Modern Demographic Methods in Epidemiology

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
http://staff.pubhealth.ku.dk/~bxc/

University of St. Andrews, Scotland
Longitudinal Studies Centre
1–3 June 2010

http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010

Rates and Survival
Tuesday 1 June 2010, morning

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre

http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010

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")

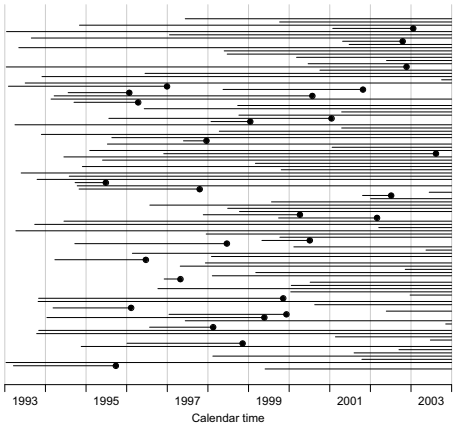
Examples of time-to-event measurements

- 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

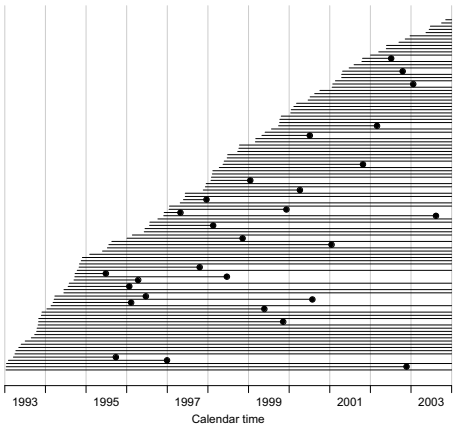
Each blob a death

Study ended at 31 Dec. 2003

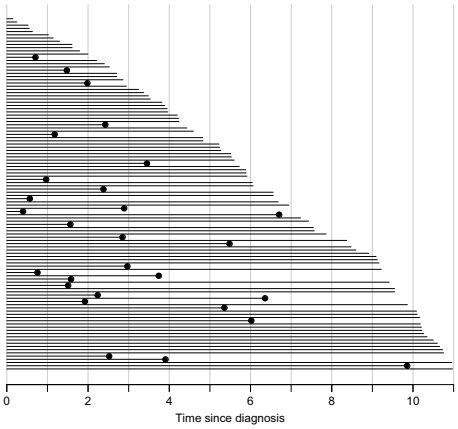


Ordered by date of entry

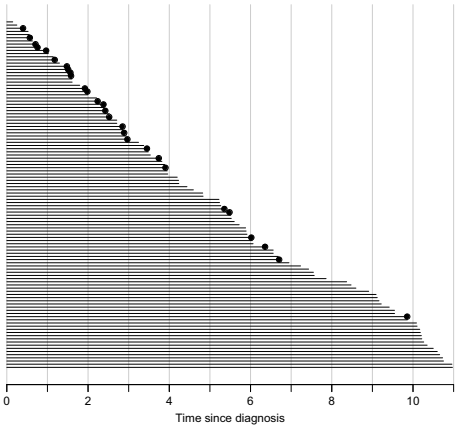
Most likely the order in your database.



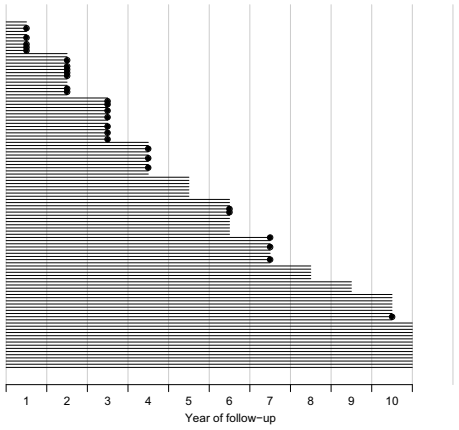
Timescale changed to "Time since diagnosis".



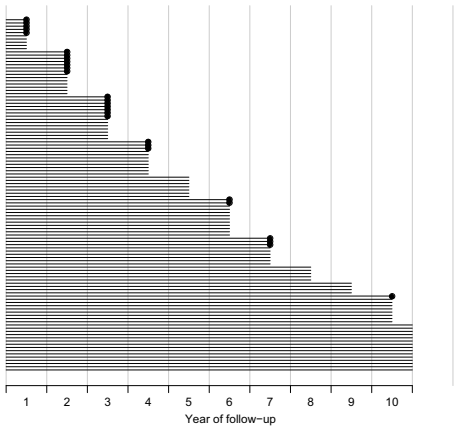
Patients ordered by survival time.



Survival times grouped into bands of survival.



Patients ordered by survival status within each band.



Survival after Cervix cancer

	Stage I			Stage II		
Year	N	D	L	N	D	L
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:

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

Intensity or rate

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

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

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

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

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

Rate and survival

$$S(t) = \exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right) \qquad \lambda(t) = \frac{S'(t)}{S(t)}$$

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

Note: A cumulative measure requires an origin!

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 truncated).
- 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.

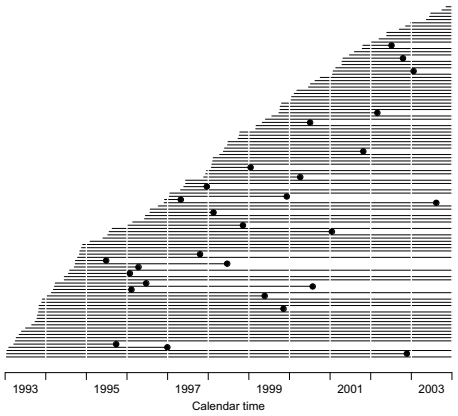
A person contributes several observations of (d, y).

Empirical rates are **responses** in survival analysis.

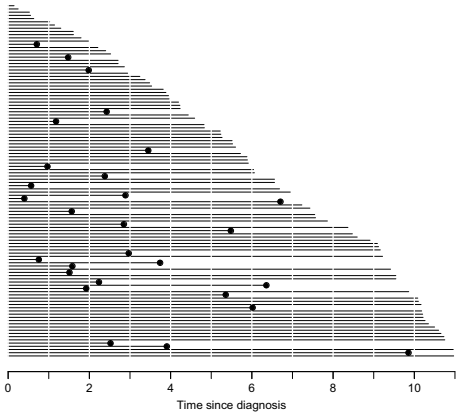
The timescale is a **covariate** — varies across empirical rates from one individual: Age, calendar time, time since diagnosis.

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.



Likelihood from one person

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

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

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).

Likelihood for an empirical rate

Model: the rate is constant in the interval we are looking at. The interval should sufficiently small for this assumption to be reasonable.

If $\pi = 1 - \mathrm{e}^{-\lambda y}$ is the death probability:

$$\begin{aligned} L(\lambda) &= \mathrm{P}\{d \text{ events during } y \text{ time }\} = \pi^d (1-\pi)^{1-d} \\ &= (1 - \mathrm{e}^{-\lambda y})^d (\mathrm{e}^{-\lambda y})^{1-d} \\ &= \left(\frac{1 - \mathrm{e}^{-\lambda y}}{\mathrm{e}^{-\lambda y}}\right)^d (\mathrm{e}^{-\lambda y}) \approx (\lambda y)^d \mathrm{e}^{-\lambda y} \end{aligned}$$

since the first term is equal to $\mathrm{e}^{-\lambda y} - 1 \approx \lambda y$.

Log-likelihood:

$$l(\lambda) = d \log(\lambda y) - \lambda y = d \log(\lambda) + d \log(y) - \lambda y$$

The term $d \log(y)$ does not include λ, so the relevant part of the log-likelihood is:

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

Poisson likelihood

The contributions from **one** individual $d_t \log(\lambda(t)) - \lambda(t) y_t$, is like 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.
- $\log(y)$ is the offset variable.

Likelihood for follow-up of many subjects

Adding empirical rates over the follow-up of persons:

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

- Persons are assumed independent
- Contribution from the same person are *conditionally* independent, hence give separate contributions to the log-likelihood.

The log-likelihood is maximal for:

$$\frac{\mathrm{d}l(\lambda)}{\mathrm{d}\lambda} = \frac{D}{\lambda} - Y = 0 \quad \Leftrightarrow \quad \hat{\lambda} = \frac{D}{Y}$$

Information about $\theta = \log(\lambda)$:

$$l(\theta|D,Y) = D\theta - \mathrm{e}^{\theta}Y, \quad l'_\theta = D - \mathrm{e}^{\theta}Y, \quad l''_\theta = -\mathrm{e}^{\theta}Y$$

so $I(\hat{\theta}) = \mathrm{e}^{\hat{\theta}}Y = \hat{\lambda}Y = D$, hence $\mathrm{var}(\hat{\theta}) = 1/D$

Standard error of log-rate: $1/\sqrt{D}$.

Note that this only depends on the no. events, **not** on the follow-up time.

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:

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

Exercise

Suppose we have 17 deaths during 843.6 years of follow-up.

Calculate the mortality rate with a 95% c.i.

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 ratio of the rates, RR, is:

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

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

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

Exercise

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.

Calculate the rate-ratio between group 1 and 0 with a 95% c.i.

Survival analysis

Response variable: Time to event, T

Censoring: We observe $(\min(T, Z), \delta = 1\{T < Z\})$.

This gives time a special status, and mixes the response variable (risk)time with the covariate time(scale).

Originates from clinical trials where everyone enters at time 0.

