

Practice in analysis of multistate models using Epi::Lexis

Bendix Carstensen Steno Diabetes Center,
Gentofte, Denmark
& Department of Biostatistics,
University of Copenhagen
bxc@steno.dk
<http://BendixCarstensen.com>

FRIAS, Freiburg, 21 September 2016

<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

Rates and Survival

Bendix Carstensen

Senior Statistician, Steno Diabetes Center

Practice in analysis of multistate models using Epi::Lexis

21 September 2016

FRIAS, Freiburg

<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

Survival data

Persons enter the study at some date.

Persons exit at a later date, either dead or alive.

Observation:

Actual time span to death (“event”)

or

Some time alive (“at least this long”)

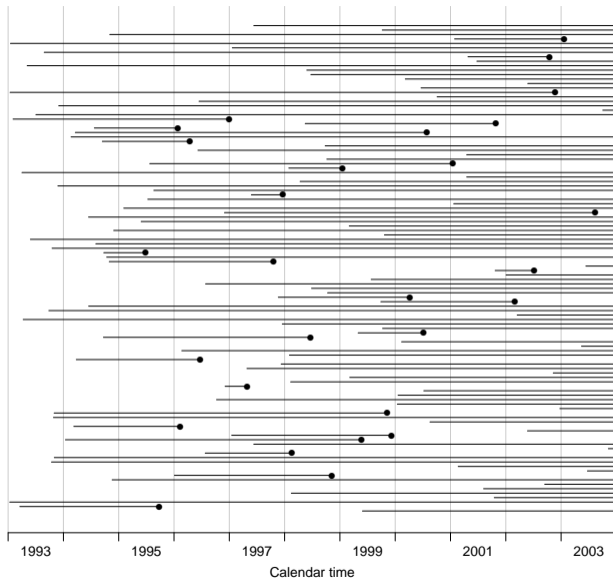
Examples of time-to-event measurements

- ▶ Time from diagnosis of cancer to death.
- ▶ Time from randomisation to death in a cancer clinical trial
- ▶ Time from HIV infection to AIDS.
- ▶ Time from marriage to 1st child birth.
- ▶ Time from marriage to divorce.
- ▶ Time to re-offending after being released from jail

Each line a
person

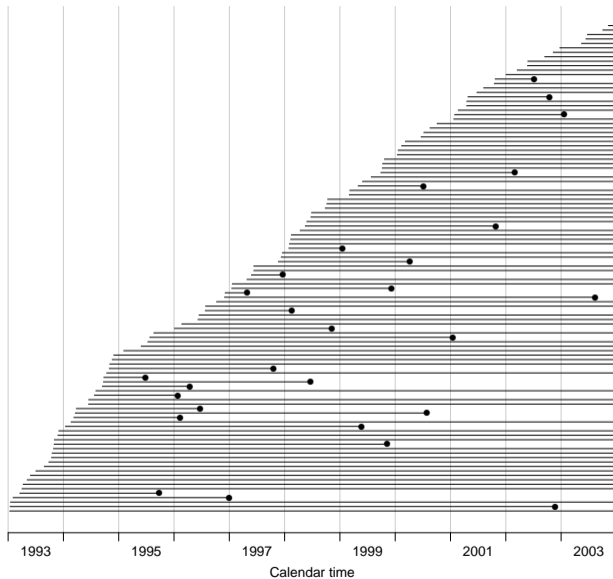
Each blob a
death

Study ended at
31 Dec. 2003

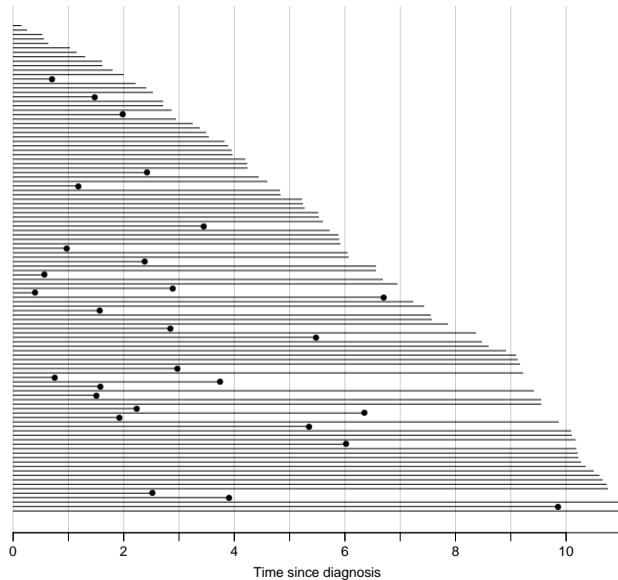


Ordered by date
of entry

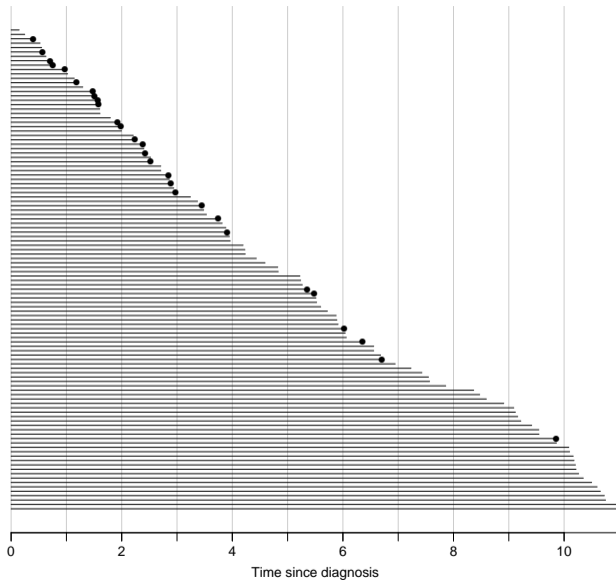
Most likely the
order in your
database.



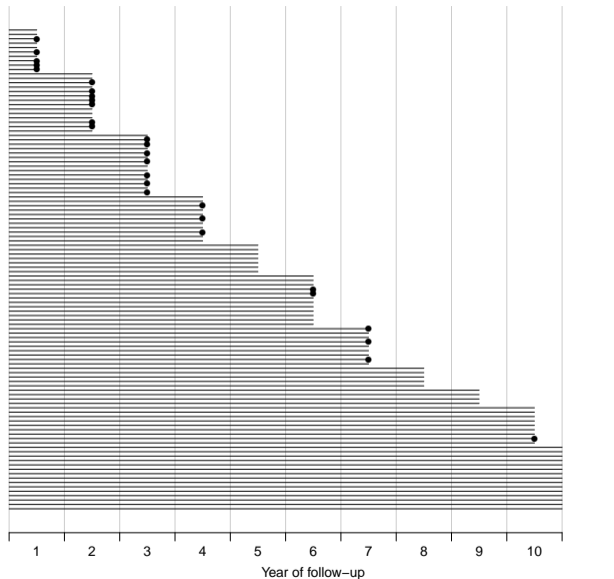
Timescale
changed to
“Time since
diagnosis”.



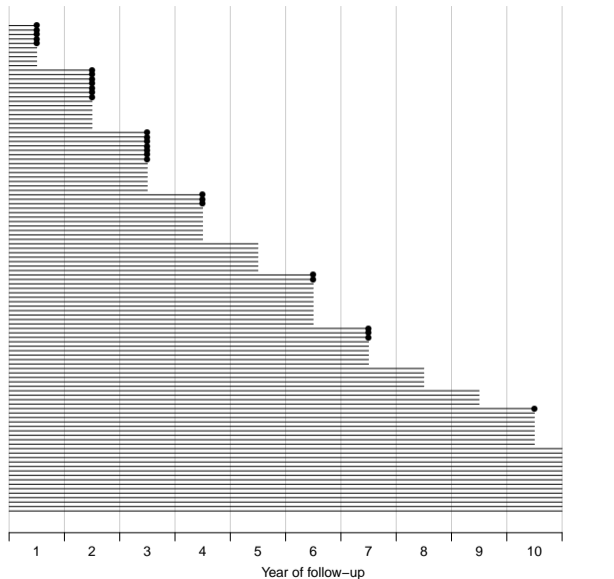
Patients ordered
by survival time.



Survival times grouped into bands of survival.



Patients ordered
by survival
status within
each band.



Survival after Cervix cancer

Year	Stage I			Stage II		
	<i>N</i>	<i>D</i>	<i>L</i>	<i>N</i>	<i>D</i>	<i>L</i>
1	110	5	5	234	24	3
2	100	7	7	207	27	11
3	86	7	7	169	31	9
4	72	3	8	129	17	7
5	61	0	7	105	7	13
6	54	2	10	85	6	6
7	42	3	6	73	5	6
8	33	0	5	62	3	10
9	28	0	4	49	2	13
10	24	1	8	34	4	6

Estimated risk in year 1 for Stage I women is $5/107.5 = 0.0465$

Estimated 1 year survival is $1 - 0.0465 = 0.9535$

Life-table estimator.

Survival function

Persons enter at time 0:

Date of birth, date of randomization, date of diagnosis.

How long do they survive?

Survival time T — a stochastic variable.

Distribution is characterized by the survival function:

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

$F(t)$ is the cumulative risk of death before time t .

Intensity or rate

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

This is the **intensity** or **hazard function** for the distribution. Characterizes the survival distribution as does f or F .

Theoretical counterpart of a **rate**.

Relationships

$$-\frac{d\log S(t)}{dt} = \lambda(t)$$

\Updownarrow

$$S(t) = \exp\left(-\int_0^t \lambda(u) du\right) = \exp(-\Lambda(t))$$

$\Lambda(t) = \int_0^t \lambda(s) ds$ is called the **integrated intensity**. **Not** an intensity, it is dimensionless.

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

Rate and survival

$$S(t) = \exp\left(-\int_0^t \lambda(s) ds\right) \quad \lambda(t) = \frac{S'(t)}{S(t)}$$

Survival is a *cumulative* measure, the rate is an *instantaneous* measure.

Note: A cumulative measure requires an origin!
... it is always survival **since** some timepoint.

Observed survival and rate

- ▶ **Survival studies:** Observation of (right censored) survival time:

$$X = \min(T, Z), \quad \delta = 1\{X = T\}$$

— sometimes conditional on $T > t_0$
(left truncation, delayed entry).

- ▶ **Epidemiological studies:**
Observation of (components of) a rate:

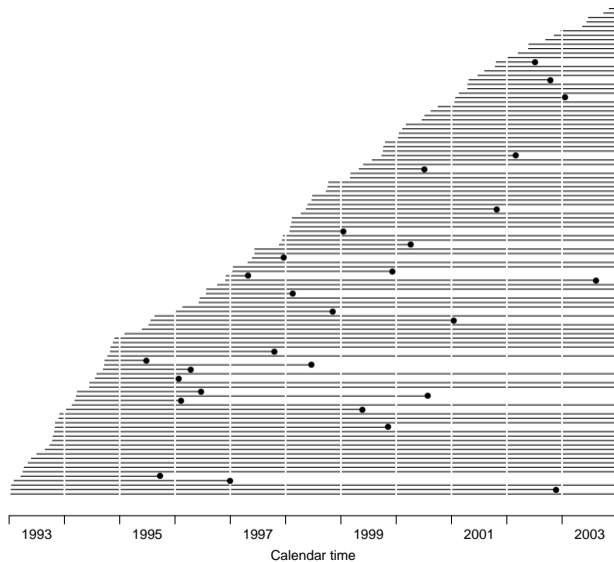
$$D/Y$$

D : no. events, Y no of person-years, in a prespecified time-frame.

