$ is $f(a) = -S'(a)$, so integration by parts gives:

$$\mathrm{EL} = \int_0^\infty a\big(-S'(a)\big)\,\mathrm{d}a = -\Big[aS(a)\Big]_0^\infty + \int_0^\infty S(a)\,\mathrm{d}a$$

The first of the resulting terms is 0 because $S(a)$ is 0 at the upper limit and $a$ by definition is 0 at the lower limit.

Hence the expected lifetime can be computed as the integral of the survival function.

The expected *residual* lifetime at age $a$ is calculated as the integral of the *conditional* survival function for a person aged $a$:

$$\mathrm{EL}(a) = \int_a^\infty S(u)/S(a)\,\mathrm{d}u$$

**Lifetime lost** due to a disease is the difference between the expected residual lifetime for a diseased person and a non-diseased (well) person at the same age. So all that is needed is a(n estimate of the) survival function in each of the two groups.

$$\mathrm{LL}(a) = \int_a^\infty S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(u)/S_{\mathrm{Diseased}}(a)\,\mathrm{d}u$$

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a decision required in order to devise an estimate of $S_{\mathrm{Well}}$.

**Lifetime lost by cause of death** is using the fact that the difference between the survival probabilities is the same as the difference between the death probabilities. If several causes of death (3, say) are considered then:

$$\begin{aligned} S(a) = 1 &- \mathrm{P}\{\text{dead from cause 1 at } a\} \\ &- \mathrm{P}\{\text{dead from cause 2 at } a\} \\ &- \mathrm{P}\{\text{dead from cause 3 at } a\} \end{aligned}$$

and hence:

$$\begin{aligned} S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(a) &= \mathrm{P}\{\text{dead from cause 1 at } a|\text{Diseased}\} \\ &+ \mathrm{P}\{\text{dead from cause 2 at } a|\text{Diseased}\} \\ &+ \mathrm{P}\{\text{dead from cause 3 at } a|\text{Diseased}\} \\ &- \mathrm{P}\{\text{dead from cause 1 at } a|\text{Well}\} \\ &- \mathrm{P}\{\text{dead from cause 2 at } a|\text{Well}\} \\ &- \mathrm{P}\{\text{dead from cause 3 at } a|\text{Well}\} \end{aligned}$$

So we can conveniently define the lifetime lost due to cause 2, say, by:

$$\begin{aligned}\mathrm{LL}_2(a) = \int_a^\infty & \mathrm{P}\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\} \\ & - \mathrm{P}\{\text{dead from cause 2 at } u|\text{Well \& alive at } a\}\, \mathrm{d}u\end{aligned}$$

These quantities have the property that their sum is the total years of life lost due to the disease:

$$\mathrm{LL}(a) = \mathrm{LL}_1(a) + \mathrm{LL}_2(a) + \mathrm{LL}_3(a)$$

The terms in the integral are computed as (see the section on competing risks):

$$\begin{aligned}\mathrm{P}\{\text{dead from cause 2 at } x|\text{Diseased \& alive at } a\} &= \int_a^x \lambda_{2,\mathrm{Dis}}(u) S_{\mathrm{Dis}}(u)/S_{\mathrm{Dis}}(a)\,\mathrm{d}u \\ \mathrm{P}\{\text{dead from cause 2 at } x|\text{Well \& alive at } a\} &= \int_a^x \lambda_{2,\mathrm{Well}}(u) S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a)\,\mathrm{d}u\end{aligned}$$

# References

[1] Martyn Plummer and Bendix Carstensen. Lexis: An R class for epidemiological studies with long-term follow-up. *Journal of Statistical Software*, 38(5):1–12, 1 2011.

[2] Bendix Carstensen and Martyn Plummer. Using Lexis objects for multi-state models in R. *Journal of Statistical Software*, 38(6):1–18, 1 2011.

[3] B Carstensen. Age-Period-Cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15):3018–3045, July 2007.

[4] L. Huo, D. J. Magliano, F. Ranciere, J. L. Harding, N. Nanayakkara, J. E. Shaw, and B. Carstensen. Impact of age at diagnosis and duration of type 2 diabetes on mortality in Australia 1997-2011. *Diabetologia*, Feb 2018.

[5] P. K. Andersen, R. B. Geskus, T. de Witte, and H. Putter. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*, Jan 2012.