The life table method

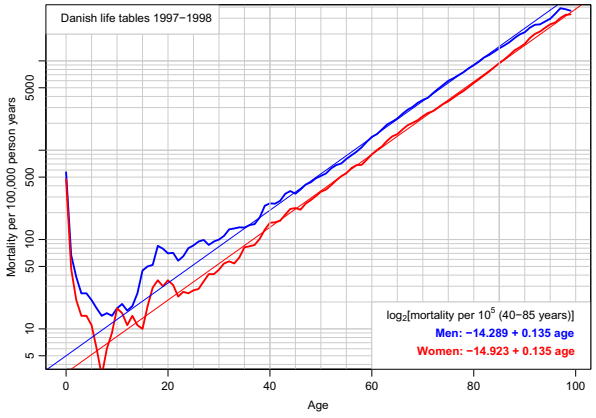
The simplest analysis is by the "life-table method":

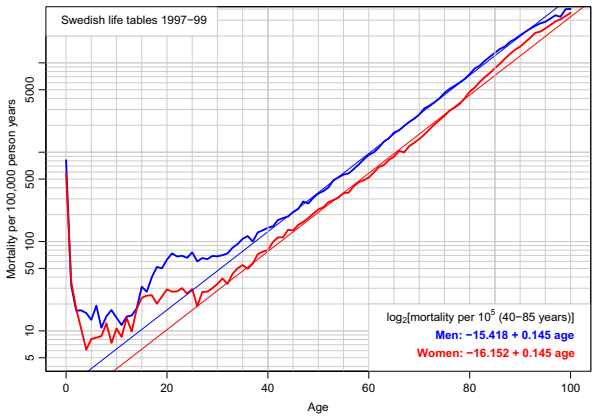
interval i	alive n_i	dead d_i	cens. l_i	p_i
1	77	5	2	$5/(77-2/2) = 0.066$
2	70	7	4	$7/(70-4/2) = 0.103$
3	59	8	1	$8/(59-1/2) = 0.137$

$$\begin{aligned} p_i &= \mathrm{P}\{\text{death in interval } i\} = 1 - d_i/(n_i - l_i/2) \\ S(t) &= (1-p_1) \times \cdots \times (1-p_t) \end{aligned}$$

Population life table, DK 1997–98

	Men			Women		
a	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathrm{res}}(a)]$	$S(a)$	$\lambda(a)$	$\mathrm{E}[\ell_{\mathrm{res}}(a)]$
0	1.00000	567	73.68	1.00000	474	78.65
1	0.99433	67	73.10	0.99526	47	78.02
2	0.99366	38	72.15	0.99479	21	77.06
3	0.99329	25	71.18	0.99458	14	76.08
4	0.99304	25	70.19	0.99444	14	75.09
5	0.99279	21	69.21	0.99430	11	74.10
6	0.99258	17	68.23	0.99419	6	73.11
7	0.99242	14	67.24	0.99413	3	72.11
8	0.99227	15	66.25	0.99410	6	71.11
9	0.99213	14	65.26	0.99404	9	70.12
10	0.99199	17	64.26	0.99395	17	69.12
11	0.99181	19	63.28	0.99378	15	68.14
12	0.99162	16	62.29	0.99363	11	67.15
13	0.99147	18	61.30	0.99352	14	66.15
14	0.99129	25	60.31	0.99338	11	65.16
15	0.99104	45	59.32	0.99327	10	64.17
16	0.99059	50	58.35	0.99317	18	63.18
17	0.99009	52	57.38	0.99299	29	62.19
18	0.98957	85	56.41	0.99270	35	61.21
19	0.98873	79	55.46	0.99235	30	60.23
20	0.98795	70	54.50	0.99205	35	59.24
21	0.98726	71	53.54	0.99170	31	58.27



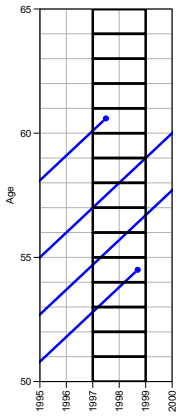


Practical

Based on the previous slides answer the following for both Danish and Swedish lifetables:

- What is the doubling time for mortality?
- What is the rate-ratio between males and females?
- How much older should a woman be in order to have the same mortality as a man?

Observations for the lifetable



Life table is based on person-years and deaths accumulated in a short period.

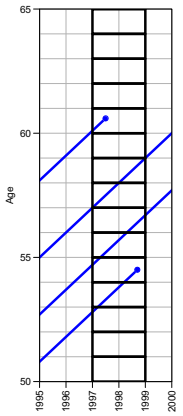
Age-specific rates — cross-sectional!

Survival function:

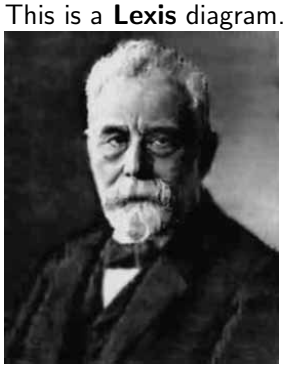
$$S(t) = \mathrm{e}^{-\int_0^t \lambda(a)\,\mathrm{d}a} = \mathrm{e}^{-\sum_0^t \lambda(a)}$$

— assumes stability of rates to be interpretable for actual persons.

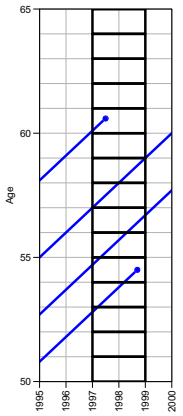
Observations for the lifetable



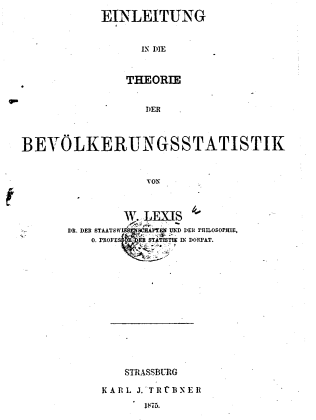
This is a **Lexis** diagram.



Observations for the lifetable



This is a **Lexis** diagram.


EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

VON

W. LEXIS

STRASSBURG

KARL J. TRÜBNER

1875.

Life table approach

The observation of interest is **not** the survival time of the **individual**.

It is the **population** experience:

- D: Deaths (events).
- Y: Person-years (risk time).

The classical lifetable analysis compiles these for prespecified intervals of age, and computes age-specific mortality **rates**.

Data are collected crossectionally, but interpreted longitudinally.

Classical estimators

Tuesday 1 June 2010, afternoon

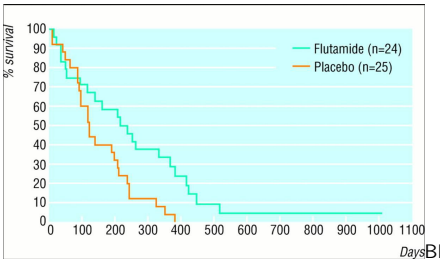
Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

The Kaplan-Meier Method 1

- The most common method of estimating the survival function.
- A non-parametric method.
- Divides time into small xintervals where the intervals are defined by the unique time points.
- 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.

Example of KM Survival Curve from BMJ



1998;316:1935-1938 Kaplan-Meier curve from an RCT of patients with pancreatic cancer

Calculating the Kaplan-Meier estimator

An estimate of $S(t_k)$ is:

$$\widehat{S}(t_k) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right)\ldots\left(1 - \frac{d_k}{n_k}\right)$$

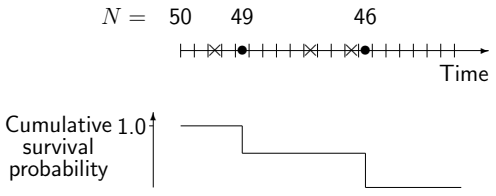
or more simply:

$$\widehat{S}(t_k) = \prod_{i=1}^{k} 1 - \frac{d_i}{n_i}$$

$$\widehat{S}(t_k) = \widehat{S}(t_{k-1})\left(1 - \frac{d_k}{n_k}\right)$$

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

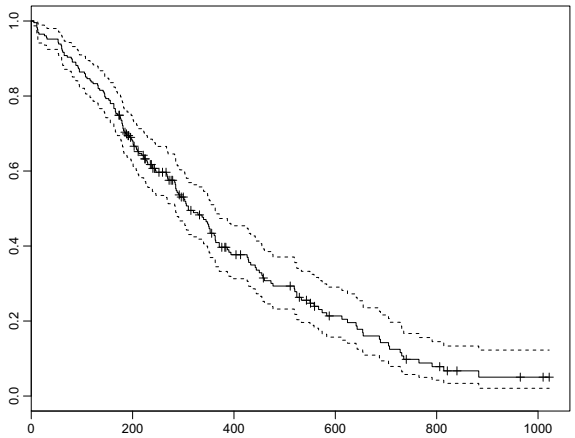
Using R: Surv()

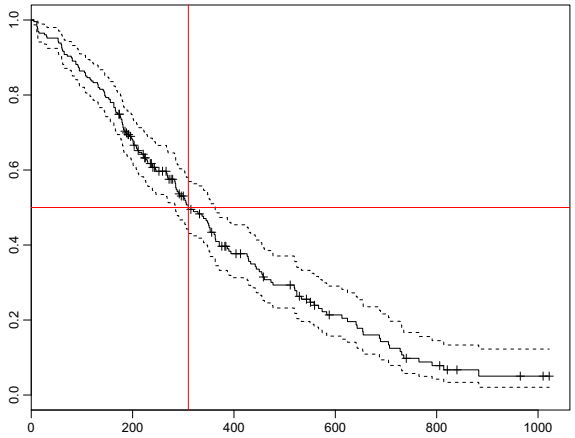
```
> with( lung, Surv( time, status==2 ) )
  [1]  306   455  1010+  210   883  1022+  310   361   218   166
 [12]  654   728    71   567   144   613   707    61    88   301
 [23]  624   371

> ( s.km <- survfit( Surv( time, status==2 ), data=lung ) )
Call: survfit(formula = Surv(time, status == 2), data = lung)

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

> plot( s.km )
```



The Cox model

Tuesday 1 June 2010, afternoon

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010

Modelling Survival Data

- As with other types of data we are interested in fitting a *statistical model* to survival data.
- Most modelling principles are the same.
- In epidemiology it is customary to model on the hazard scale. For example, by how much does being exposed to factor X increase/decrease the hazard rate.

Proportional Hazards model

Consider the following model:

$$\lambda_i(t, \mathbf{x}_i) = \lambda_0(t)\exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\right)$$

- $\lambda_i(t, \mathbf{x}_i)$ is the hazard rate for the i^{th} subject.
- $\lambda_0(t)$ is the *baseline hazard* function - a non-linear effect of the *covariate* t.
- $\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$ is the linear predictor.

The proportional hazards 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.
- Conceptually it is less clear — t is but a covariate that varies *within* each individual.

Cox-likelihood

The partial likelihood for the regression parameters:

$$\ell(\beta) = \sum_{\text{death times}} \log\left(\frac{\mathrm{e}^{x_{\text{death}}\beta}}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{x_i\beta}}\right)$$

- This is David Cox's invention.
- Extremely efficient from a computaionel point of view.
- The baseline hazard is bypassed.

Proportional Hazards model

- The baseline hazard rate, $\lambda_0(t)$, is the hazard rate when all the covariates are 0.
- The form of the above equation means that covariates act *multiplicatively* on the baseline hazard rate.
- The baseline hazard is a function of time and thus varies with time.
  Time is a covariate (albeit with special status).
- The proportionality assumption means that the difference between two groups can be summarised by one number. This is because the (relative) effect of a covariate is assumed to be the same throughout the time-scale.