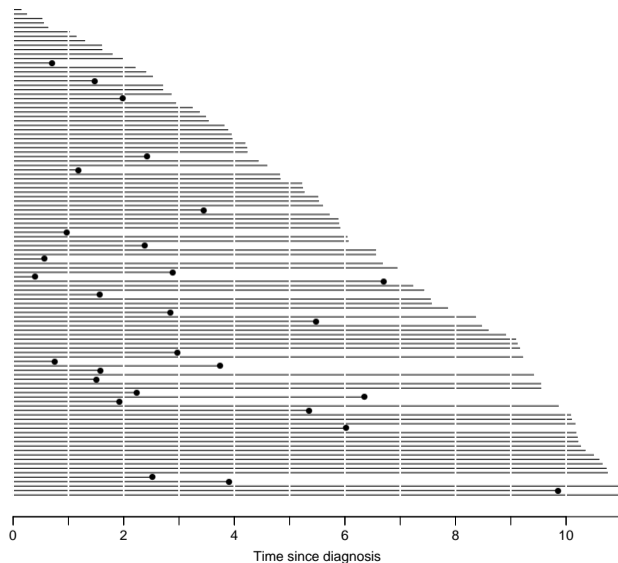
Empirical rates for individuals

- ▶ At the *individual* level we introduce the **empirical rate**: (d, y) ,
— number of events ($d \in \{0, 1\}$) during y risk time.
- ▶ A person contributes several observations of (d, y) , with associated covariate values.
- ▶ Empirical rates are **responses** in survival analysis.
- ▶ The timescale t is a **covariate** — varies within each individual:
 t : age, time since diagnosis, calendar time.
- ▶ Don't confuse with y — difference between two points on **any** timescale we may choose.

Empirical rates
by
calendar time.



Empirical rates
by
time since
diagnosis.



Statistical inference: Likelihood

Two things needed:

- ▶ **Data** — what did we actually observe
Follow-up for each person:
Entry time, exit time, exit status, covariates
- ▶ **Model** — how was data generated
Rates as a function of time:
Probability machinery that generated data

Likelihood is the probability of observing the **data**, assuming the **model** is correct.

Maximum likelihood estimation is choosing **parameters** of the model that makes the likelihood maximal.

Likelihood from one person

The likelihood from several empirical rates from one individual is a product of conditional probabilities:

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

Log-likelihood from one individual is a sum of terms.

Each term refers to one empirical rate (d, y)

— $y = t_i - t_{i-1}$ and mostly $d = 0$.

t_i is the timescale (covariate).

Poisson likelihood

The log-likelihood contributions from follow-up of **one** individual:

$$d_t \log(\lambda(t)) - \lambda(t) y_t, \quad t = t_1, \dots, t_n$$

is also the log-likelihood from several independent Poisson observations with mean $\lambda(t) y_t$, i.e. log-mean $\log(\lambda(t)) + \log(y_t)$

Analysis of the rates, (λ) can be based on a Poisson model with log-link applied to empirical rates where:

- ▶ d is the response variable.
- ▶ $\log(\lambda)$ is modelled by covariates
- ▶ $\log(y)$ is the offset variable.

Likelihood for follow-up of many persons

Adding empirical rates over the follow-up of persons:

$$D = \sum d \quad Y = \sum y \quad \Rightarrow \quad D \log(\lambda) - \lambda Y$$

- ▶ Persons are assumed independent
- ▶ Contribution from the same person are **conditionally** independent, hence give separate contributions to the log-likelihood.
- ▶ Therefore equivalent to likelihood for independent Poisson variates
- ▶ No need to correct for dependent observations; the likelihood is a product.

Likelihood

Probability of the data and the parameter:

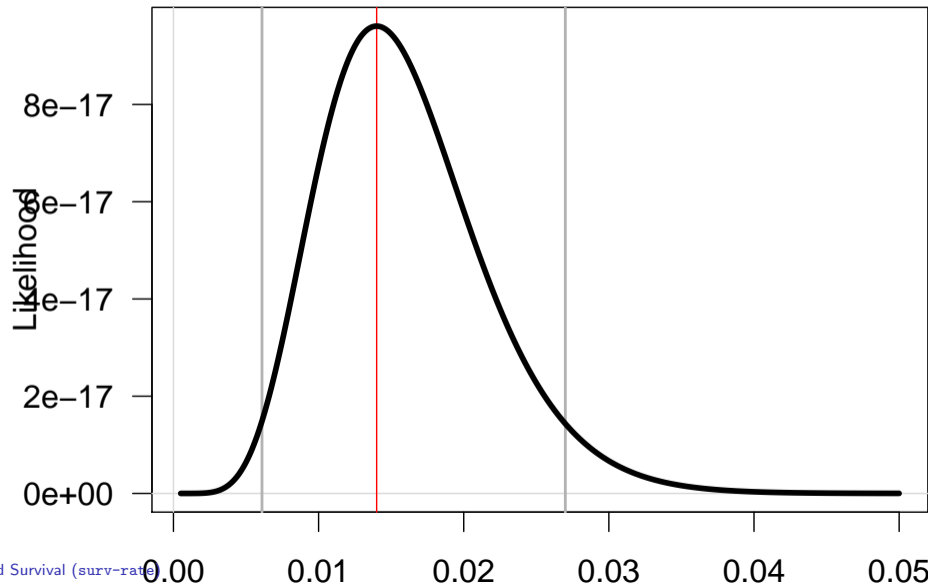
Assuming the rate (intensity) is constant, λ , the probability of observing 7 deaths in the course of 500 person-years:

$$\begin{aligned}P\{D = 7, Y = 500|\lambda\} &= \lambda^D e^{-\lambda Y} \times K \\ &= \lambda^7 e^{-\lambda 500} \times K \\ &= L(\lambda|\text{data})\end{aligned}$$

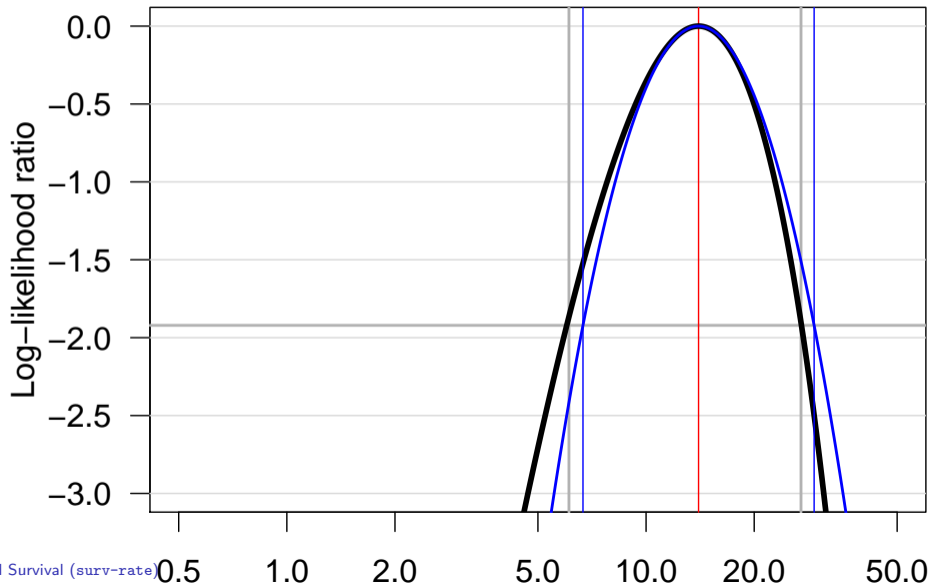
Best guess of λ is where this function is as large as possible.

Confidence interval is where it is not too far from the maximum

Likelihood function



Likelihood function



Confidence interval for a rate

A 95% confidence interval for the log of a rate is:

$$\hat{\theta} \pm 1.96/\sqrt{D} = \log(\lambda) \pm 1.96/\sqrt{D}$$

Take the exponential to get the confidence interval for the rate:

$$\lambda \times \underbrace{\exp(1.96/\sqrt{D})}_{\text{error factor, erf}}$$

Example

Suppose we have 17 deaths during 843.6 years of follow-up.
The rate is computed as:

$$\hat{\lambda} = D/Y = 17/843.7 = 0.0201 = 20.1 \text{ per 1000 years}$$

The confidence interval is computed as:

$$\hat{\lambda} \times \text{erf} = 20.1 \times \exp(1.96/\sqrt{D}) = (12.5, 32.4)$$

per 1000 person-years.

Ratio of two rates

If we have observations two rates λ_1 and λ_0 , based on (D_1, Y_1) and (D_0, Y_0) , the variance of the difference of the log-rates, the $\log(\text{RR})$, is:

$$\begin{aligned}\text{var}(\log(\text{RR})) &= \text{var}(\log(\lambda_1/\lambda_0)) \\ &= \text{var}(\log(\lambda_1)) + \text{var}(\log(\lambda_0)) \\ &= 1/D_1 + 1/D_0\end{aligned}$$

As before a 95% c.i. for the RR is then:

$$\text{RR} \times \underbrace{\exp\left(1.96\sqrt{\frac{1}{D_1} + \frac{1}{D_0}}\right)}_{\text{error factor}}$$

Example

Suppose we in group 0 have 17 deaths during 843.6 years of follow-up in one group, and in group 1 have 28 deaths during 632.3 years.

The rate-ratio is computed as:

$$\begin{aligned} \text{RR} &= \hat{\lambda}_1 / \hat{\lambda}_0 = (D_1 / Y_1) / (D_0 / Y_0) \\ &= (28 / 632.3) / (17 / 843.7) = 0.0443 / 0.0201 = 2.198 \end{aligned}$$

The 95% confidence interval is computed as:

$$\begin{aligned} \hat{\text{RR}} \times_{\div} \text{erf} &= 2.198 \times_{\div} \exp(1.96 \sqrt{1/17 + 1/28}) \\ &= 2.198 \times_{\div} 1.837 = (1.20, 4.02) \end{aligned}$$

Example using R

Poisson likelihood, for one rate,
based on 17 events in 843.7 PY:

```
library( Epi )  
D <- 17 ; Y <- 843.7  
m1 <- glm( D ~ 1, offset=log(Y/1000), family=poisson)  
ci.exp( m1 )
```

```
                exp(Est.)      2.5%      97.5%  
(Intercept)  20.14934 12.52605 32.41213
```

Poisson likelihood, two rates, or one rate and RR:

```
D <- c(17,28) ; Y <- c(843.7,632.3) ; gg <- factor(0:1)  
m2 <- glm( D ~ gg, offset=log(Y/1000), family=poisson)  
ci.exp( m2 )
```

```
                exp(Est.)      2.5%      97.5%  
(Intercept)  20.149342 12.526051 32.412130  
gg1          2.197728  1.202971  4.015068
```

Example using R

Poisson likelihood, two rates, or one rate and RR:

```
D <- c(17,28) ; Y <- c(843.7,632.3) ; gg <- factor(0:1)
m2 <- glm( D ~ gg, offset=log(Y/1000), family=poisson)
ci.exp( m2 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	20.149342	12.526051	32.412130
gg1	2.197728	1.202971	4.015068