The Cox Proportional Hazards likelihood

- By far the most common model applied to time-to-event outcomes.
- The Cox PH model does not make any assumption about the shape of the underlying hazard function.
- However, it does make the assumption that the hazard rates for patient subgroups are proportional over time.
- The Cox model models the hazard function, $\lambda_i(t; x_i)$ where x_i denotes the covariate vector.

Proportional Hazards Model

- Parameters are estimated on log scale:

$$\lambda_i(t) = \lambda_0(t)\exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\right)$$

$$\log\left(\lambda_i(t)\right) = \log\left(\lambda_0(t)\right) + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$$

- The baseline hazard is the hazard rate when all covariate values are equal to zero.
- Estimates of the parameters, β, are obtained by maximizing the partial likelihood.

Interpreting Regression Coefficients

- How do we interpret the parameters of interest?
- In a Cox model the baseline hazard $\lambda_0(t)$ is not included in the partial likelihood and so we only obtain estimates of the regression coefficients associated with each of the covariates.
- Consider a binary covariate x_1 which takes the values 0 and 1.

Interpreting Regression Coefficients

- The model is

$$\lambda_i(t) = \lambda_0(t)\exp(\beta_1 x_{1i})$$

- The hazard rate when $x_1 = 0$ is $\lambda_0(t)$.
- The hazard rate when $x_1 = 1$ is $\lambda_0(t)\exp(\beta_1)$.
- The hazard ratio is therefore

$$\frac{\lambda_0(t)\exp(\beta)}{\lambda_0(t)}$$

- The $\lambda_0(t)$ cancels: β_1 is the log hazard ratio.
- Exponentiate β_1 to get the hazard ratio.

Interpreting Regression Coefficients

- If x_j is binary $\exp(\beta_j)$ is the estimated hazard ratio for subjects corresponding to $x_j = 1$ compared to those where $x_j = 0$.
- If x_j is continuous $\exp(\beta_j)$ is the estimated increase/decrease in the hazard rate for a unit change in x_j.
- With more than one covariate interpretation is similar, i.e. $\exp(\beta_j)$ is the hazard ratio for subjects who *only* differ with respect to covariate x_j.

Fitting a Cox- model in R

```
> data(bladder)
> bladder <- subset( bladder, enum<2 )
> head( bladder)
   id rx number size stop event enum
1   1  1      1    3    1     0    1
5   2  1      2    1    4     0    1
9   3  1      1    1    7     0    1
13  4  1      5    1   10     0    1
17  5  1      4    1    6     1    1
21  6  1      1    1   14     0    1
```


Fitting a Cox- model in R

```
> c0 <- coxph( Surv(stop,event) ~ number + size, data=bladder )
> c0
Call:
coxph(formula = Surv(stop, event) ~ number + size, data = bladde

         coef exp(coef) se(coef)     z      p
number 0.2049      1.23   0.0704 2.912 0.0036
size   0.0613      1.06   0.1033 0.594 0.5500

Likelihood ratio test=7.04  on 2 df, p=0.0296  n= 85
```


Plotting the base survival in R

```
> plot( survfit(c0) )
> lines( survfit(c0), conf.int=F, lwd=3 )
```

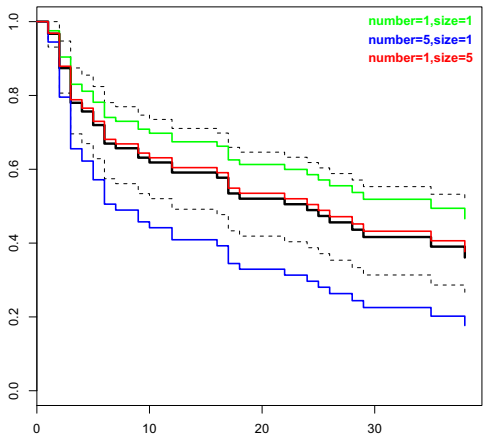
The `plot.coxph` plots the survival curve for a person with an average covarite value.

— which is not the average survival for the population considered. . .

Plotting the base survival in R

```
> plot( survfit(c0) )
> lines( survfit(c0), conf.int=F, lwd=3 )
> lines( survfit(c0,newdata=data.frame(number=1,size=1)), lwd=2,
> text( par("usr")[2]*0.98, 1.00, "number=1,size=1", col="green"
```

You can plot the survival curve for specific values of the covariates, using the `newdata=` argument.



Follow-up data

Tuesday 1 June 2010, afternoon

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

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:
  - Along age
  - Along calendar time
- Multiple timescales.

Representation of follow-up data

In a cohort study we have records of:
Events and **Risk time**.

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

- Date of entry — date variable.
- Date of exit — date variable
- Status at exit — indicator-variable (0/1)

Specific for each *type* of outcome.

Aim of dividing time into bands:

Put $\begin{matrix} D \text{ — events} \\ Y \text{ — risk time} \end{matrix}$ in intervals on the timescale:

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?

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

- Define strata: 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
Status at exit:	Dead	Alive	Dead
Y	31.89	22.70	6.58
D	1	0	1

	subj. 1		subj. 2		subj. 3		$\sum$	
Age	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
$\sum$	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

- but what if we want to keep 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.

Representation of follow-up on several timescales

- The time followed is the same on all timescales.
- Only use 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.

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

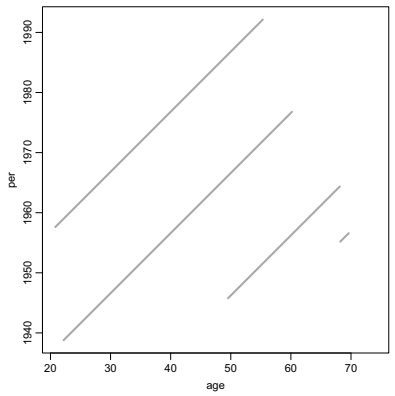
Definition of `Lexis` object

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

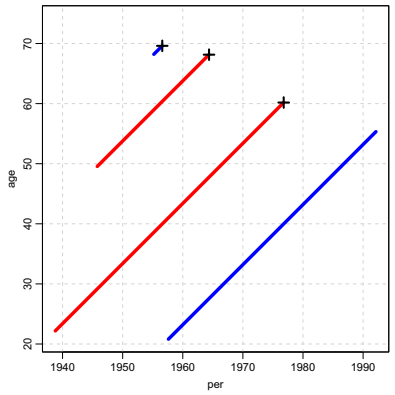
`entry` is defined on **three** timescales,
but `exit` is only defined on **one** timescale:
Follow-up time is the same on all timescales.

The looks of a `Lexis` object

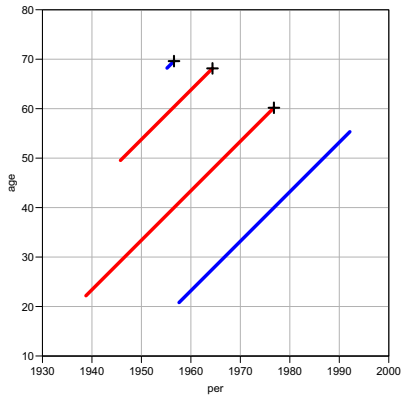
```
> round( thL[,c(1:8,14,15)], 2 )
    age     per tfi lex.dur lex.Cst lex.Xst lex.id   ic
1 22.18 1938.79   0   38.00       0       1      1    1
2 49.55 1945.77   0   18.60       0       1      2  640
3 68.21 1955.18   0    1.40       0       1      3 3425
4 20.80 1957.61   0   34.52       0       0      4 4017
```


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


```
> plot( thL, 2:1, lwd=5, col=c("red","blue")[thL$contrast], grid=T )
> points( thL, 2:1, pch=c(NA,3)[thL$lex.Xst+1],lwd=3, cex=1.5 )
```


```
> 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, "age", breaks=seq(0,100,20) )
> round( spl1, 2 )
  lex.id   age     per   tfi lex.dur lex.Cst lex.Xst   id sex bi
1      1 22.18 1938.79  0.00   17.82       0       0    1   2  1
2      1 40.00 1956.61 17.82   20.00       0       0    1   2  1
3      1 60.00 1976.61 37.82    0.18       0       1    1   2  1
4      2 49.55 1945.77  0.00   10.45       0       0  640   2  1
5      2 60.00 1956.23 10.45    8.14       0       1  640   2  1
6      3 68.21 1955.18  0.00    1.40       0       1 3425   1  1
7      4 20.80 1957.61  0.00   19.20       0       0 4017   2  1
8      4 40.00 1976.81 19.20   15.33       0       0 4017   2  1
```

Split on another timescale

```
> # Split further on tfi:
> spl2 <- splitLexis( spl1, "tfi", breaks=c(0,1,5,20,100) )
> round( spl2, 2 )
   lex.id   age     per   tfi lex.dur lex.Cst lex.Xst   id sex b
1       1 22.18 1938.79  0.00    1.00       0       0    1   2
2       1 23.18 1939.79  1.00    4.00       0       0    1   2
3       1 27.18 1943.79  5.00   12.82       0       0    1   2
4       1 40.00 1956.61 17.82    2.18       0       0    1   2
5       1 42.18 1958.79 20.00   17.82       0       0    1   2
6       1 60.00 1976.61 37.82    0.18       0       1    1   2
7       2 49.55 1945.77  0.00    1.00       0       0  640   2
8       2 50.55 1946.77  1.00    4.00       0       0  640   2
9       2 54.55 1950.77  5.00    5.45       0       0  640   2
10      2 60.00 1956.23 10.45    8.14       0       1  640   2
11      3 68.21 1955.18  0.00    1.00       0       0 3425   1
12      3 69.21 1956.18  1.00    0.40       0       1 3425   1
13      4 20.80 1957.61  0.00    1.00       0       0 4017   2
14      4 21.80 1958.61  1.00    4.00       0       0 4017   2
15      4 25.80 1962.61  5.00   14.20       0       0 4017   2
```


The Poisson likelihood for time-split data

Split records (one per person-**interval** (i,t)):

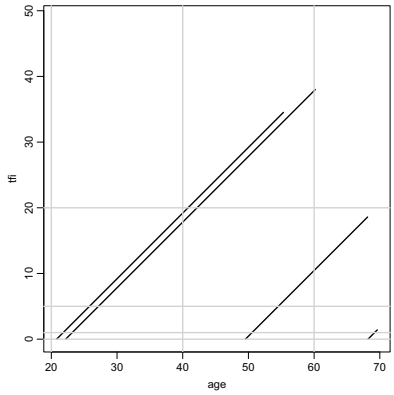
$$D\ln(\lambda) - \lambda Y = \sum_{i,t}\left(d_{it}\ln(\lambda) - \lambda y_{it}\right)$$

Assume that the death indicator ($d_i \in \{0,1\}$) is Poisson, with log-offset y_i will give the same result.