```
m3 <- glm( D ~ gg - 1, offset=log(Y/1000), family=poisson)
ci.exp( m3 )
```

	exp(Est.)	2.5%	97.5%
gg0	20.14934	12.52605	32.41213
gg1	44.28278	30.57545	64.13525

Representation of follow-up data

Bendix Carstensen

Senior Statistician, Steno Diabetes Center

Practice in analysis of multistate models using Epi::Lexis

21 September 2016

FRIAS, Freiburg

<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

Follow-up and rates

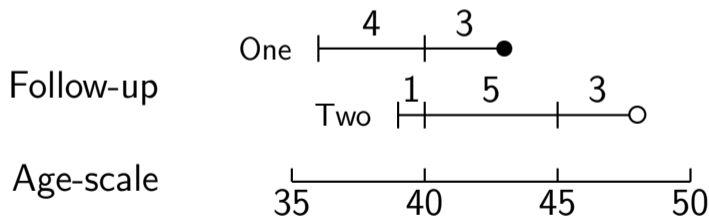
- ▶ Follow-up studies:
 - ▶ D — events, deaths
 - ▶ Y — person-years
 - ▶ $\lambda = D/Y$ rates
- ▶ Rates differ between persons.
- ▶ Rates differ **within** persons:
 - ▶ By age
 - ▶ By calendar time
 - ▶ By disease duration
 - ▶ ...
- ▶ Multiple timescales.
- ▶ Multiple states (little boxes — later)

Stratification by age

If follow-up is rather short, age at entry is OK for age-stratification.

If follow-up is long, use stratification by categories of **current age**, both for:

No. of events, D , and Risk time, Y .



Representation of follow-up data

A cohort or follow-up study records:

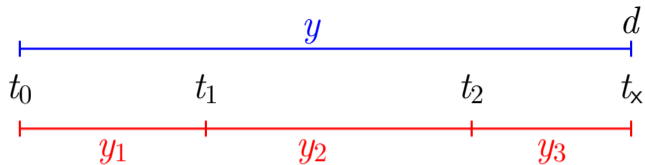
Events and **Risk time**.

The outcome is thus **bivariate**: (d, y)

Follow-up **data** for each individual must therefore have (at least) three variables:

Date of entry	entry	date variable
Date of exit	exit	date variable
Status at exit	fail	indicator (0/1)

Specific for each **type** of outcome.



Probability

$$P(d \text{ at } t_x | \text{entry } t_0)$$

$$= P(\text{surv } t_0 \rightarrow t_1 | \text{entry } t_0)$$

$$\times P(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1)$$

$$\times P(d \text{ at } t_x | \text{entry } t_2)$$

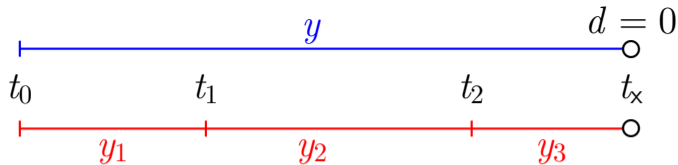
log-Likelihood

$$d \log(\lambda) - \lambda y$$

$$= 0 \log(\lambda) - \lambda y_1$$

$$+ 0 \log(\lambda) - \lambda y_2$$

$$+ d \log(\lambda) - \lambda y_3$$



Probability

$$P(\text{surv } t_0 \rightarrow t_x | \text{entry } t_0)$$

$$= P(\text{surv } t_0 \rightarrow t_1 | \text{entry } t_0)$$

$$\times P(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1)$$

$$\times P(\text{surv } t_2 \rightarrow t_x | \text{entry } t_2)$$

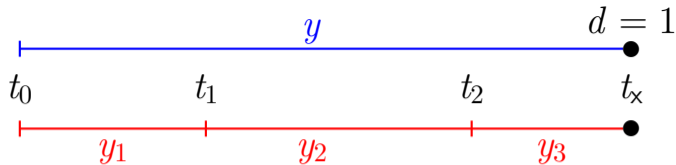
log-Likelihood

$$0 \log(\lambda) - \lambda y$$

$$= 0 \log(\lambda) - \lambda y_1$$

$$+ 0 \log(\lambda) - \lambda y_2$$

$$+ 0 \log(\lambda) - \lambda y_3$$



Probability

$P(\text{event at } t_x | \text{entry } t_0)$

$= P(\text{surv } t_0 \rightarrow t_1 | \text{entry } t_0)$

$\times P(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1)$

$\times P(\text{event at } t_x | \text{entry } t_2)$

log-Likelihood

$1 \log(\lambda) - \lambda y$

$= 0 \log(\lambda) - \lambda y_1$

$+ 0 \log(\lambda) - \lambda y_2$

$+ 1 \log(\lambda) - \lambda y_3$

Dividing time into bands:

If we want to put D and Y into intervals on the timescale we must know:

Origin: The date where the time scale is 0:

- ▶ Age — 0 at date of birth
- ▶ Disease duration — 0 at date of diagnosis
- ▶ Occupation exposure — 0 at date of hire

Intervals: How should it be subdivided:

- ▶ 1-year classes? 5-year classes?
- ▶ Equal length?

Aim: Separate rate in each interval

Example: cohort with 3 persons:

Id	Bdate	Entry	Exit	St
1	14/07/1952	04/08/1965	27/06/1997	1
2	01/04/1954	08/09/1972	23/05/1995	0
3	10/06/1987	23/12/1991	24/07/1998	1

- ▶ Age bands: 10-years intervals of current age.
- ▶ Split Y for every subject accordingly
- ▶ Treat each segment as a separate unit of observation.
- ▶ Keep track of exit status in each interval.

Splitting the follow up

	subj. 1	subj. 2	subj. 3
Age at E ntry:	13.06	18.44	4.54
Age at e X it:	44.95	41.14	11.12
S tatus at exit:	Dead	Alive	Dead
<hr/>			
<i>Y</i>	31.89	22.70	6.58
<i>D</i>	1	0	1

Age	subj. 1		subj. 2		subj. 3		Σ	
	Y	D	Y	D	Y	D	Y	D
0-	0.00	0	0.00	0	5.46	0	5.46	0
10-	6.94	0	1.56	0	1.12	1	8.62	1
20-	10.00	0	10.00	0	0.00	0	20.00	0
30-	10.00	0	10.00	0	0.00	0	20.00	0
40-	4.95	1	1.14	0	0.00	0	6.09	1
Σ	31.89	1	22.70	0	6.58	1	60.17	2

Splitting the follow-up

id	Bdate	Entry	Exit	St	risk	int
1	14/07/1952	03/08/1965	14/07/1972	0	6.9432	10
1	14/07/1952	14/07/1972	14/07/1982	0	10.0000	20
1	14/07/1952	14/07/1982	14/07/1992	0	10.0000	30
1	14/07/1952	14/07/1992	27/06/1997	1	4.9528	40
2	01/04/1954	08/09/1972	01/04/1974	0	1.5606	10
2	01/04/1954	01/04/1974	31/03/1984	0	10.0000	20
2	01/04/1954	31/03/1984	01/04/1994	0	10.0000	30
2	01/04/1954	01/04/1994	23/05/1995	0	1.1417	40
3	10/06/1987	23/12/1991	09/06/1997	0	5.4634	0
3	10/06/1987	09/06/1997	24/07/1998	1	1.1211	10

Keeping track of calendar time too?

Timescales

- ▶ A timescale is a variable that varies **deterministically** *within* each person during follow-up:
 - ▶ Age
 - ▶ Calendar time
 - ▶ Time since treatment
 - ▶ Time since relapse
- ▶ All timescales advance at the same pace (1 year per year . . .)
- ▶ Note: Cumulative exposure is **not** a timescale.

Follow-up on several timescales

- ▶ The risk-time is the same on all timescales
- ▶ Only need the entry point on each time scale:
 - ▶ Age at entry.
 - ▶ Date of entry.
 - ▶ Time since treatment at entry.
 - if time of treatment is the entry, this is 0 for all.
- ▶ Response variable in analysis of rates:

(d, y) (event, duration)

- ▶ Covariates in analysis of rates:
 - ▶ timescales
 - ▶ other (fixed) measurements

Follow-up data in Epi — Lexis objects

A follow-up study:

```
> round( th, 2 )
      id sex birthdat contrast injecdat volume exitdat exitstat
1     1  2  1916.61         1  1938.79      22  1976.79         1
2    640  2  1896.23         1  1945.77      20  1964.37         1
3   3425  1  1886.97         2  1955.18       0  1956.59         1
4   4017  2  1936.81         2  1957.61       0  1992.14         2
...
```

Timescales of interest:

- ▶ Age
- ▶ Calendar time
- ▶ Time since injection

Definition of Lexis object

```
> thL <- Lexis( entry = list( age = injecdat-birthdat,  
+                          per = injecdat,  
+                          tfi = 0 ),  
+              exit = list( per = exitdat ),  
+              exit.status = as.numeric(exitstat==1),  
+              data = th )
```

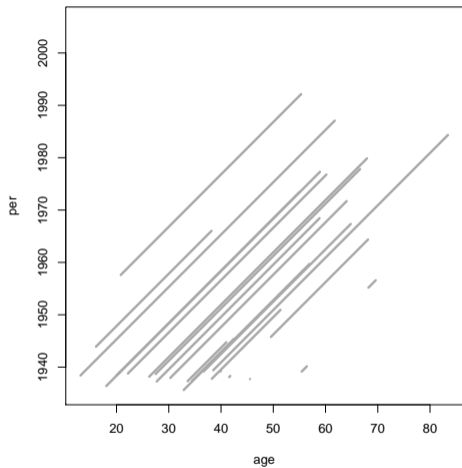
`entry` is defined on **three** timescales,
but `exit` is only defined on **one** timescale:
`Follow-up time` is the same on all timescales:

`exitdat - injecdat`

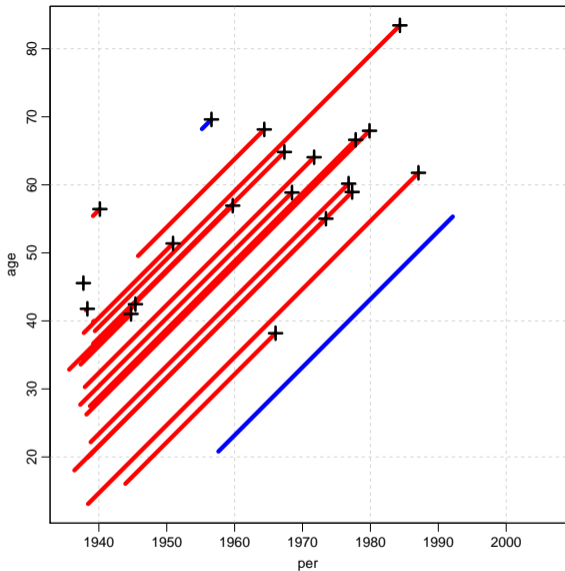
The looks of a Lexis object

```
> thL[,1:9]
  age      per tfi lex.dur lex.Cst lex.Xst lex.id
1 22.18 1938.79  0  37.99      0      1      1
2 49.54 1945.77  0  18.59      0      1      2
3 68.20 1955.18  0   1.40      0      1      3
4 20.80 1957.61  0  34.52      0      0      4
...
```

```
> summary( thL )
Transitions:
      To
From 0  1 Records:  Events:  Risk time:  Persons:
    0 3 20         23         20         512.59         23
```

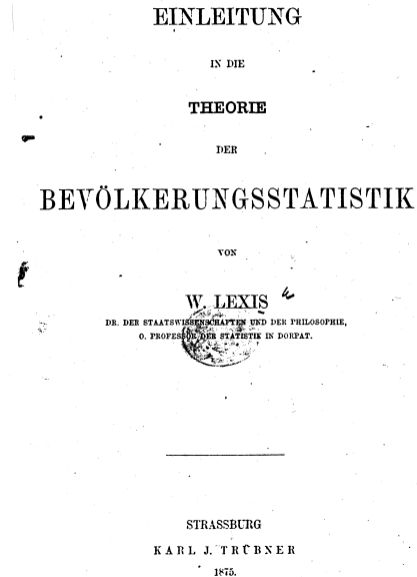


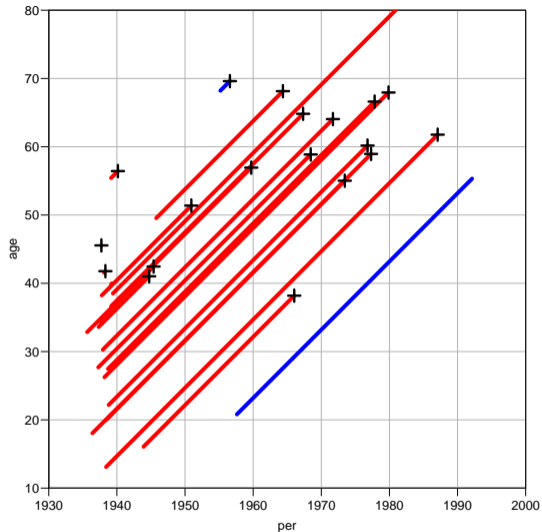
```
> plot( thL, lwd=3 )
```



Representation of follow-up data (time-split)

Lexis diagram





```

> plot( thL, 2:1, lwd=5, col=c("red","blue")[thL$contrast],
+       grid=TRUE, lty.grid=1, col.grid=gray(0.7),
+       xlim=1930+c(0,70), xaxs="i", ylim= 10+c(0,70), yaxs="i", las=1 )
> points( thL, 2:1, pch=c(NA,3)[thL$lex.Xst+1], lwd=3, cex=1.5 )

```

Splitting follow-up time

```
> spl1 <- splitLexis( thL, breaks=seq(0,100,20),  
>                               time.scale="age" )  
> round(spl1,1)
```

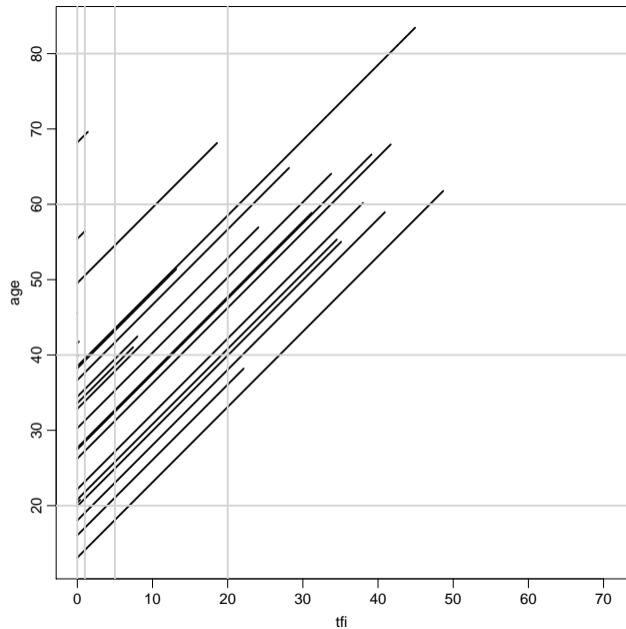
	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	birthdat	contrast	injecdat	vol
1	22.2	1938.8	0.0	17.8	0	0	1	2	1916.6	1	1938.8	
2	40.0	1956.6	17.8	20.0	0	0	1	2	1916.6	1	1938.8	
3	60.0	1976.6	37.8	0.2	0	1	1	2	1916.6	1	1938.8	
4	49.5	1945.8	0.0	10.5	0	0	640	2	1896.2	1	1945.8	
5	60.0	1956.2	10.5	8.1	0	1	640	2	1896.2	1	1945.8	
6	68.2	1955.2	0.0	1.4	0	1	3425	1	1887.0	2	1955.2	
7	20.8	1957.6	0.0	19.2	0	0	4017	2	1936.8	2	1957.6	
8	40.0	1976.8	19.2	15.3	0	0	4017	2	1936.8	2	1957.6	
...												

Split on another timescale

```
> spl2 <- splitLexis( spl1, time.scale="tfi",  
                      breaks=c(0,1,5,20,100) )
```

```
> round( spl2, 1 )
```

	lex.id	age	per	tfi	lex.dur	lex.Cst	lex.Xst	id	sex	birthdat	contrast	inje
1	1	22.2	1938.8	0.0	1.0	0	0	1	2	1916.6	1	19
2	1	23.2	1939.8	1.0	4.0	0	0	1	2	1916.6	1	19
3	1	27.2	1943.8	5.0	12.8	0	0	1	2	1916.6	1	19
4	1	40.0	1956.6	17.8	2.2	0	0	1	2	1916.6	1	19
5	1	42.2	1958.8	20.0	17.8	0	0	1	2	1916.6	1	19
6	1	60.0	1976.6	37.8	0.2	0	1	1	2	1916.6	1	19
7	2	49.5	1945.8	0.0	1.0	0	0	640	2	1896.2	1	19
8	2	50.5	1946.8	1.0	4.0	0	0	640	2	1896.2	1	19
9	2	54.5	1950.8	5.0	5.5	0	0	640	2	1896.2	1	19
10	2	60.0	1956.2	10.5	8.1	0	1	640	2	1896.2	1	19
11	3	68.2	1955.2	0.0	1.0	0	0	3425	1	1887.0	2	19
12	3	69.2	1956.2	1.0	0.4	0	1	3425	1	1887.0	2	19
13	4	20.8	1957.6	0.0	1.0	0	0	4017	2	1936.8	2	19
14	4	21.8	1958.6	1.0	4.0	0	0	4017	2	1936.8	2	19
15	4	25.8	1962.6	5.0	14.2	0	0	4017	2	1936.8	2	19
16	4	40.0	1976.8	19.2	0.8	0	0	4017	2	1936.8	2	19
17	4	40.8	1977.6	20.0	14.5	0	0	4017	2	1936.8	2	19



age	tfi	lex.dur	lex.Cst	lex.Xst
22.2	0.0	1.0	0	0
23.2	1.0	4.0	0	0
27.2	5.0	12.8	0	0
40.0	17.8	2.2	0	0
42.2	20.0	17.8	0	0
60.0	37.8	0.2	0	1

Likelihood for a piecewise constant rate

- ▶ This setup is for a situation where it is assumed that rates are constant in each of the intervals.
- ▶ Each observation in the dataset contributes a term to a “Poisson” likelihood.
- ▶ Models can include fixed covariates, as well as the timescales (the left end-points of the intervals) as continuous variables.
- ▶ Rates are assumed to vary by timescales:
 - ▶ continuously
 - ▶ non-linearly
- ▶ Rates can vary along several timescales simultaneously.

Where is (d_{pi}, y_{pi}) in the split data?

Likelihood is $d_{pi} \log(\lambda_{pi}) - \lambda_{pi} y_{pi}$

```
> round( spl2, 1 )
  lex.id  age    per   tfi lex.dur lex.Cst lex.Xst   id sex birthdat contrast
1      1  22.2 1938.8  0.0    1.0     0     0    1  2   1916.6         1
2      1  23.2 1939.8  1.0    4.0     0     0    1  2   1916.6         1
3      1  27.2 1943.8  5.0   12.8     0     0    1  2   1916.6         1
4      1  40.0 1956.6 17.8    2.2     0     0    1  2   1916.6         1
5      1  42.2 1958.8 20.0   17.8     0     0    1  2   1916.6         1
6      1  60.0 1976.6 37.8    0.2     0     1    1  2   1916.6         1
7      2  49.5 1945.8  0.0    1.0     0     0   640  2   1896.2         1
8      2  50.5 1946.8  1.0    4.0     0     0   640  2   1896.2         1
9      2  54.5 1950.8  5.0    5.5     0     0   640  2   1896.2         1
10     2  60.0 1956.2 10.5    8.1     0     1   640  2   1896.2         1
...
```

— and what are **covariates** for the rates?

Analysis of results

- ▶ d_{pi} — events in the variable: `lex.Xst`:
In the model as response: `lex.Xst==1`
- ▶ y_{pi} — risk time: `lex.dur` (duration):
In the model as offset `log(y)`, `log(lex.dur)`.
- ▶ Covariates are:
 - ▶ timescales (age, period, time in study)
 - ▶ other variables for this person (constant or *assumed* constant in each interval).
- ▶ Model rates using the covariates in `glm`:
— no difference between time-scales and other covariates.

Classical estimators: Kaplan-Meier

Bendix Carstensen

Senior Statistician, Steno Diabetes Center

Practice in analysis of multistate models using Epi::Lexis

21 September 2016

FRIAS, Freiburg

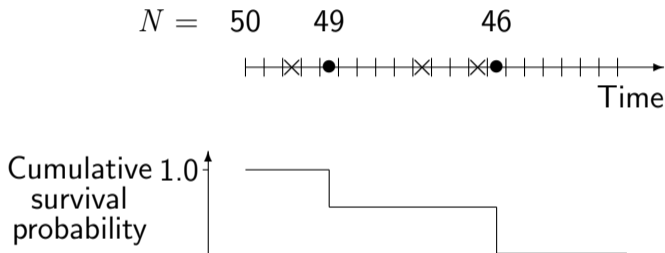
<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

The Kaplan-Meier Method

- ▶ The most common method of estimating the survival function.
- ▶ A non-parametric method.
- ▶ Divides time into small intervals where the intervals are defined by the unique times of failure (death).
- ▶ Based on conditional probabilities as we are interested in the probability a subject surviving the next time interval given that they have survived so far.

Kaplan–Meier method illustrated

(● = failure and × = censored):



- ▶ Steps caused by multiplying by $(1 - 1/49)$ and $(1 - 1/46)$ respectively
- ▶ Late entry can also be dealt with

Using R: Surv()

```
library( survival )  
data( lung )  
head( lung, 3 )
```

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	NA
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	3	90	90	NA	15

```
with( lung, Surv( time, status==2 ) )[1:10]
```

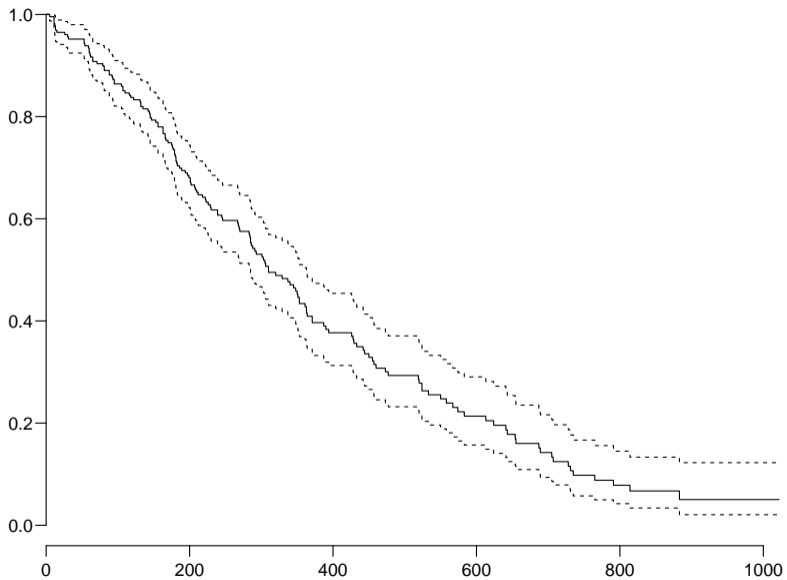
```
[1] 306 455 1010+ 210 883 1022+ 310 361 218 166
```

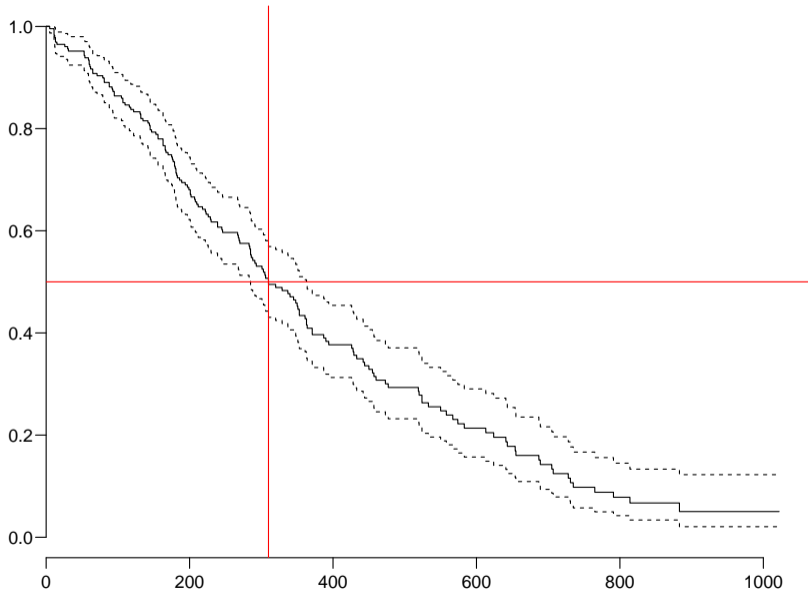
```
( s.km <- survfit( Surv( time, status==2 ) ~ 1 , data=lung ) )
```

```
Call: survfit(formula = Surv(time, status == 2) ~ 1, data = lung)
```

n	events	median	0.95LCL	0.95UCL
228	165	310	285	363