Model assumes that rates are constant.

But the split data allows models that assume different rates for different (d_{it}, y_{it}).

Where are the (d_{it}, y_{it}) in the split data?




```
plot( spl2, c(1,3), col="black", lwd=2 )
```


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

```
> round( spl2, 2 )
   lex.id   age     per   tfi lex.dur lex.Cst lex.Xst   id sex b
1       1 22.18 1938.79  0.00    1.00       0       0    1   2
2       1 23.18 1939.79  1.00    4.00       0       0    1   2
3       1 27.18 1943.79  5.00   12.82       0       0    1   2
4       1 40.00 1956.61 17.82    2.18       0       0    1   2
5       1 42.18 1958.79 20.00   17.82       0       0    1   2
6       1 60.00 1976.61 37.82    0.18       0       1    1   2
7       2 49.55 1945.77  0.00    1.00       0       0  640   2
8       2 50.55 1946.77  1.00    4.00       0       0  640   2
9       2 54.55 1950.77  5.00    5.45       0       0  640   2
10      2 60.00 1956.23 10.45    8.14       0       1  640   2
11      3 68.21 1955.18  0.00    1.00       0       0 3425   1
12      3 69.21 1956.18  1.00    0.40       0       1 3425   1
13      4 20.80 1957.61  0.00    1.00       0       0 4017   2
14      4 21.80 1958.61  1.00    4.00       0       0 4017   2
15      4 25.80 1962.61  5.00   14.20       0       0 4017   2
16      4 40.00 1976.81 19.20    0.80       0       0 4017   2
17      4 40.80 1977.61 20.00   14.52       0       0 4017   2
```


Analysis of results

- d_i — events in the variable: `lex.Xst`.
- y_i — risk time: `lex.dur` (Δ_t!).
  Enters in the model via $\log(y)$ as offset.
- 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.

Poisson model for split data

- Each interval contribute λY to the log-likelihood.
- All intervals with the same set of covariate values (age,exposure,...) have the same λ.
- The log-likelihood contribution from these is $\lambda \sum Y$ — the same as from aggregated data.
- The event intervals contribute each $D\log\lambda$.
- The log-likelihood contribution from those with the same lambda is $\sum D\log\lambda$ — the same as from aggregated data.
- The log-likelihood is the same for split data and aggregated data — no need to tabulate first.

Who needs the Cox-model anyway?

Wednesday 2 June 2010, morning

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

The proportional hazards 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.
- Conceptually it is less clear — t is but a covariate that varies within individual.

Cox-likelihood

The partial likelihood for the regression parameters:

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

is also a *profile likelihood* in the model where observation time has been subdivided in small pieces (empirical rates) and each small piece provided with its own parameter:

$$\log\big(\lambda(t,x)\big) = \log\big(\lambda_0(t)\big) + x'\beta = \alpha_t + \eta$$

The Cox-likelihood as profile likelihood

Regression parameters describing the effect of covariates (other than the chosen underlying time scale).
One parameter per death time to describe the effect of time (i.e. the chosen timescale).

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

Suppose the time scale has been divided into small intervals with at most one death in each.

Assume w.l.o.g. the ys 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:

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

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

The derivative w.r.t. α_t is:

$$\mathrm{D}_{\alpha_t}\ell(\alpha_t,\beta) = 1-\mathrm{e}^{\alpha}_t\sum_{i\in\mathcal{R}_t}\mathrm{e}^{\eta_i} = 0 \quad\Leftrightarrow\quad \mathrm{e}^{\alpha}_t = \frac{1}{\sum_{i\in\mathcal{R}_t}\mathrm{e}^{\eta_i}}$$

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

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

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

What the Cox-model really is

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

- dividing time as finely as possible,
- modelling one covariate, the time-scale, with one parameter per distinct value,
- profiling these parameters out by maximizing the profile likelihood

Subsequently, one may recover the effect of the timescale by smoothing an estimate of the cumulative sum of these.

Sensible modelling

Replace the α_ts by a parmetric function $f(t)$ with a limited number of parameters, for example:

- Piecewise constant
- Splines (linear, quadratic or cubic)
- Fractional polynomials

Use Poisson modelling software on a dataset of empirical rates for small intervals (ys).

Splitting the dataset

The Poisson approach needs a dataset of empirical rates with small values of y.

Larger than the original: each individual contributes many empirical rates. From each empirical rate we get:

- Poisson-response d
- Risk time y
- Covariate value for the timescale (time since entry, current age, current date, ...)
- other covariates

Example: Mayo Clinic lung cancer

```
  time status age sex
1  306      2  74   1
2  455      2  68   1

> Lx <- Lexis( exit=list( tfd=time), exit.status=(status==2), da
NOTE: entry is assumed to be 0 on the tfd timescale.

> tab(Lx,scale=365.25)
States:
     #records:
     To
From    FALSE TRUE Sum  #events:  #risk time:      Rate      (95
  FALSE    63  165 228       165     190.5352 0.8659815 0.743432

> dx <- splitLexis( Lx, "tfd", breaks=c(0,unique(Lx$time)) )
> tab( dx, scale=365.25 )
States:
     #records:
     To
From    FALSE TRUE   Sum  #events:  #risk time:      Rate      (
  FALSE 19857  165 20022       165     190.5352 0.8659815 0.7434
```


The baseline hazard and survival functions

Using a parametric function to model the baseline hazard gives the possibility to plot this with confidence intervals for a given set of covariate values, x_0

The survival function in a multiplicative Poisson model has the form:

$$S(t) = \exp\Big(-\sum_{\tau<t} \exp(g(\tau) + x_0'\gamma)\Big)$$

This is just a non-linear function of the parameters in the model, g and γ. So the variance can be computed using the δ-method.

δ-method for survival function

1. Select timepoints t_i (fairly close).
2. Get estimates of log-rates $f(t_i) = g(t_i) + x_0'\gamma$ for these points:

   $$\hat{f}(t_i) = \mathbf{B}\,\hat{\beta}$$

   where β is the total parameter vector in the model.
3. Variance-covariance matrix of $\hat{\beta}$: $\hat{\Sigma}$.
4. Variance-covariance of $\hat{f}(t_i)$: $\mathbf{B}\Sigma\mathbf{B}'$.
5. Transformation to the rates is the coordinate-wise exponential function, with derivative $\mathrm{diag}\big[\exp\big(\hat{f}(t_i)\big)\big]$

6. Variance-covariance matrix of the rates at the points t_i:

   $$\mathrm{diag}(\mathrm{e}^{\hat{f}(t_i)})\,\mathbf{B}\,\hat{\Sigma}\,\mathbf{B}'\,\mathrm{diag}(\mathrm{e}^{\hat{f}(t_i)})'$$

7. Transformation to cumulative hazard (ℓ is interval length):

$$\ell\times\begin{bmatrix}1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\end{bmatrix}\begin{bmatrix}\mathrm{e}^{\hat{f}(t_1))}\\\mathrm{e}^{\hat{f}(t_2))}\\\mathrm{e}^{\hat{f}(t_3))}\\\mathrm{e}^{\hat{f}(t_4))}\end{bmatrix} = \mathbf{L}\begin{bmatrix}\mathrm{e}^{\hat{f}(t_1))}\\\mathrm{e}^{\hat{f}(t_2))}\\\mathrm{e}^{\hat{f}(t_3))}\\\mathrm{e}^{\hat{f}(t_4))}\end{bmatrix}$$

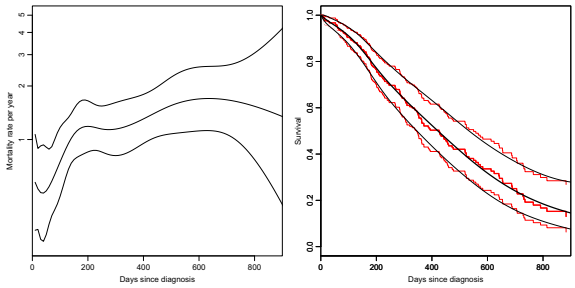
8. Variance-covariance matrix for the cumulative hazard is:

   $$\mathbf{L}\,\mathrm{diag}(\mathrm{e}^{\hat{f}(t_i)})\,\mathbf{B}\,\hat{\Sigma}\,\mathbf{B}'\,\mathrm{diag}(\mathrm{e}^{\hat{f}(t_i)})'\,\mathbf{L}'$$

This is all implemented in the `ci.cum()` function in `Epi`.

Mayo clinic lung cancer data

Smoothing by natural splines with 7 parameters; knots at $0, 25, 75, 150, 250, 500, 1000$ days



Computational tools for time-splitting

R: A function `splitLexis`, written by Martyn Plummer, included in the package `Epi` available at `http://wwww.biostat.ku.dk/~bxc/Epi` or CRAN.

Stata: The function `stsplit` (part of standard Stata).
Descendant of `stlexis` written by Michael Hills & David Clayton.

SAS: A macro `%Lexis`, available at `http://wwww.biostat.ku.dk/~bxc/Lexis`.

Modelling rates

Wednesday 2 June 2010, morning

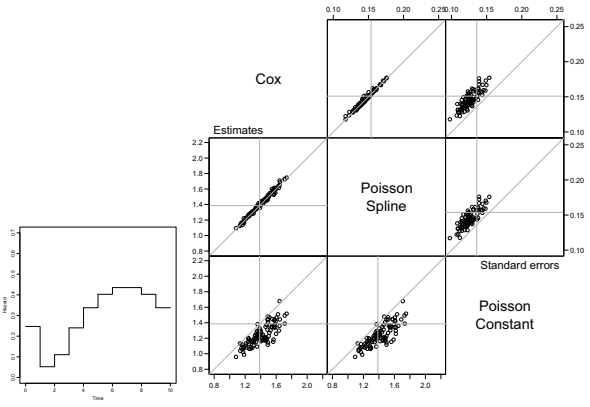
Bendix Carstensen

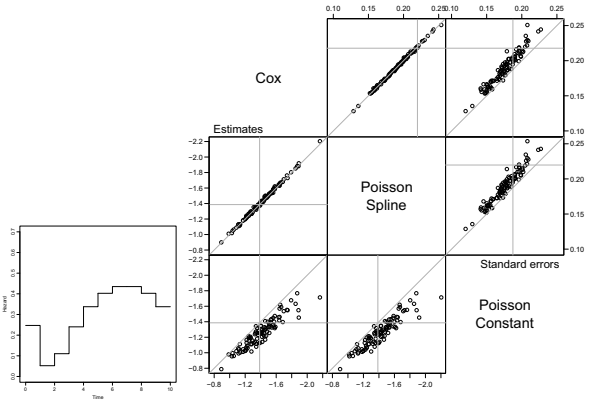
Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

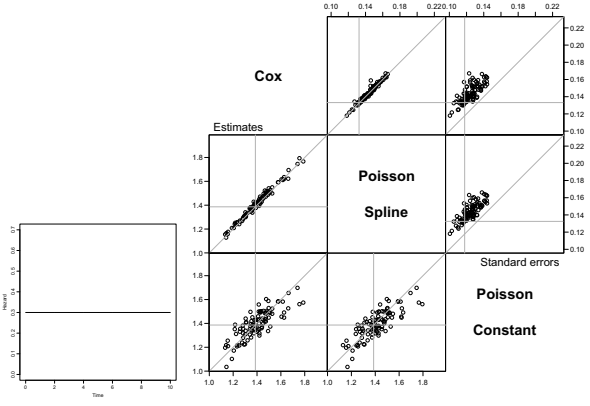
Any difference in covariate effects?