```
plot( s.km )  
abline( v=310, h=0.5, col="red" )
```





Who needs the Cox-model anyway?

Bendix Carstensen

Senior Statistician, Steno Diabetes Center

Practice in analysis of multistate models using Epi::Lexis

21 September 2016

FRIAS, Freiburg

<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

A look at the Cox model

$$\lambda(t, x) = \lambda_0(t) \times \exp(x'\beta)$$

A model for the rate as a function of t and x .

The covariate t has a special status:

- ▶ Computationally, because all individuals contribute to (some of) the range of t .
- ▶ ... the scale along which time is split (the risk sets)
- ▶ Conceptually t is just a covariate that varies within individual.
- ▶ Cox's approach profiles $\lambda_0(t)$ out from the model

The Cox-likelihood as profile likelihood

- ▶ One parameter per death time to describe the effect of time (i.e. the chosen timescale).

$$\log(\lambda(t, x_i)) = \log(\lambda_0(t)) + \beta_1 x_{1i} + \dots + \beta_p x_{pi} = \alpha_t + \eta_i$$

- ▶ Profile likelihood:
 - ▶ Derive estimates of α_t as function of data and β s
— assuming constant rate between death times
 - ▶ Insert in likelihood, now only a function of data and β s
 - ▶ Turns out to be Cox's partial likelihood

The Cox-likelihood: mechanics of computing

- ▶ The likelihood is computed by summing over risk-sets:

$$\ell(\eta) = \sum_t \log \left(\frac{e^{\eta_{\text{death}}}}{\sum_{i \in \mathcal{R}_t} e^{\eta_i}} \right)$$

- ▶ this is essentially splitting follow-up time at event- (and censoring) times
- ▶ ... repeatedly in every cycle of the iteration
- ▶ ... simplified by not keeping track of risk time
- ▶ ... but only works along **one** time scale

$$\log(\lambda(t, x_i)) = \log(\lambda_0(t)) + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} = \alpha_t + \eta_i$$

- ▶ Suppose the time scale has been divided into small intervals with at most one death in each:
- ▶ Empirical rates: (d_{it}, y_{it}) — each t has at most one $d_{it} = 0$.
- ▶ Assume w.l.o.g. the y s in the empirical rates all are 1.
- ▶ Log-likelihood contributions that contain information on a specific time-scale parameter α_t will be from:
 - ▶ the (only) empirical rate $(1, 1)$ with the death at time t .
 - ▶ all other empirical rates $(0, 1)$ from those who were at risk at time t .

Note: There is one contribution from each person at risk to this part of the log-likelihood:

$$\begin{aligned} \ell_t(\alpha_t, \beta) &= \sum_{i \in \mathcal{R}_t} d_i \log(\lambda_i(t)) - \lambda_i(t) y_i \\ &= \sum_{i \in \mathcal{R}_t} \{ d_i(\alpha_t + \eta_i) - e^{\alpha_t + \eta_i} \} \\ &= \alpha_t + \eta_{\text{death}} - e^{\alpha_t} \sum_{i \in \mathcal{R}_t} e^{\eta_i} \end{aligned}$$

where η_{death} is the linear predictor for the person that died.

The derivative w.r.t. α_t is:

$$D_{\alpha_t} \ell_t(\alpha_t, \beta) = 1 - e^{\alpha_t} \sum_{i \in \mathcal{R}_t} e^{\eta_i} = 0 \quad \Leftrightarrow \quad e^{\alpha_t} = \frac{1}{\sum_{i \in \mathcal{R}_t} e^{\eta_i}}$$

If this estimate is fed back into the log-likelihood for α_t , we get the **profile likelihood** (with α_t “profiled out”):

$$\log \left(\frac{1}{\sum_{i \in \mathcal{R}_t} e^{\eta_i}} \right) + \eta_{\text{death}} - 1 = \log \left(\frac{e^{\eta_{\text{death}}}}{\sum_{i \in \mathcal{R}_t} e^{\eta_i}} \right) - 1$$

which is the same as the contribution from time t to Cox’s partial likelihood.

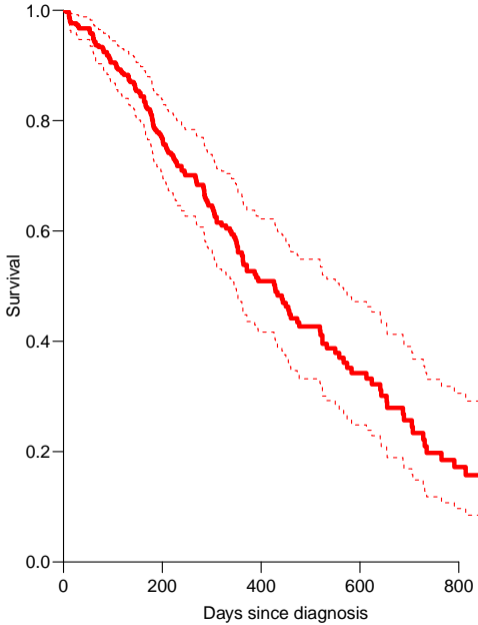
Splitting the dataset a priori

- ▶ The Poisson approach needs a dataset of empirical rates (d, y) with suitably small values of y .
- ▶ — each individual contributes many empirical rates
- ▶ (one per risk-set contribution in Cox-modelling)
- ▶ From each empirical rate we get:
 - ▶ Poisson-response d
 - ▶ Risk time $y \rightarrow \log(y)$ as offset
 - ▶ Covariate value for the timescale (time since entry, current age, current date, ...)
 - ▶ other covariates
- ▶ Contributions not independent, but likelihood is a product
- ▶ Same likelihood as for independent Poisson variates
- ▶ Modelling is by standard `glm` Poisson

Example: Mayo Clinic lung cancer

- ▶ Survival after lung cancer
- ▶ Covariates:
 - ▶ Age at diagnosis
 - ▶ Sex
 - ▶ Time since diagnosis
- ▶ Cox model
- ▶ Split data:
 - ▶ Poisson model, time as factor
 - ▶ Poisson model, time as spline

Mayo Clinic lung cancer 60 year old woman



Example: Mayo Clinic lung cancer I

```
> library( survival )
> library( Epi )
> Lung <- Lexis( exit = list( tfe=time ),
+               exit.status = factor(status,labels=c("Alive","Dead")),
+               data = lung )
```

NOTE: entry.status has been set to "Alive" for all.

NOTE: entry is assumed to be 0 on the tfe timescale.

Example: Mayo Clinic lung cancer II

```
> mL.cox <- coxph( Surv( tfe, tfe+lex.dur, lex.Xst=="Dead" ) ~
+                   age + factor( sex ),
+                   method="breslow", eps=10^-8, iter.max=25, data=Lung )
> Lung.s <- splitLexis( Lung,
+                      breaks=c(0,sort(unique(Lung$time))),
+                      time.scale="tfe" )
> Lung.S <- splitLexis( Lung,
+                      breaks=c(0,sort(unique(Lung$time[Lung$lex.Xst=="Dead"]))),
+                      time.scale="tfe" )
> summary( Lung.s )
```

Transitions:

To

From	Alive	Dead	Records:	Events:	Risk time:	Persons:
Alive	19857	165	20022	165	69593	228

```
> summary( Lung.S )
```

Example: Mayo Clinic lung cancer III

Transitions:

To

From	Alive	Dead	Records:	Events:	Risk time:	Persons:
Alive	15916	165	16081	165	69593	228

```
> subset( Lung.s, lex.id==96 )[,1:11]
```

	lex.id	tfe	lex.dur	lex.Cst	lex.Xst	inst	time	status	age	sex	ph.ecog
9235	96	0	5	Alive	Alive	12	30	2	72	1	2
9236	96	5	6	Alive	Alive	12	30	2	72	1	2
9237	96	11	1	Alive	Alive	12	30	2	72	1	2
9238	96	12	1	Alive	Alive	12	30	2	72	1	2
9239	96	13	2	Alive	Alive	12	30	2	72	1	2
9240	96	15	11	Alive	Alive	12	30	2	72	1	2
9241	96	26	4	Alive	Dead	12	30	2	72	1	2

```
> nlevels( factor( Lung.s$tfe ) )
```

```
[1] 186
```

Example: Mayo Clinic lung cancer IV

```
> system.time(  
+ mLs.pois.fc <- glm( lex.Xst=="Dead" ~ - 1 + factor( tfe ) +  
+                   age + factor( sex ),  
+                   offset = log(lex.dur),  
+                   family=poisson, data=Lung.s, eps=10^-8, maxit=25 )  
+ )
```

```
user  system elapsed  
10.905  0.016  10.919
```

```
> length( coef(mLs.pois.fc) )
```

```
[1] 188
```

```
> system.time(  
+ mLs.pois.fc <- glm( lex.Xst=="Dead" ~ - 1 + factor( tfe ) +  
+                   age + factor( sex ),  
+                   offset = log(lex.dur),  
+                   family=poisson, data=Lung.S, eps=10^-8, maxit=25 )  
+ )
```

Example: Mayo Clinic lung cancer V

```
user system elapsed
3.286  0.012  3.297
```

```
> length( coef(mLS.pois.fc) )
```

```
[1] 142
```

```
> t.kn <- c(0,25,100,500,1000)
> dim( Ns(Lung.s$tfe,knots=t.kn) )
```

```
[1] 20022      4
```

```
> system.time(
+ mLS.pois.sp <- glm( lex.Xst=="Dead" ~ Ns( tfe, knots=t.kn ) +
+                               age + factor( sex ),
+                               offset = log(lex.dur),
+                               family=poisson, data=Lung.s, eps=10^-8, maxit=25 )
+ )
```

Example: Mayo Clinic lung cancer VI

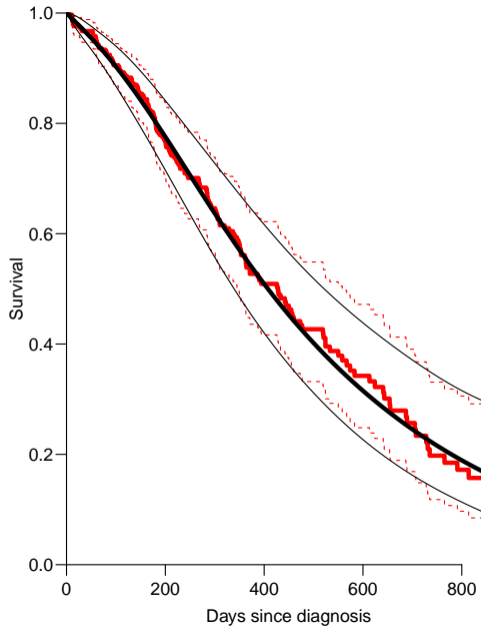
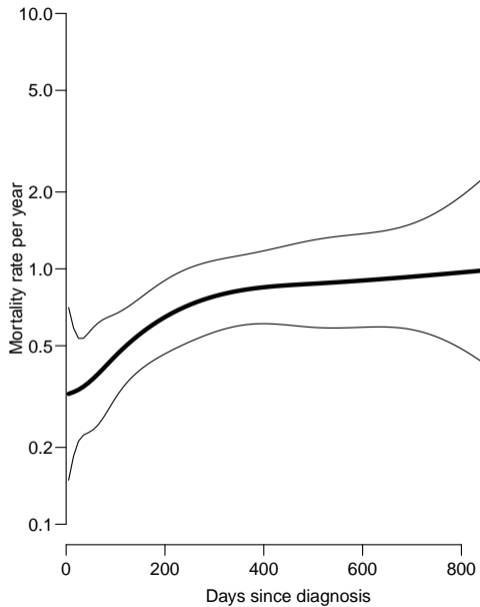
```
user  system elapsed
0.177  0.000  0.176
```

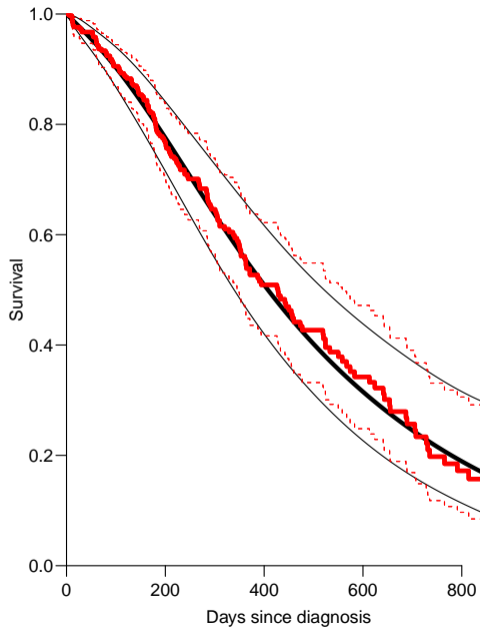
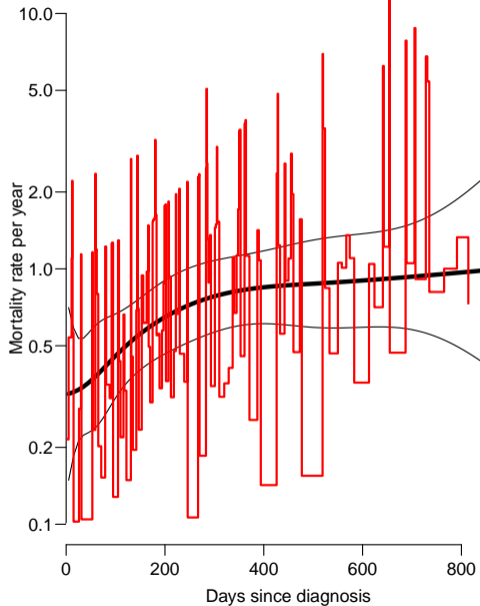
```
> ests <-
+ rbind( ci.exp(mL.cox),
+        ci.exp(mLs.pois.fc,subset=c("age","sex")),
+        ci.exp(mLS.pois.fc,subset=c("age","sex")),
+        ci.exp(mLs.pois.sp,subset=c("age","sex")) )
> cmp <- cbind( ests[c(1,3,5,7)  ],
+              ests[c(1,3,5,7)+1,] )
> rownames( cmp ) <- c("Cox","Poisson-factor","Poisson-factor (D)","Poisson-spline")
> colnames( cmp )[c(1,4)] <- c("age","sex")

> round( cmp, 7 )
```


Example: Mayo Clinic lung cancer VII

	age			sex		
		2.5%	97.5%		2.5%	97.5%
Cox	1.017158	0.9989388	1.035710	0.5989574	0.4313720	0.8316487
Poisson-factor	1.017158	0.9989388	1.035710	0.5989574	0.4313720	0.8316487
Poisson-factor (D)	1.017332	0.9991211	1.035874	0.5984794	0.4310150	0.8310094
Poisson-spline	1.016189	0.9980329	1.034676	0.5998287	0.4319932	0.8328707





Deriving the survival function

```
> mLs.pois.sp <- glm( lex.Xst=="Dead" ~ Ns( tfe, knots=t.kn ) +
+                   age + factor( sex ),
+                   offset = log(lex.dur),
+                   family=poisson, data=Lung.s, eps=10^-8, maxit=25 )

> CM <- cbind( 1, Ns( seq(10,1000,10)-5, knots=t.kn ), 60, 1 )
> lambda <- ci.exp( mLs.pois.sp, ctr.mat=CM )
> Lambda <- ci.cum( mLs.pois.sp, ctr.mat=CM, intl=10 )[, -4]
> survP <- exp(-rbind(0, Lambda))
```

Code and output for the entire example available in
<http://bendixcarstensen.com/AdvCoh/WNtCMA/>

What the Cox-model really is

Taking the life-table approach *ad absurdum* by:

- ▶ dividing time very finely and
- ▶ modeling one covariate, the time-scale, with one parameter per distinct value.
- ▶ the **model** for the time scale is really with exchangeable time-intervals.
- ▶ \Rightarrow difficult to access the baseline hazard (which looks terrible)
- ▶ \Rightarrow uninitiated tempted to show survival curves where irrelevant

Models of this world

- ▶ Replace the α_t s by a parametric function $f(t)$ with a limited number of parameters, for example:
 - ▶ Piecewise constant
 - ▶ Splines (linear, quadratic or cubic)
 - ▶ Fractional polynomials
- ▶ the two latter brings model into “this world”:
 - ▶ smoothly varying rates
 - ▶ parametric closed form representation of baseline hazard
 - ▶ finite no. of parameters
- ▶ Makes it really easy to use rates directly in calculations of
 - ▶ expected residual life time
 - ▶ state occupancy probabilities in multistate models
 - ▶ ...

Likelihood for multistate follow-up

Bendix Carstensen

Senior Statistician, Steno Diabetes Center

Practice in analysis of multistate models using Epi::Lexis

21 September 2016

FRIAS, Freiburg

<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

Likelihood for transition through states

$$\mathbf{A} \longrightarrow \mathbf{B} \longrightarrow \mathbf{C} \longrightarrow$$

- ▶ given start of observation in **A** at time t_0
- ▶ transitions at times t_B and t_C
- ▶ survival in **C** till (at least) time t_x :

$$L = P\{\text{survive } t_0 \rightarrow t_B \text{ in } \mathbf{A}\}$$

$$\times P\{\text{transition } \mathbf{A} \rightarrow \mathbf{B} \text{ at } t_B \mid \text{alive in } \mathbf{A}\}$$

$$\times P\{\text{survive } t_B \rightarrow t_C \text{ in } \mathbf{B} \mid \text{entered } \mathbf{B} \text{ at } t_B\}$$

$$\times P\{\text{transition } \mathbf{B} \rightarrow \mathbf{C} \text{ at } t_C \mid \text{alive in } \mathbf{B}\}$$

$$\times P\{\text{survive } t_C \rightarrow t_x \text{ in } \mathbf{C} \mid \text{entered } \mathbf{C} \text{ at } t_C\}$$

- ▶ Product of likelihood contributions for each transition
— each one as for a survival model

Likelihood contributions reflected in Lexis object

$$\begin{aligned} L = & P\{\text{survive } t_0 \rightarrow t_B \text{ in } \mathbf{A}\} \\ & \times P\{\text{transition } \mathbf{A} \rightarrow \mathbf{B} \text{ at } t_B \mid \text{alive in } \mathbf{A}\} \\ & \times P\{\text{survive } t_B \rightarrow t_C \text{ in } \mathbf{B} \mid \text{entered } \mathbf{B} \text{ at } t_B\} \\ & \times P\{\text{transition } \mathbf{B} \rightarrow \mathbf{C} \text{ at } t_C \mid \text{alive in } \mathbf{B}\} \\ & \times P\{\text{survive } t_C \rightarrow t_x \text{ in } \mathbf{C} \mid \text{entered } \mathbf{C} \text{ at } t_C\} \end{aligned}$$

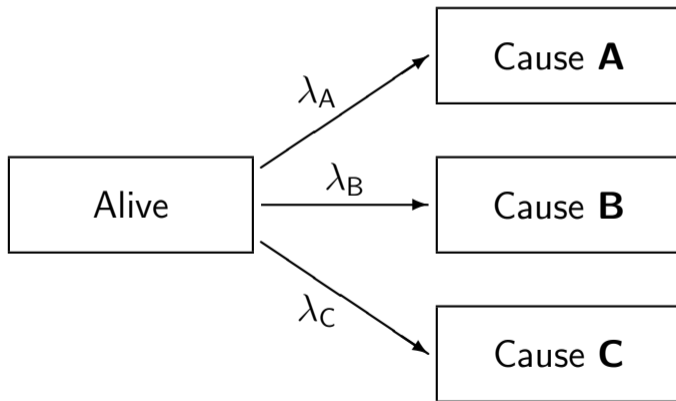
lex.id	time	lex.dur	lex.Cst	lex.Xst
1	t_0	t_B-t_0	A	B
1	t_B	t_C-t_B	B	C
1	t_C	t_x-t_C	C	C

constant rate in interval \Rightarrow log-likelihood term is Poisson:

$$d\log(\lambda) - \lambda y = (\text{lex.Xst!} = \text{lex.Cst}) \times \log(\lambda) - \lambda \times \text{lex.dur}$$

Competing risks

But you may die from more than one cause
(move to one of more possible states):



Cause-specific intensities

$$\lambda_A(t) = \lim_{h \rightarrow 0} \frac{P \{ \text{death from cause A in } (t, t + h] \mid \text{alive at } t \}}{h}$$

$$\lambda_B(t) = \lim_{h \rightarrow 0} \frac{P \{ \text{death from cause B in } (t, t + h] \mid \text{alive at } t \}}{h}$$

$$\lambda_C(t) = \lim_{h \rightarrow 0} \frac{P \{ \text{death from cause C in } (t, t + h] \mid \text{alive at } t \}}{h}$$

Total mortality rate:

$$\lambda_{\text{Total}}(t) = \lim_{h \rightarrow 0} \frac{P \{ \text{death from any cause in } (t, t + h] \mid \text{alive at } t \}}{h}$$

Cause-specific intensities

For small h , $P \{2 \text{ events in } (t, t + h]\} \approx 0$, so:

$$\begin{aligned} & P \{ \text{death from any cause in } (t, t + h] \mid \text{alive at } t \} \\ &= P \{ \text{death from cause A in } (t, t + h] \mid \text{alive at } t \} + \\ & \quad P \{ \text{death from cause B in } (t, t + h] \mid \text{alive at } t \} + \\ & \quad P \{ \text{death from cause C in } (t, t + h] \mid \text{alive at } t \} \\ & \implies \lambda_{\text{Total}}(t) = \lambda_A(t) + \lambda_B(t) + \lambda_C(t) \end{aligned}$$

Intensities are additive,
if they all refer to the
same risk set, in this case “Alive”.

Likelihood for competing risks

Data:

Y - person years in “Alive”

D_A - deaths from cause A

D_B - deaths from cause B

D_C - deaths from cause C

Now, assume for a start that transition rates between states are constant.

Likelihood for competing risks

A survivor contributes to the log-likelihood:

$$\log(P \{\text{Survival for a time of } y\}) = -(\lambda_A + \lambda_B + \lambda_C)y$$

A death from cause **A** contributes an additional $\log(\lambda_A)$, from cause **B** an additional $\log(\lambda_B)$ etc.

The total log-likelihood is then:

$$\begin{aligned}\ell(\lambda_A, \lambda_B, \lambda_C) &= D_A \log(\lambda_A) + D_B \log(\lambda_B) + D_C \log(\lambda_C) \\ &\quad - (\lambda_A + \lambda_B + \lambda_C) Y \\ &= [D_A \log(\lambda_A) - \lambda_A Y] + \\ &\quad [D_B \log(\lambda_B) - \lambda_B Y] + \\ &\quad [D_C \log(\lambda_C) - \lambda_C Y]\end{aligned}$$

Components of the likelihood

The log-likelihood is made up of three contributions:

- ▶ one for cause A,
- ▶ one for cause B and
- ▶ one for cause C

Deaths are the cause-specific deaths,

but the **person-years** are the **same** in all contributions.

The person-years appear once for each transition **out** of a state.

Likelihood for multiple states

- ▶ **Product** of likelihoods for each transition
— each one as for a survival model
- ▶ **conditional** on being alive at (observed) entry to current state
- ▶ **Risk time** is the risk time in the current (“From”, $lex.Cst$) state
- ▶ **Events** are transitions to the “To” state ($lex.Xst$)
- ▶ All other transitions out of “From” are treated as **censorings** (but they are not)
- ▶ Fit models separately for each transition or jointly for all

Time varying rates:

- ▶ The same type of analysis as with a constant rates
- ▶ ... but data must be split in intervals sufficiently small to justify an assumption of constant rate (intensity),
- ▶ the model should allow for a separate rate for each interval,
- ▶ but these can be constrained to follow model with a smooth effect of the time-scale values allocated to each interval.

Practical implications

- ▶ Empirical rates $((d, y)$ from each individual) will be the same for all analyses except for those where deaths occur.
- ▶ Analysis of cause **A**:
 - ▶ Contributions $(1, y)$ only for those intervals where a cause **A** death occurs.
 - ▶ Intervals with cause **B** or **C** deaths (or no deaths) contribute only $(0, y)$ — treated as censorings.

original						
id	time	cause	xx	d.A	d.B	d.C
1	1	B	0.50	0	1	0
2	1	NA	1.00	0	0	0
3	8	B	-1.74	0	1	0
4	3	A	-0.55	1	0	0
5	7	NA	-0.58	0	0	0
6	7	C	-0.04	0	0	1

expanded				
id	time	dd	xx	Tr
1	1	0	0.50	A
2	1	0	1.00	A
3	8	0	-1.74	A
4	3	1	-0.55	A
5	7	0	-0.58	A
6	7	0	-0.04	A
1	1	1	0.50	B
2	1	0	1.00	B
3	8	1	-1.74	B
4	3	0	-0.55	B
5	7	0	-0.58	B
6	7	0	-0.04	B
1	1	0	0.50	C
2	1	0	1.00	C
3	8	0	-1.74	C
4	3	0	-0.55	C
5	7	0	-0.58	C
6	7	1	-0.04	C

... accomplished by `stack.Lexis`

Lexis objects (data frame)

- ▶ Represents the **follow-up**
- ▶ `lex.dur` contains the total time at risk for (any) event
- ▶ `lex.Cst` is the state in which this time is spent
- ▶ `lex.Xst` is the state to which a transition occurs
— if no transition, the same as `lex.Cst`.

This is used for modelling of single transitions between states — and multiple transitions with no two originating in the same state.

stacked.Lexis **objects (data frame)**

- ▶ Represents the **likelihood** contributions
- ▶ `lex.dur` contains the total time at risk for (any) event
- ▶ `lex.Tr` is the transition to which the record contributes
- ▶ `lex.Fail` is the event (failure) indicator for the transition in question.

This is used for joint modelling of **all** transition in a multistate set-up.

Particularly with several rates originating in the **same** state (competing risks).

Implemented in the `stack.Lexis` function:

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

	sex	dobth	dodm	dodth	doad	doins	dox
50185	F	1940.256	1998.917	NA	NA	NA	2009.997
307563	M	1939.218	2003.309	NA	2007.446	NA	2009.997
294104	F	1918.301	2004.552	NA	NA	NA	2009.997
336439	F	1965.225	2009.261	NA	NA	NA	2009.997
245651	M	1932.877	2008.653	NA	NA	NA	2009.997
216824	F	1927.870	2007.886	2009.923	NA	NA	2009.923