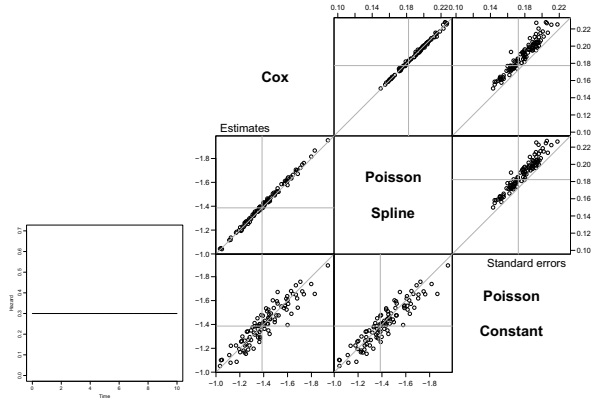
Simulation study:
100 survival datasets, 200 individuals in each.
Baseline hazard varying, censoring at time 10.
Two covariates, one standard normal with rate-ratio of 4 and the other log-normal with rate-ratio of 0.25.

For each dataset three models fitted:
1. standard Cox-model.
2. Poisson model using natural splines, 6 baseline parameters.
3. Poisson-model using constant baseline, 1 parameter.









Computational aspects

- Cox model:
  - Only one timescale.
  - Each person contributes one (or very few) records.
  - Computationally simple, because time (risk / covariate) is profiled out in the estimation.
- Poisson modelling:
  - Many records per person.
  - Very large datasets.
  - Any number of timescales.
  - Timeconsuming due to the full modelling of the rates.

Historical aspects

Whitehead J: Fitting Cox's regression model to survival data using GLIM. Applied Statistics, 29(3):268–275, 1980.[1]

Set up tables of event counts and person-years, classified by event times and covariate patterns.

Even with moderate datasets this can be large, albeit smaller than some 100 separate records per person.

[1] Recall **Keiding's law**: "Any result was published earlier than you think, even if you take Keiding's law into account."

Computational practicalities

Early 1980s: Fitting of Poisson models on datasets with 50,000 records were out of the question.
In particular with 100+ parameters.

Computationally feasible approaches to cohort studies were:

- Cox modelling — tanks to computational elegance.
- Time-splitting and tabulation before modelling.

Time-splitting and tabulation.

Man-years and PYRS programs:

Follow-up of each person was put into a table of (current) age-class by calendar time: Cut by the grid in a Lexis diagram. Possibly also classified by time since entry.

The tables of (D, Y) generated directly (disk space limitations prevented storage of the split dataset).

Used for SMR analysis, by merging with tables of population mortailty rates. Analyses based on a manageable number of analytical units.

The tabulation legacy (curse)

The **computational** need for tabulation has influenced thinking in epidemiology / demography:

- Life-tables in 1-year intervals.
- Rates are regarded in 5-year age by period intervals. Used for analysis of mortality and incidence rates based on registers. Age-period-cohort models with one parameter per level of the age/period factor.
- Yet, survival analysis is largely based on "time to event" methods (Kaplan-Meier, Cox), even from cancer registries.

The period method for survival analysis

H. Brenner, O. Gefeller & T. Hakulinen: Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications European Journal of Cancer **40**, (2004), pp. 326–335

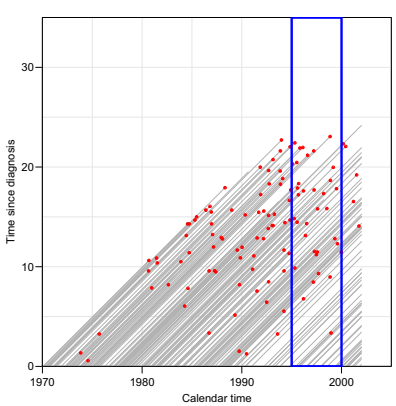
This method of survival analysis is designed to take interactions between two time-scale into account:

Mortality rates at a given time since entry into the study (usually diagnosis of cancer) depends on the current calendar time.

Brenner *et al.* propose to restrict analysis to the most recent period and then report results by survival curves.

Period analysis reports survival curve based on data from the blue rectangle.

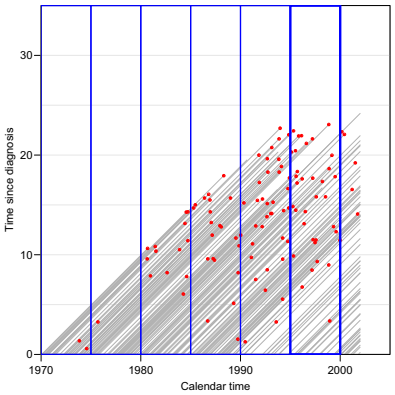
Interaction between current date and time since diagnosis.



Interaction between current date and time since diagnosis.

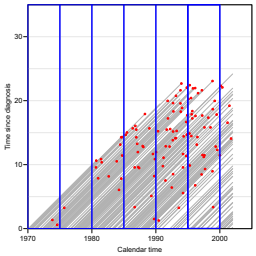
Separate survival curves for each period.

Period analysis reports the last set of parameters, because it is *clinically* the most relevant.



Interaction between current date and time since diagnosis:

- Separate survival curves for each period.
- Stratified Cox-model with time-dependent strata.
- In practical terms, data are split by (current) calendar time (period), and interactions with this are introduced throughout the model.



Using the Lexis diagram today

Rates are observed as little *empirical rates* (d, y), several per individual.

These vary by several *timescales*

- current age
- calendar time
- time since entry

and fixed covariates

- age at entry
- date of entry
- date of birth
- sex
- ...

Stratified Cox-model

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

The key is the "s" — separate baseline for each stratum.

In plain words:
The effect of time depends on s — an interaction between time and stratum.

Test of "proportionality" is merely a test of interaction between time and some (categorical) covariate.

Age at entry as covariate

t: time since entry
e: age at entry
$a = e + t$: current age

$$\log\big(\lambda(a,t)\big) = f(t) + \beta e = (f(t) - \beta t) + \beta a$$

Immaterial whether a or e is used as (log)-*linear* covariate as long as t is in the model.

In a Cox-model with time since entry as time-scale, only the baseline hazard will change if age at entry is replaced by current age (a time-dependent variable).

Non-linear effects of time-scales

Arbitrary effects of the three variables t, a and e:
$\Longrightarrow$ genuine extension of the model.

$$\log\big(\lambda(a,t,x_i)\big) = f(t) + g(a) + h(e) + \eta_i$$

Three quantities can be arbitrarily moved between the three functions:

$$\begin{aligned}\tilde{f}(t) &= f(a) - \mu_a - \mu_e + \gamma t\\ \tilde{g}(a) &= g(p) + \mu_a \qquad - \gamma a\\ \tilde{h}(e) &= h(c) \qquad + \mu_a + \gamma e\end{aligned}$$

because $t - a + e = 0$.
This is the age-period-cohort modelling problem again.

"Controlling for age"

— is not a well defined statement.

Mostly it means that age *at entry* is included in the model.

But ideally one would check whether there were non-linear effects of age at entry and current age.

This would require modelling of multiple timescales.

Which is best accomplished by splitting time.

SMR

Wednesday 2 June 2010, afternoon

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

Cohorts where all are exposed

When there is no comparison group we may ask:
Do mortality rates in cohort differ from those of an **external** population, for example:

Rates from:

- Occupational cohorts
- Patient cohorts

compared with reference rates obtained from:

- Population statistics (mortality rates)
- Disease registers (hospital discharge registers)

Log-likelihood

Cohort rates proportional to reference rates:
$\lambda(a) = \theta \times \lambda_R(a)$ — the same in all age-bands.

D_a deaths during Y_a person-years an age-band a gives the likelihood:

$$\begin{aligned}D_a\log(\lambda(a)) - \lambda(a)Y_a &= D_a\log(\theta\lambda_R(a)) - \theta\lambda_R(a)Y_a\\ &= D_a\log(\theta) + D_a\log(\lambda_R(a))\\ &\quad -\theta(\lambda_R(a)Y_a)\end{aligned}$$

The constant $D_a\log(\lambda_R(a))$ does not involve θ, and so can be dropped.

The term $\lambda_R(a)Y_a = E_a$ is the "expected" number of cases in age a, so the log-likelihood for age a is:

$$D_a\log(\theta) - \theta(\lambda_R(a)Y_a) = D_a\log(\theta) - \theta(E_a)$$

Note: $\lambda_R(a)$ is known for all values of a. The total log-likelihood is:

$$D\log(\theta) - \theta E$$

Therefore:

$$\hat{\theta} = \frac{D}{\lambda_R Y} = \frac{D}{E} = \frac{\text{Observed}}{\text{Expected}} = \text{SMR}$$

SMR is the maximum likelihood estimator of the relative mortality in the cohort.

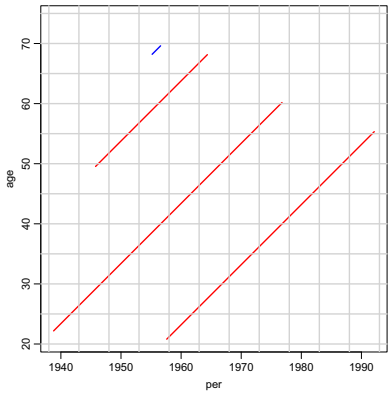
Accounting for age composition

- Compare rates in a study group with a standard set of age–specific rates.
- Reference rates are normally based on large numbers of cases, — assumed known.
- Calculate "expected" number of cases, $E_a = \lambda_R(a)Y_a$, and compare this with the observed number of cases, D:
- SMR is based on a log-likelihood similar to that for a rate — Y is replaced by E:

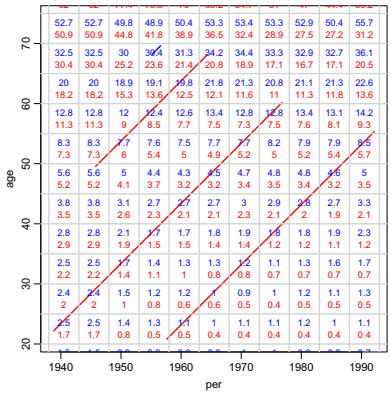
$$\text{SMR} = \frac{D}{E}, \qquad \text{s.d.}\big(\log(\text{SMR})\big) = \frac{1}{\sqrt{D}}$$

Modelling the SMR

- As for the rates, the SMR can be modelled using individual data.
- Response is d_i, the event indicator (`lex.Xst`).
- $\log$-offset is the expected value for each piece of follow-up, $e_i = y_i \times \lambda_R$.
- λ_R is the population rate corresponding to the age, period and sex of the follow-up period y_i.




```
plot( thap, 2:1, col=c("blue","red")[thap$sex], lwd=2 )
```


```
plot( thap, 2:1, col=c("blue","red")[thap$sex], lwd=2 )
...
```


Split the data to fit with population data

```
> # Split the data for SMR-analysis
> tha  <- splitLexis(thL, "age", breaks=seq(0,90,5) )
> thap <- splitLexis(tha, "per", breaks=seq(1938,2038,5) )
> dim( thap )
[1] 41 15
> # Create variables to fit with the population data
> thap$agr <- timeBand( thap, "age", "left" )
> thap$cal <- timeBand( thap, "per", "left" )
> round( thap[,c("lex.id","age","agr","per","cal","lex.dur","lex
   lex.id   age agr     per  cal lex.dur lex.Xst sex
1       1 22.18  20 1938.79 1938    2.82       0   2
2       1 25.00  25 1941.61 1938    1.39       0   2
3       1 26.39  25 1943.00 1943    3.61       0   2
4       1 30.00  30 1946.61 1943    1.39       0   2
5       1 31.39  30 1948.00 1948    3.61       0   2
6       1 35.00  35 1951.61 1948    1.39       0   2
7       1 36.39  35 1953.00 1953    3.61       0   2
8       1 40.00  40 1956.61 1953    1.39       0   2
9       1 41.39  40 1958.00 1958    3.61       0   2
```


Merge with population data

```
> thapx <- merge( thap, gmortDK[,c("agr","cal","sex","rt")] )
> str( thapx )
Classes 'Lexis' and 'data.frame': 41 obs. of  18 variables:
 $ sex     : num  1 2 2 2 2 2 2 2 2 2 ...
 $ agr     : num  65 20 20 20 25 25 25 25 30 30 ...
 $ cal     : num  1953 1938 1953 1958 1938 ...
 $ lex.id  : int  3 1 4 4 1 1 4 4 1 1 ...
 $ age     : num  68.2 22.2 20.8 21.2 25.0 ...
 $ per     : num  1955 1939 1958 1958 1942 ...
 $ tfi     : num  0.000 0.000 0.000 0.389 2.818 ...
 $ lex.dur : num  1.405 2.818 0.389 3.806 1.391 ...
 $ lex.Cst : num  0 0 0 0 0 0 0 0 0 0 ...
 $ lex.Xst : num  1 0 0 0 0 0 0 0 0 0 ...
 $ id      : num  3425    1 4017 4017    1 ...
 $ birthdat: num  1887 1917 1937 1937 1917 ...
 $ contrast: num  2 1 2 2 1 1 2 2 1 1 ...
 $ injecdat: num  1955 1939 1958 1958 1939 ...
 $ volume  : num  0 22 0 0 22 22 0 0 22 22 ...
 $ exitdat : num  1957 1977 1992 1992 1977 ...
```


Calculation of the SMR

```
> thapx$E <- thapx$lex.dur * thapx$rt / 1000
> stat.table( contrast,
+             list( D = sum( lex.Xst ),
+                   Y = sum( lex.dur ),
+                   E = sum( E ),
+                 SMR = ratio( lex.Xst, E ) ),
+              margin = TRUE,
+                data = thapx )
 ----------------------------------------------
 contrast         D       Y       E     SMR
 ----------------------------------------------
 1             2.00   56.59    0.33    6.02
 2             1.00   35.93    0.11    8.70

 Total         3.00   92.52    0.45    6.71
 ----------------------------------------------
```


Modelling the SMR

```
> m.SMR <- glm( lex.Xst ~ factor(contrast)-1+offset(log(E)),
+               family=poisson, data=thapx )
> round( ci.lin( m.SMR, Exp=TRUE )[,5:7], 3 )
                  exp(Est.)  2.5%  97.5%
factor(contrast)1     6.023 1.506 24.082
factor(contrast)2     8.698 1.225 61.745
```

- Analysis of SMR is like analysis of rates:
- Replace Y with E — that's all!

Interactions and timescales

Thursday 3 June 2010, morning

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

Computational aspects of fitting models

- Cox model:
  - Only one timescale.
  - Each person contributes one (or very few) records.
  - Computationally simple, because time (risk / covariate) is profiled out in the estimation.
- Poisson modelling:
  - Many records per person.
  - Very large datasets.
  - Any number of timescales.
  - Timeconsuming due to the full modelling of the rates.

Historical aspects

Whitehead J: Fitting Cox's regression model to survival data using GLIM. Applied Statistics, 29(3):268–275, 1980.[**?**][2]

Set up tables of event counts and person-years, classified by event times and covariate patterns.

Even with moderate datasets this can be large, albeit smaller than some 100 separate records per person.

[2] Recall **Keiding's law**: "Any result was published earlier than you think, even if you take Keiding's law into account."

Computational practicalities

Early 1980s: Fitting of Poisson models on datasets with 50,000 records were out of the question.
In particular with 100+ parameters.

Computationally feasible approaches to cohort studies were:

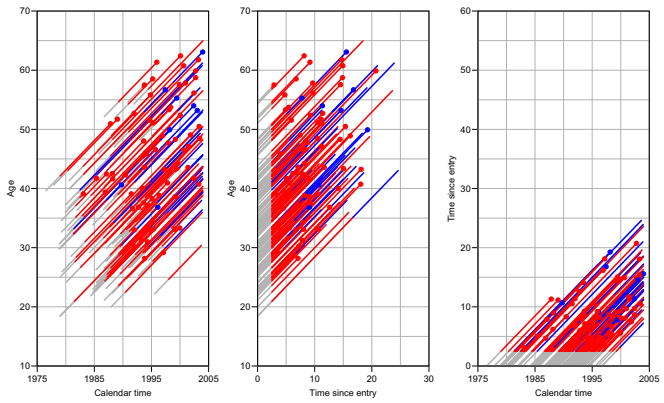
- Cox modelling — thanks to computational elegance.
- Time-splitting and tabulation before modelling.

The tabulation legacy (curse)

The **computational** need for tabulation has influenced thinking in epidemiology / demography:

- Life-tables in 1-year intervals.
- Rates are regarded in 5-year age by period intervals. Used for analysis of mortality and incidence rates based on registers.
  Age-period-cohort models with one parameter per level of the age/period factor.
- Yet, survival analysis is largely based on "time to event" methods (Kaplan-Meier, Cox), even from cancer registries — only one timescale.

Representation of follow-up



Age at entry as covariate

t: time since entry
e: age at entry
$a = e + t$: current age

$$\log\big(\lambda(a,t)\big) = f(t) + \beta e = \big(f(t) - \beta t\big) + \beta a$$

Immaterial whether a or e is used as (log)-*linear* covariate as long as t is in the model.

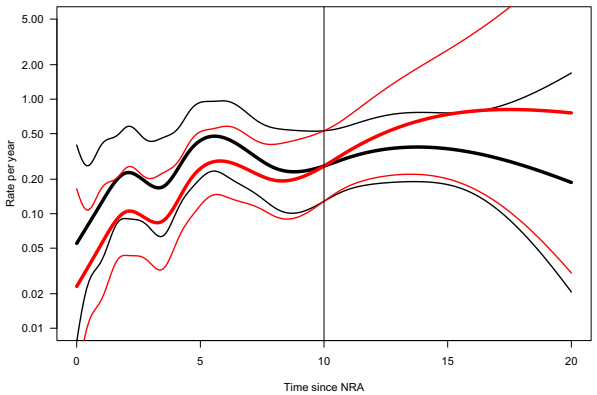
In a Cox-model with time since entry as time-scale, only the baseline hazard will change if age at entry is replaced by current age (a time-dependent variable).

"Controlling for age"

Including age at entry:

- Linear effect.
- Grouped variable.
- Parametric function.

— still only controls for the *linear* effect of *current age*.



Current age as covariate
Age at entry as covariate

Non-linear effects of time-scales

Arbitrary effects of the three variables t, a and e: Genuine extension of the model.

$$\log\big(\lambda(a,t,x_i)\big) = f(t) + g(a) + h(e) + \eta_i$$

Three quantities can be arbitrarily moved between the three functions:

$$\begin{aligned}\tilde{f}(t) &= f(a) - \mu_a - \mu_e + \gamma t\\ \tilde{g}(a) &= g(p) + \mu_a \quad\quad - \gamma a\\ \tilde{h}(e) &= h(c) \quad\quad + \mu_a + \gamma e\end{aligned}$$

because $t - a + e = 0$.
How many timescales in this model?

"Controlling for age"

— is not a well defined statement.

Mostly it means that age *at entry* is included in the model.

But ideally one would check whether there were non-linear effects of age at entry and current age.

This would require modelling of multiple timescales.

Which is best accomplished by splitting time and modelling the timescales explicitly.

Several timescales: Caveat

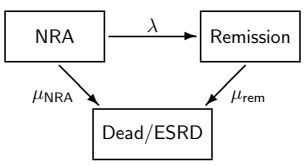
As an example, consider:
t: time since entry
e: age at entry
$a = e + t$: current age

The relation: $a = t + e$ must hold for all units of analysis.

In general: The difference between two time-scales must be constant within individuals.

The Boyle-Robertson fallacy from age-period-cohort models, where units with identical values of (current) age, a, and (current) period p had varying values of cohort, date of birth $c = p - a$! [**?**].

Several timescales



Cox-model:
— One dataset per transition.
— Combine datasets and make relevant interactions.
— Timescale must be the same.

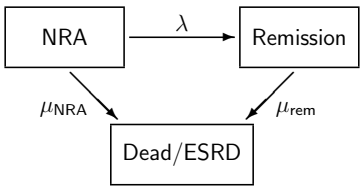
Poisson-model:
— One time-split dataset per transition.
— Combine datasets and make relevant interactions.
— Timescales can be different, and multiple timsecales can be accomodated simultaneously; duration of NRA, for example.