```
> dml <- Lexis( entry = list(Per = dodm,
+                             Age = dodm-dobth,
+                             DMdur = 0 ),
+               exit = list(Per = dox ),
+               exit.status = factor(!is.na(dodth),
+                                   labels=c("DM", "Dead")),
+               data = DMlate )
```

NOTE: `entry.status` has been set to "DM" for all.

Implemented in the `stack.Lexis` function:

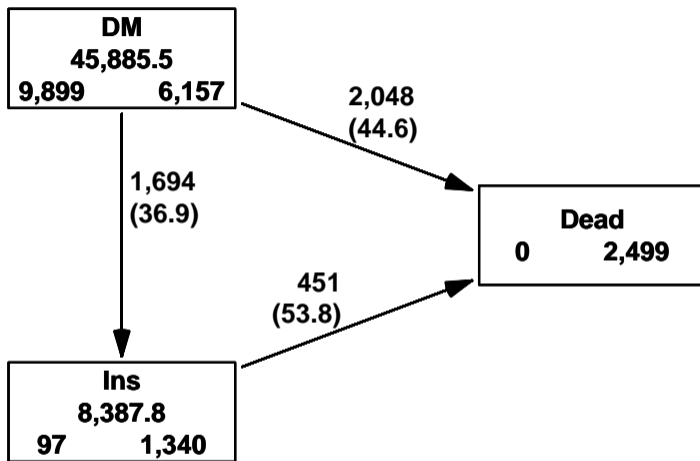
```
> dmi <- cutLexis( dml, cut = dml$doins,  
+                 new.state = "Ins",  
+                 precursor = "DM" )  
> summary( dmi )
```

Transitions:

To

From	DM	Ins	Dead	Records:	Events:	Risk time:	Persons:
DM	6157	1694	2048	9899	3742	45885.49	9899
Ins	0	1340	451	1791	451	8387.77	1791
Sum	6157	3034	2499	11690	4193	54273.27	9996

```
> boxes( dmi, boxpos = list(x=c(20,20,80),  
+                           y=c(80,20,50)),  
+       scale.R=1000, show.BE=TRUE, hmult=1.2, wmult=1.1 )
```



Implemented in the `stack.Lexis` function:

```
> options( digits=3, width=200 )  
> st.dmi <- stack( dmi )  
> print( st.dmi[1:6,], row.names=F )
```

Per	Age	DMdur	lex.dur	lex.Cst	lex.Xst	lex.Tr	lex.Fail	lex.id	sex	dobth	dodm	do
1999	58.7	0	11.080	DM	DM	DM->Ins	FALSE	1	F	1940	1999	
2003	64.1	0	6.689	DM	DM	DM->Ins	FALSE	2	M	1939	2003	
2005	86.3	0	5.446	DM	DM	DM->Ins	FALSE	3	F	1918	2005	
2009	44.0	0	0.736	DM	DM	DM->Ins	FALSE	4	F	1965	2009	
2009	75.8	0	1.344	DM	DM	DM->Ins	FALSE	5	M	1933	2009	
2008	80.0	0	2.037	DM	Dead	DM->Ins	FALSE	6	F	1928	2008	2

```
> str( st.dmi )
```

Classes 'stacked.Lexis' and 'data.frame': 21589 obs. of 16 variables:

```
$ Per      : num  1999 2003 2005 2009 2009 ...  
$ Age      : num  58.7 64.1 86.3 44 75.8 ...  
$ DMdur    : num  0 0 0 0 0 0 0 0 0 0 ...  
$ lex.dur  : num  11.08 6.689 5.446 0.736 1.344 ...  
$ lex.Cst  : Factor w/ 3 levels "DM","Ins","Dead": 1 1 1 1 1 1 1 1 1 1 ...  
$ lex.Xst  : Factor w/ 3 levels "DM","Ins","Dead": 1 1 1 1 1 3 1 1 3 1 ...  
$ lex.Tr   : Factor w/ 3 levels "DM->Ins","DM->Dead",...: 1 1 1 1 1 1 1 1 1 1 ...  
$ lex.Fail : logi  FALSE FALSE FALSE FALSE FALSE FALSE
```

Implemented in the `stack.Lexis` function:

```
> print( subset( dmi, lex.id %in% c(13,15,28) ), row.names=FALSE )
```

Per	Age	DMdur	lex.dur	lex.Cst	lex.Xst	lex.id	sex	dobth	dodm	dodth	doad	doins
1997	59.4	0.0	0.890	DM	Dead	13	M	1938	1997	1998	NA	NA
2003	58.1	0.0	2.804	DM	Ins	15	M	1944	2003	NA	NA	2005
2005	60.9	2.8	4.643	Ins	Ins	15	M	1944	2003	NA	NA	2005
1999	73.7	0.0	8.701	DM	Ins	28	F	1925	1999	2008	2001	2007
2007	82.4	8.7	0.977	Ins	Dead	28	F	1925	1999	2008	2001	2007

```
> print( subset( st.dmi, lex.id %in% c(13,15,28) ), row.names=FALSE )
```

Per	Age	DMdur	lex.dur	lex.Cst	lex.Xst	lex.Tr	lex.Fail	lex.id	sex	dobth	dodm
1997	59.4	0.0	0.890	DM	Dead	DM->Ins	FALSE	13	M	1938	1997
2003	58.1	0.0	2.804	DM	Ins	DM->Ins	TRUE	15	M	1944	2003
1999	73.7	0.0	8.701	DM	Ins	DM->Ins	TRUE	28	F	1925	1999
1997	59.4	0.0	0.890	DM	Dead	DM->Dead	TRUE	13	M	1938	1997
2003	58.1	0.0	2.804	DM	Ins	DM->Dead	FALSE	15	M	1944	2003
1999	73.7	0.0	8.701	DM	Ins	DM->Dead	FALSE	28	F	1925	1999
2005	60.9	2.8	4.643	Ins	Ins	Ins->Dead	FALSE	15	M	1944	2003
2007	82.4	8.7	0.977	Ins	Dead	Ins->Dead	TRUE	28	F	1925	1999

Analysis of rates in multistate models

- ▶ Interactions between all covariates (including time) and state (`lex.Cst`):
 - ⇔ separate analyses of all transition rates.
- ▶ Only interaction between state (`lex.Cst`) and time(scales):
 - ⇔ same covariate effects for all causes transitions, but separate baseline hazards — “stratified model”.
- ▶ Main effect of state only (`lex.Cst`):
 - ⇔ proportional hazards
- ▶ No effect of state:
 - ⇔ identical baseline hazards — hardly ever relevant.

Analysis approaches and data representation

- ▶ Lexis objects represents the precise follow-up in the cohort, in states and along timescales
- ▶ — used for analysis of single transition rates.
- ▶ `stacked.Lexis` objects represents contributions to the total likelihood
- ▶ — used for joint analysis of (all) rates in a multistate setup
- ▶ ... which is the case if you want to specify common effects between different transitions.

Assumptions in competing risks

“Classical” way of looking at survival data:
description of the distribution of time to death.

For competing risks that would require three variables:

T_A , T_B and T_C , representing times to death from each of the three causes.

But at most one of these is observed.

Often it is stated that these must be assumed independent in order to make the likelihood machinery work

1. It is not necessary.
2. Independence can never be assessed from data.

An account of these problems is given in:

PK Andersen, SZ Abildstrøm & S Rosthøj:

Competing risks as a multistate model,

Statistical Methods in Medical Research; **11**, 2002: pp. 203–215

Per Kragh Andersen, Ronald B Geskus, Theo de Witte & Hein Putter:

Competing risks in epidemiology: possibilities and pitfalls,

International Journal of Epidemiology; 2012: pp. 1–10

Contains examples where both dependent and independent “cause specific survival times” gives rise to the same set of cause specific rates.

Reporting a multistate model

Bendix Carstensen

Senior Statistician, Steno Diabetes Center

Practice in analysis of multistate models using Epi::Lexis

21 September 2016

FRIAS, Freiburg

<http://BendixCarstensen/AdvCoh/courses/Frias-2016>

Multistate models

- ▶ Outcomes are transitions between states, with times
- ▶ Covariates are measurements and timescales
- ▶ Models describe the single transition rates
- ▶ Results are:
 - ▶ Description of rates — how do they depend time etc.
 - ▶ Prediction of state occupancy:
What is the probability that a person is in a given state at a given time?
- ▶ This illustrates the latter.

Diabetes patient mortality

```
> library(Epi)
> data(DMlate)
> dml <- Lexis( entry = list(Per=dodm, Age=dodm-dobth, DMdur=0 ),
+               exit = list(Per=dox),
+               exit.status = factor(!is.na(dodth),labels=c("DM","Dead")),
+               data = DMlate )
```

NOTE: entry.status has been set to "DM" for all.

```
> summary(dml)
```

Transitions:

	To					
From	DM	Dead	Records:	Events:	Risk time:	Persons:
DM	7497	2499	9996	2499	54273.27	9996

... subdivided by insulin status

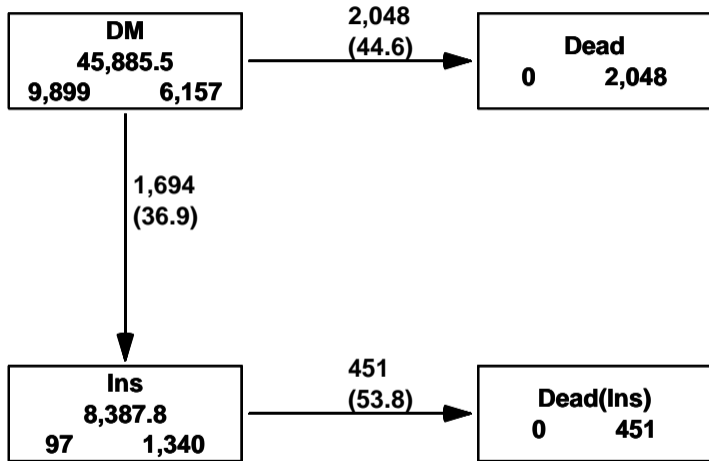
Split follow-up at insulin, introduce a new timescale and split non-precursor states:

```
> dmi <- cutLexis( dml, cut = dml$doins,  
+                 pre = "DM",  
+                 new.state = "Ins",  
+                 new.scale = "t.Ins",  
+                 split.states = TRUE )  
> summary( dmi )
```

Transitions:

	To							
From	DM	Ins	Dead	Dead(Ins)	Records:	Events:	Risk time:	Persons:
DM	6157	1694	2048	0	9899	3742	45885.49	9899
Ins	0	1340	0	451	1791	451	8387.77	1791
Sum	6157	3034	2048	451	11690	4193	54273.27	9996

```
> boxes( dmi, boxpos=list(x=c(20,20,80,80),y=c(80,20,80,20)),  
+       scale.R=1000, show.BE=TRUE, hmult=1.2, wmult=1.2 )
```



Split the follow in 3-month intervals for modelling

```
> Si <- splitLexis( dmi, 0:60/4, "DMdur" )  
> summary( Si )
```

Transitions:

	To							
From	DM	Ins	Dead	Dead(Ins)	Records:	Events:	Risk time:	Persons:
DM	184986	1694	2048	0	188728	3742	45885.49	9899
Ins	0	34707	0	451	35158	451	8387.77	1791
Sum	184986	36401	2048	451	223886	4193	54273.27	9996

```
> summary( dmi )
```

Transitions:

	To							
From	DM	Ins	Dead	Dead(Ins)	Records:	Events:	Risk time:	Persons:
DM	6157	1694	2048	0	9899	3742	45885.49	9899
Ins	0	1340	0	451	1791	451	8387.77	1791
Sum	6157	3034	2048	451	11690	4193	54273.27	9996

Define knots for spline modelling of the rates:

```
> nk <- 4
> ( ai.kn <- with( subset(Si,lex.Xst=="Ins"),
+                 quantile( Age+lex.dur, probs=(1:nk-0.5)/nk ) ) )

    12.5%    37.5%    62.5%    87.5%
27.68241 49.61893 61.88364 75.56211

> ( ad.kn <- with( subset(Si,lex.Xst=="Dead"),
+                 quantile( Age+lex.dur, probs=(1:nk-0.5)/nk ) ) )