Time dependent variable

How does remission influence the mortality?

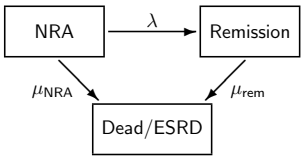
$$\lambda(t) = \lambda_0(t)\exp\big(1\{\text{remission}\}(t) \times \beta\big)$$

i.e. when remission occurs, mortality increase by e^{β}.



What transitions are modelled here?

Time-dependent variable



If we take

$$1\{\text{remission}\}(t)$$

as time-dependent variable, we assume that μ_{NRA} and μ_{rem} are proportional on the same timescale — no disease duration!.

— and λ is not modelled at all.

Stratified model

A popular version of the Cox-model allowing for non-proportionality is the **stratified model**:

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

where s refers to levels of a factor S.

This is but a completely general **interaction** between the factor S and the chosen timescale.

A better approach to interactions would be to specify a clinically founded form of interaction, so that test for interaction is against a specific (and sensible) alternative.

Time varying coefficients

This is a concept introduced by letting (some of) the parameters depend on time:

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

This is also an interaction, but restricted:
The effect of a covariate is linear for any value of t.

If the covariate is a factor, then we just have a reparametrization of the stratified model.

Poisson modelling of interactions

When interactions are needed (or desired):

- use the familiar terminology of interaction as known from (generalized) linear models.
- use clinical judgement of which interactions are relevant.
- use clinical judgement of which forms of interaction are relevant.
- are interactions with time of special interest?

Poisson model for time-split data

- Clarifies the destinction between (risk) time as response variable and time(scales) as covariates.
- Multiple timescales easily handled.
- Hazard rates by standard methods.
- More credible estimates of survival functions.
- Sensible modelling of interactions between timescales and other variables (and between timescales).
- Interactions are called interactions.

Multistate models

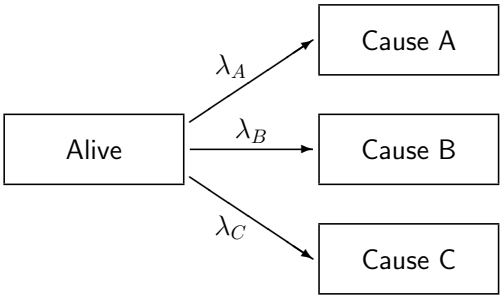
Thursday 3 June 2010, afternoon

Bendix Carstensen

Modern Demographic Methods in Epidemiology
1–3 June 2010
University of St. Andrews, Scotland
Longitudinal Studies Centre
`http://www.biostat.ku.dk/~bxc/AdvCoh/StAn-2010`

Competing risks

You may die from more than one cause:



Cause-specific intensities

$$\lambda_A(t) = \lim_{h\to 0}\frac{\mathrm{P}\{\text{death from cause A in } (t,t+h] \mid \text{alive at } t\}}{h}$$
$$\lambda_B(t) = \lim_{h\to 0}\frac{\mathrm{P}\{\text{death from cause B in } (t,t+h] \mid \text{alive at } t\}}{h}$$
$$\lambda_C(t) = \lim_{h\to 0}\frac{\mathrm{P}\{\text{death from cause C in } (t,t+h] \mid \text{alive at } t\}}{h}$$

Total mortality rate:

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

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

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.

Assume for simplicity that rates are constant.

A survivor contributes to the log-likelohood:

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

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

The total log-likelihood is then:

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

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.

Time varying rates:

This is the same business as with one rate; use time intervals sufficiently small to justify an assumption of constant rate (intensity).

Practical implications

Analysis of the individual cause-specific rates effectively uses the same dataset for all causes, because the person-years are the same.

Thus the little "atoms" of data (the 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 $(0, y)$
— for the analysis of cause A treated as censorings.

Analysis of competing risks

Competing risks are analysed by considering the cause specific rates separately.

Joint modelling: Take the datasets for analysis of each of the causes, stack them including an indicator.


```
          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 type
 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
```


Implemented in the `stack.Lexis` function:

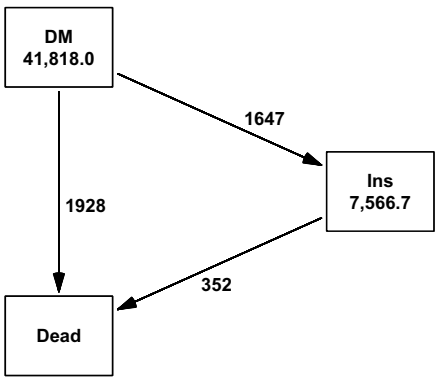
```
> data(DMlate)
> str(DMlate)
'data.frame': 10000 obs. of  6 variables:
 $ sex  : Factor w/ 2 levels "M","F": 1 1 2 1 1 1 1 1 1 2 ...
 $ dobth: num  1952 1951 1926 1923 1914 ...
 $ dodm : num  2006 2001 1996 1996 2002 ...
 $ dodth: num  NA NA 1996 NA 2002 ...
 $ doins: num  NA 2006 NA NA NA ...
 $ dox  : num  2008 2008 1996 2008 2002 ...
> 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",
                        pre="DM" )
> summary( dmi )

Transitions:
     To
From    DM  Ins Dead  Records:  Events: Risk time:  Persons:
  DM  6319 1647 1928      9894     3575   41817.98      9894
  Ins    0 1399  352      1751      352    7566.72      1751
  Sum 6319 3046 2280     11645     3927   49384.71      9998

boxes( dmi, boxpos=list(x=c(20,20,80),y=c(80,20,50)) )
```

Implemented in the `stack.Lexis` function:

```
> ls.dmi <- stack( dmi )

> str( ls.dmi )
Classes 'stacked.Lexis' and 'data.frame': 21539 obs. of  15 vari
 $ Per     : num  2006 2001 1996 1996 2002 ...
 $ Age     : num  53.3 50.6 70 72.5 87.7 ...
 $ DMdur   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ lex.dur : num  2.4586 4.7036 0.063 12.4709 0.0219 ...
 $ lex.Cst : Factor w/ 3 levels "DM","Ins","Dead": 1 1 1 1 1 1 1
 $ lex.Xst : Factor w/ 3 levels "DM","Ins","Dead": 1 2 3 1 3 1 2
 $ lex.Tr  : Factor w/ 3 levels "DM->Ins","DM->Dead",..: 1 1 1 1
 $ lex.Fail: logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
 $ lex.id  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex     : Factor w/ 2 levels "M","F": 1 1 2 1 1 1 1 1 1 2 ...
 $ dobth   : num  1952 1951 1926 1923 1914 ...
 $ dodm    : num  2006 2001 1996 1996 2002 ...
 $ dodth   : num  NA NA 1996 NA 2002 ...
 $ doins   : num  NA 2006 NA NA NA ...
 $ dox     : num  2008 2008 1996 2008 2002 ...
```

Implemented in the `stack.Lexis` function:

```
> options(digits=2)

> subset(dmi,lex.id==2)
 Per Age DMdur lex.dur lex.Cst lex.Xst lex.id sex dobth
2001  51   0.0     4.7      DM     Ins      2   M  1951
2006  55   4.7     2.2     Ins     Ins      2   M  1951

> subset(ls.dmi,lex.id==2)
 Per Age DMdur lex.dur lex.Cst lex.Xst    lex.Tr lex.Fail lex.id
2001  51   0.0     4.7      DM     Ins   DM->Ins     TRUE      2
2001  51   0.0     4.7      DM     Ins  DM->Dead    FALSE      2
2006  55   4.7     2.2     Ins     Ins Ins->Dead    FALSE      2
```


Analysis

- Interactions between all covariates (including time) and type:
  The same as separate analyses of the rates λ_A, λ_B and λ_C.
- No interaction with time:
  Same underlying baseline hazard.
- Only interaction with time:
  Same covariate effects for all causes of death.

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 likelihoods machinery work.

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

An excellent 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

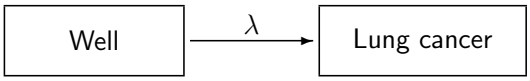
The paper includes a guide for the practitioner.

Also contains en example where both dependent and independent "cause specific survival times" gives rise to the same set of cause specific rates.

Competing risk problems

The problems with competing risk models comes when estimated intensities (rates) are used to produce probability statements.

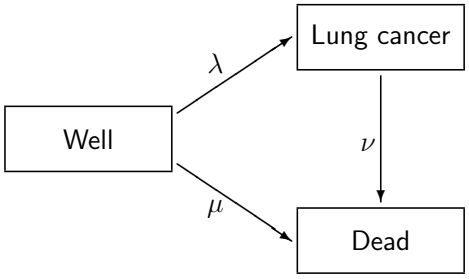
Classical set-up in cancer-registries:



$$\mathrm{P}\left\{\text{Lung cancer before age 75}\right\} = 1 - \mathrm{e}^{-\Lambda(75)}$$

This is not quite right.

How the world really looks



Illness-death model. Little boxes with arrows. (The mortality of lung cancer patients (ν) not relevant here).

How many get lung cancer before age a?

$$\mathrm{P}\left\{\text{Lung cancer before age 75}\right\} \neq 1 - \mathrm{e}^{-\Lambda(75)}$$

does not take the possibility of death prior to lung cancer into account.

$1 - \mathrm{e}^{-\Lambda(75)}$ often stated as the probability of lung cancer before age 75, assuming all other acuses of death absent.

Lung cancer rates are however observed in a mortal population.