    12.5%    37.5%    62.5%    87.5%
63.61875 74.98700 81.38501 89.26831

> ( di.kn <- with( subset(Si,lex.Xst=="Ins"),
+                 quantile( DMdur+lex.dur, probs=(1:nk-0.5)/nk ) ) )

12.5% 37.5% 62.5% 87.5%
 1.50  4.25  7.00 10.50

> ( dd.kn <- with( subset(Si,lex.Xst=="Dead"),
+                 quantile( DMdur+lex.dur, probs=(1:nk-0.5)/nk ) ) )

    12.5%    37.5%    62.5%    87.5%
0.3778234 1.9582478 4.3370979 8.0232717
```

Fit Poisson models to transition rates

```
> DM.Ins <- glm( (lex.Xst=="Ins") ~ Ns( Age , knots=ai.kn ) +
+               Ns( DMdur, knots=di.kn ) +
+               I(Per-2000) + sex,
+               family=poisson, offset=log(lex.dur),
+               data = subset(Si,lex.Cst=="DM") )
> DM.Dead <- glm( (lex.Xst=="Dead") ~ Ns( Age , knots=ad.kn ) +
+               Ns( DMdur, knots=dd.kn ) +
+               I(Per-2000) + sex,
+               family=poisson, offset=log(lex.dur),
+               data = subset(Si,lex.Cst=="DM") )
> Ins.Dead <- glm( (lex.Xst=="Dead(Ins)") ~ Ns( Age , knots=ad.kn ) +
+               Ns( DMdur, knots=dd.kn ) +
+               Ns( t.Ins, knots=td.kn ) +
+               I(Per-2000) + sex,
+               family=poisson, offset=log(lex.dur),
+               data = subset(Si,lex.Cst=="Ins") )
```

Put the fitted models into an object representing the transitions

```
> Tr <- list( "DM" = list( "Ins"      = DM.Ins,  
+                          "Dead"    = DM.Dead ),  
+            "Ins" = list( "Dead(Ins)" = Ins.Dead ) )  
> lapply( Tr, names )
```

```
$DM  
[1] "Ins" "Dead"
```

```
$Ins  
[1] "Dead(Ins)"
```

Define an initial object

— note the combination of `select=` and `NULL` which ensures that the relevant attributes from the Lexis object `Si` are carried over to `ini` (using `Si[NULL,1:9]` will lose essential attributes)

```
> ini <- subset(Si,select=1:9)[NULL,]
> ini[1:2,"lex.Cst"] <- "DM"
> ini[1:2,"Per"] <- 1995
> ini[1:2,"Age"] <- 60
> ini[1:2,"DMdur"] <- 5
> ini[1:2,"sex"] <- c("M","F")
> ini
```

	lex.id	Per	Age	DMdur	t.Ins	lex.dur	lex.Cst	lex.Xst	sex
1	NA	1995	60	5	NA	NA	DM	<NA>	M
2	NA	1995	60	5	NA	NA	DM	<NA>	F

Simulate 10,000 of each sex using the estimated models in Tr:

```
> system.time(  
+ simL <- simLexis( Tr, ini, time.pts=seq(0,11,0.5), N=10000 ) )
```

```
   user  system elapsed  
25.111   0.100  25.208
```

```
> summary( simL )
```

Transitions:

	To								
From	DM	Ins	Dead	Dead(Ins)	Records:	Events:	Risk time:	Persons:	
DM	8817	6167	5016	0	20000	11183	150485.05	20000	
Ins	0	4456	0	1711	6167	1711	33773.71	6167	
Sum	8817	10623	5016	1711	26167	12894	184258.76	20000	

```
> subset( simL, lex.id < 3 )
```

	lex.id	Per	Age	DMdur	t.Ins	lex.dur	lex.Cst	lex.Xst	sex	cens
1	1	1995.000	60.00000	5.00000	NA	1.050103	DM	Dead	M	2006
2	2	1995.000	60.00000	5.00000	NA	6.118532	DM	Ins	M	2006
3	2	2001.119	66.11853	11.11853	0	2.324054	Ins	Dead(Ins)	M	2006

We now have a dataframe (Lexis object) with simulated follow-up of 10,000 men and 10,000 women.

We then find the number of persons in each state at a specified set of times.

```
> nSt <- nState( subset(simL,sex=="M"),  
+               at=seq(0,10,0.1), from=1995, time.scale="Per" )  
> nSt
```

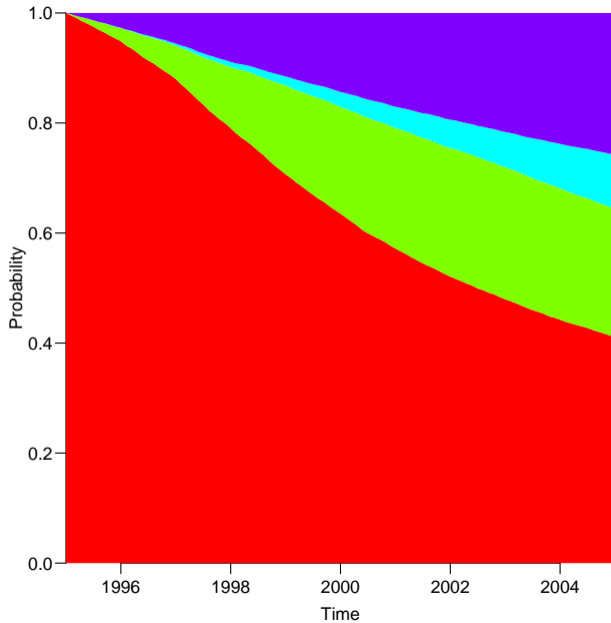
	State			
when	DM	Ins	Dead	Dead(Ins)
1995	10000	0	0	0
1995.1	9950	24	26	0
1995.2	9904	40	56	0
1995.3	9847	72	81	0
1995.4	9801	92	105	2
1995.5	9749	115	134	2
1995.6	9692	140	165	3
1995.7	9645	167	184	4
1995.8	9588	192	214	6
1995.9	9537	211	245	7
1996	9488	225	260	8

Show the cumulative prevalences in a different order than that of the state-level ordering and plot them using all defaults:

```
> pp <- pState( nSt, perm=c(1,2,4,3) )  
> head( pp )
```

	State			
when	DM	Ins	Dead(Ins)	Dead
1995	1.0000	1.0000	1.0000	1
1995.1	0.9950	0.9974	0.9974	1
1995.2	0.9904	0.9944	0.9944	1
1995.3	0.9847	0.9919	0.9919	1
1995.4	0.9801	0.9893	0.9895	1
1995.5	0.9749	0.9864	0.9866	1

```
> plot( pp )
```



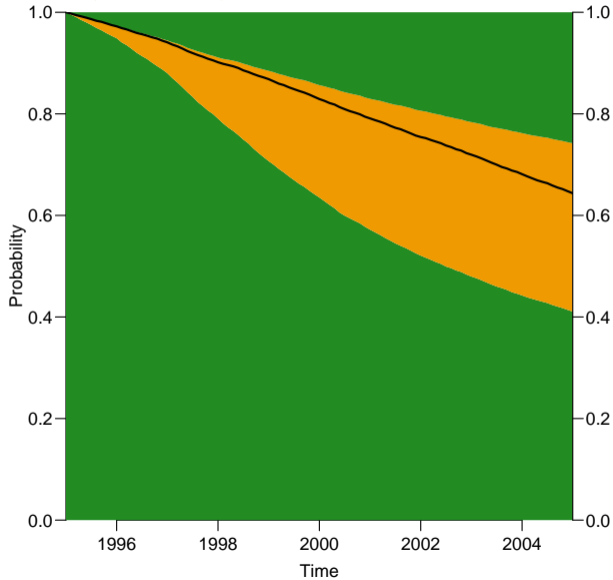
We can show the results in a clearer way, but choosing colors wiser:

```
> clr <- c("orange2","forestgreen")
> par( las=1, mar=c(3,3,3,3) )
> plot( pp, col=clr[c(2,1,1,2)] )
> lines( as.numeric(rownames(pp)), pp[,2], lwd=2 )
> mtext( "60 year old male, diagnosed 1995", side=3, line=2.5, adj=0 )
> mtext( "Survival curve", side=3, line=1.5, adj=0 )
> mtext( "DM, no insulin    DM, Insulin", side=3, line=0.5, adj=0, col=clr[1] )
> mtext( "DM, no insulin", side=3, line=0.5, adj=0, col=clr[2] )
> axis( side=4 )
```

60 year old male, diagnosed 1993

Survival curve

DM, no insulin DM, Insulin

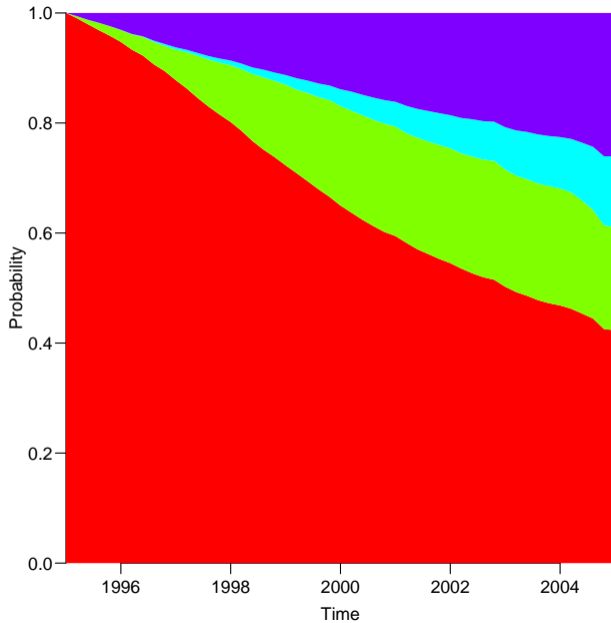


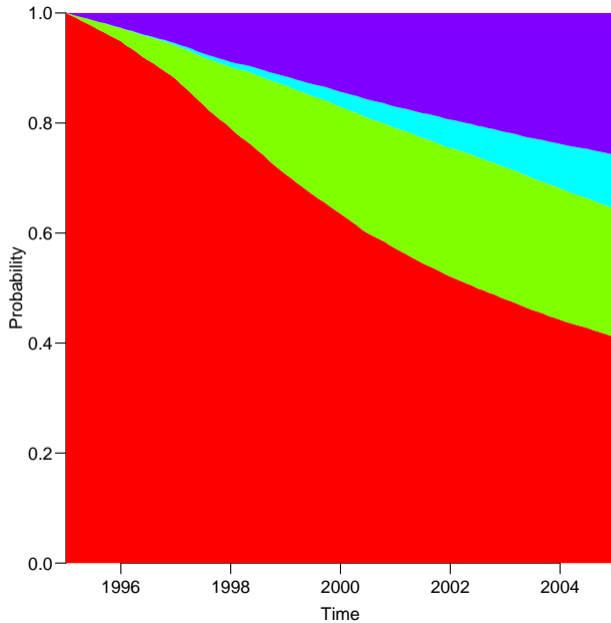
We could also use a Cox-model for the mortality rates assuming the two mortality rates to be proportional:

When we fit a Cox-model, `lex.dur` must be used in the `Surv()` function, and the `I()` construction must be used when specifying intermediate states as covariates, since factors with levels not present in the data will create NAs in the parameter vector returned by `coxph`, which in return will crash the simulation machinery.

```
> library( survival )
> Cox.Dead <- coxph( Surv( DMdur, DMdur+lex.dur,
+                          lex.Xst %in% c("Dead(Ins)", "Dead")) ~
+                          Ns( Age-DMdur, knots=ad.kn ) +
+                          I(lex.Cst=="Ins") +
+                          I(Per-2000) + sex,
+                          data = Si )
```

```
> Cr <- list( "DM" = list( "Ins"      = DM.Ins,
+                          "Dead"    = Cox.Dead ),
+            "Ins" = list( "Dead(Ins)" = Cox.Dead ) )
> simL <- simLexis( Cr, ini, time.pts=seq(0,11,0.2), N=10000 )
> nSt <- nState( subset(simL,sex=="M"),
+               at=seq(0,10,0.2), from=1995, time.scale="Per" )
> pp <- pState( nSt, perm=c(1,2,4,3) )
> plot( pp )
```



Now your turn...

References