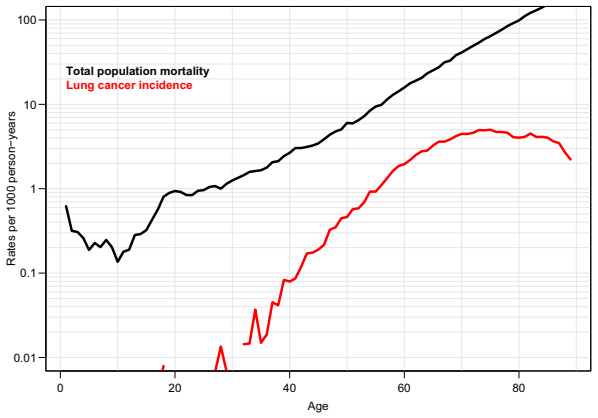
If all other causes of death were absent, this would assume that lung cancer rates remained the same.

$$\begin{aligned}
&\mathrm{P}\left\{\text{Lung cancer before age } a\right\} \\
&= \int_0^a \mathrm{P}\left\{\text{Lung cancer at age } u\right\}\,\mathrm{d}u \\
&= \int_0^a \mathrm{P}\left\{\text{Lung cancer in age } (u, u+\mathrm{d}u] \mid \text{alive at } u\right\} \\
&\qquad \times \mathrm{P}\left\{\text{alive at } u \text{ without lung cancer}\right\}\,\mathrm{d}u \\
&= \int_0^a \lambda(u)\exp\left(-\int_0^u \mu(s)+\lambda(s)\,\mathrm{d}s\right)\,\mathrm{d}u
\end{aligned}$$

Probability of lungcancer

The rates are easily plotted for inspection in R:

```
matplot( age, 1000*cbind( D/Y, lung/Y ),
         log="y", type="l", lty=1, lwd=3,
         ylim=c(0.01,100), xlab="Age",
         ylab="Rates per 1000 person-years" )
```

The probablility that a person contracts lung cancer before age a is (cf. the lecture notes):

$$\int_0^a \lambda(u)\exp\left(-\int_0^u \mu(s)+\lambda(s)\,\mathrm{d}s\right)\,\mathrm{d}u$$

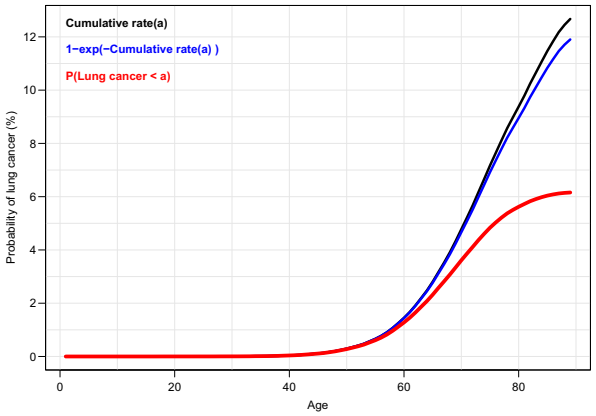
$$=\int_0^a \lambda(u)\exp\Big(-\big(\mathrm{M}(u)+\Lambda(u)\big)\Big)\,\mathrm{d}u$$

$\mathrm{M}(u)$ is the cumulative mortality rate.

$\Lambda(u)$ is the cumulative lung cancer incidence rate.

R-commands needed to do the calculations:

```
cr.death <- cumsum( D/Y )
cr.lung <- cumsum( lung/Y )
p.simple <- 1 - exp( -cr.lung )
p.lung <- cumsum( lung/Y *
                  exp( -(cr.death+cr.lung) ) )
matlines( age, 100*cbind( cr.lung, p.simple, p.lung ),
          type="l", lty=1, lwd=2*c(2,2,3),
          col=c("black","blue","red") )
```

Assumptions

The assumption behind the calculation and the statement "6% of Danish males will get lung cancer" is that the lung cancer rates and the mortality rates in the file applies to a cohort of men.

But they are cross-sectional rates, so the assumption is one of steady state of
1: mortality rates (which is dubious)
and
2: lung cancer incidence rates (which is appaling).

However the machinery can be applied to any set of rates for competing risks, regardless of how they were estimated.

Example: Renal failure data from Steno

Hovind P, Tarnow L, Rossing P, Carstensen B, and Parving H-H: Improved survival in patients obtaining remission of nephrotic range albuminuria in diabetic nephropathy. *Kidney Int.*, 66(3):1180–1186, 2004.

96 patients entering at nephrotic range albuminuria (NRA), i.e. U-alb>300mg/day.

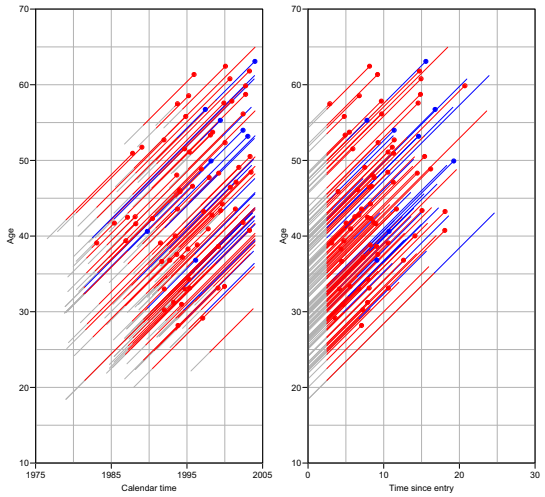
Is remission from this condition (i.e return to U-alb<300mg/day) predictive of the prognosis?

Endpoint of interest: Death or end stage renal disease (ESRD), i.e. dialysis or kidney transplant.

		Remission	
	Total	Yes	No
No. patients	125	32	93
No. events	77	8	69
Follow-up time (years)	1084.7	259.9	824.8

Cox-model:
Timescale: Time since nephrotic range albuminuria (NRA)
Entry: 2.5 years of GFR-measurements after NRA
Outcome: ESRD or Death

Estimates:	RR	95% c.i.	p
Fixed covariates:			
Sex (F vs. M):	0.92	(0.53,1.57)	0.740
Age at NRA (per 10 years):	1.42	(1.08,1.87)	0.011
Time-dependent covariate:			
Obtained remission:	0.28	(0.13,0.59)	0.001



Features of the analysis

- Remission is included a a time-dependent variable.
- Age at entry is included as a fixed variable.

```
renal[1:5,]
id       dob       doe       dor       dox event
17 1967.944 1996.013        NA 1997.094     2
26 1959.306 1989.535 1989.814 1996.136     1
27 1962.014 1987.846        NA 1993.239     3
33 1950.747 1995.243 1995.717 2003.993     0
42 1961.296 1987.884 1996.650 2003.955     0
```

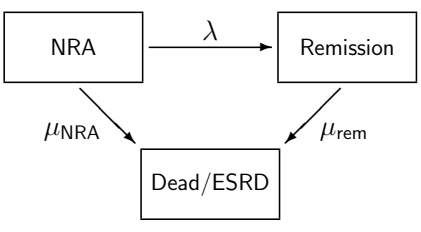
Note patient 26, 33 and 42 obtain remission.


```
renal[1:5,]
id       dob       doe       dor       dox event
17 1967.944 1996.013        NA 1997.094     2
26 1959.306 1989.535 1989.814 1996.136     1
27 1962.014 1987.846        NA 1993.239     3
33 1950.747 1995.243 1995.717 2003.993     0
42 1961.296 1987.884 1996.650 2003.955     0

> Lr <- Lexis( entry = list( per=doe,
+                            age=doe-dob,
+                            tfi=0 ),
+              exit = list( per=dox ),
+              exit.status = factor( event>0,
+                                    labels=c("NRA","ESRD") ),
+            data = renal )
NOTE: entry.status has been set to "NRA" for all.
> round( tab( Lr, scale=100 ), 2 )

States:
     #records:
     To
From  NRA ESRD Sum  #events:  #risk time:   Rate  (95% c.i.)
  NRA  48   77 125        77        10.85    7.1  5.68  8.88
```


Illness-death model



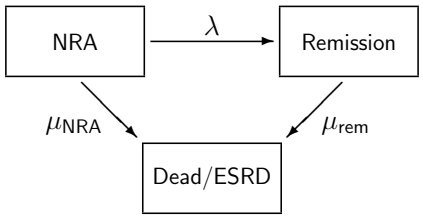
λ: remission rate.
μ_{NRA}: mortality/ESRD rate **before** remission.
μ_{rem}: mortality/ESRD rate **after** remission.

Cutting follow-up at remission: `cutLexis`

```
> Lc <- cutLexis( Lr, cut = Lr$dor,
+                 timescale = "per",
+                 new.state = "Rem",
+           precursor.states = "NRA" )

> subset( Lr[,-(8:11)], lex.id<3 )
       per      age tfi           lex.dur lex.Cst lex.Xst lex.id
1 1996.013 28.06879   0          1.081109     NRA    ESRD      1
2 1989.535 30.22895   0          6.600616     NRA    ESRD      2
> subset( Lc[,-(8:11)], lex.id<3 )
         per      age       tfi   lex.dur lex.Cst lex.Xst lex.id
1   1996.013 28.06879 0.0000000 1.0811088     NRA    ESRD      1
2   1989.535 30.22895 0.0000000 0.2789185     NRA     Rem      2
123 1989.814 30.50787 0.2789185 6.3216975     Rem    ESRD      2

> round( tab( Lc, scale=100), 2 )
States:
     #records:
     To
From  NRA ESRD Rem Sum  #events:  #risk time:   Rate  (95% c.i.)
  NRA  24   69  29 122        98         8.25  11.88  9.75 14.48
  Rem   0    8  24  32         8         2.60   3.08  1.54  6.16
  Sum  24   77  53 154       106        10.85   9.77  8.08 11.82
```

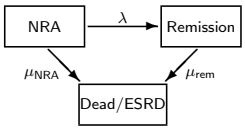
Cox-analysis with remission as time-dependent covariate:

— Ignores λ, the remission rate.

— Assumes μ_{NRA} and μ_{rem} use the same timescale.

— Duration, and timing of NRA modelled as covariates.

Model for all transitions



Cox-model:

One dataset per transition.

Combine datasets and make relevant interactions.

Same timescale.

Poisson-model:

One time-split dataset per transition.

Combine datasets and make relevant interactions.

Timescales can be different.

Multiple timescales can be accomodated simultaneously.

Calculus of probabilities

$$\mathrm{P}\{\text{Remission } \textbf{before} \text{ time } t\} = \int_0^t \lambda(u)\exp\left(-\int_0^u \lambda(s) + \mu_{\text{NRA}}\,\mathrm{d}s\right)\,\mathrm{d}u$$

$$\mathrm{P}\{\text{Being in remission } \textbf{at} \text{ time } t\} = \int_0^t \lambda(u)\exp\left(-\int_0^u \lambda(s) + \mu_{\text{NRA}}(s)\,\mathrm{d}s\right) \times \exp\left(-\int_u^t \mu_{\text{rem}}(s)\,\mathrm{d}s\right)\,\mathrm{d}u$$

Note μ_{rem} could also depend on u, time since obtained remission.

Sketch of programming:

```
c.rem      <- cumsum( lambda )
c.mort.nra <- cumsum( mu.nra )
c.mort.rem <- cumsum( mu.rem )
pr1  <- cumsum( lambda * exp( -( c.rem + c.mort.nra ) ) )

intgr(t,s) <- function(t,s){
           lambda[s] * exp( -( c.rem[s] + c.mort.nra[s] ) ) *
                     exp( -( c.mort.rem[t]-c.mort.rem[s] ) ) }
for( t in 1:100 ) p2[t] <- sum( intgr(t,1:t) )
```

If μ_{rem} depends on time of remission, then `c.mort.rem` should have an extra argument.

More complicated models: Simulation of probabilities.

(Outside the scope of this course).