Survival
Multiple timescales
Competing risks

Bendix Carstensen Steno Diabetes Center Copenhagen,
Gentofte, Denmark
b@bxc.dk
http://BendixCarstensen.com

IDEG 2019 training day, Seoul,
29 November 2019
http://BendixCarstensen/Epi/Courses/IDEG2019

Steno Diabetes Center Copenhagen

From /home/bendix/teach/Epi/IDEG2019/slides/slides.tex

Rates and Survival

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

`http://BendixCarstensen/Epi/Courses/IDEG2019`

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 ("censoring")

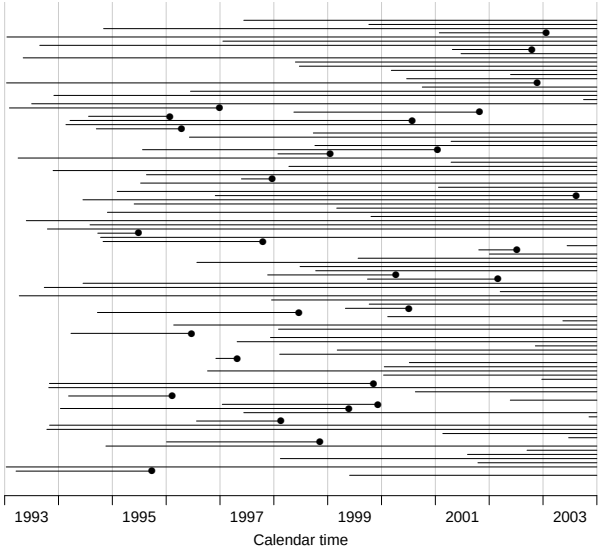
Examples of time-to-event measurements

- 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

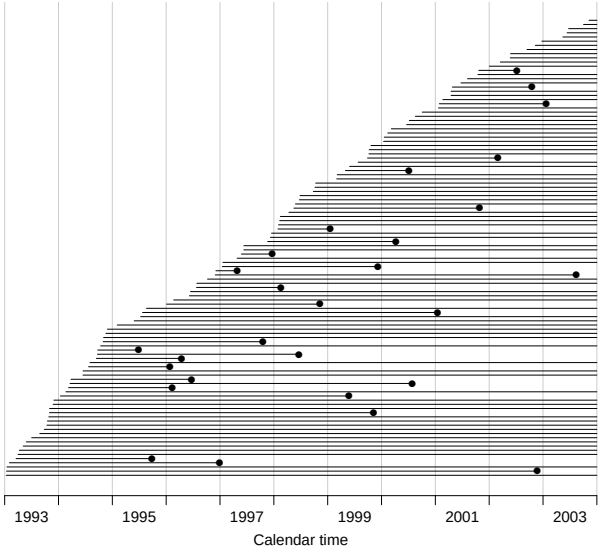
Each blob a death

Study ended at 31 Dec. 2003

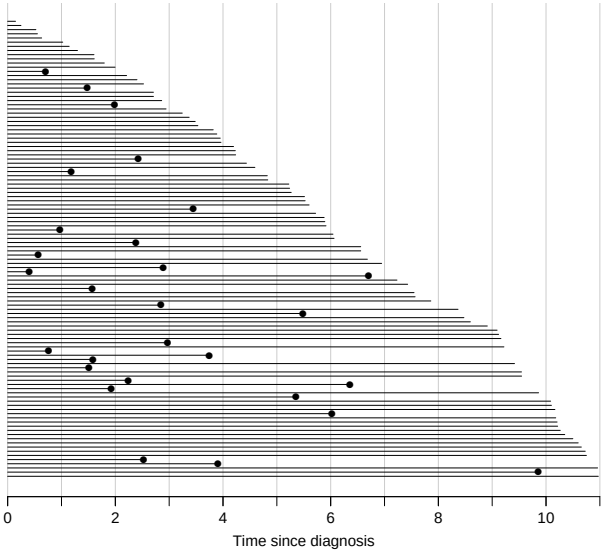


Ordered by date of entry

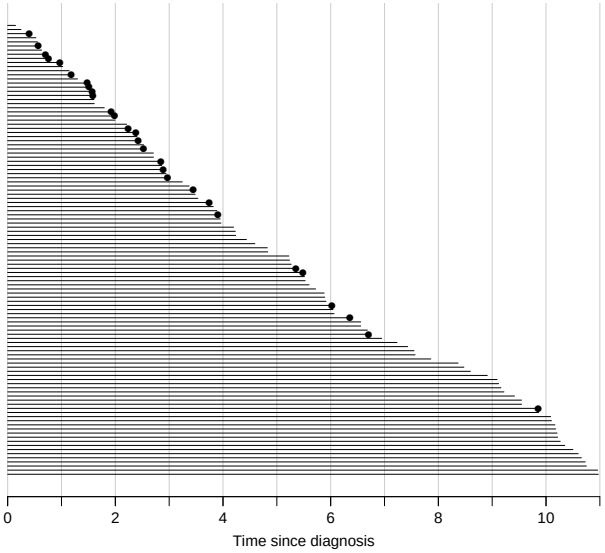
Most likely the order in your database.



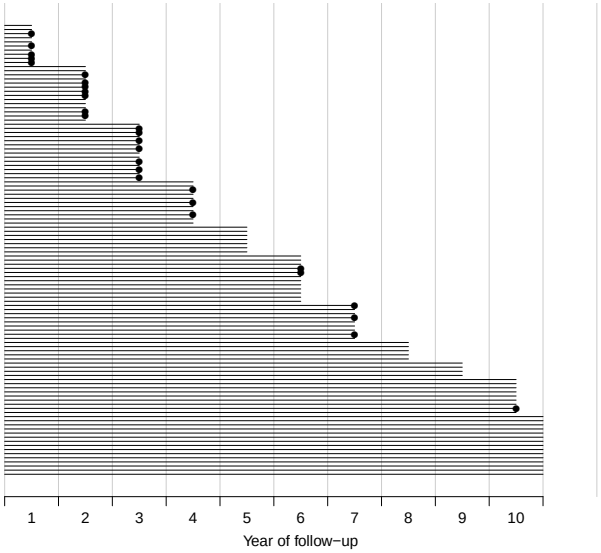
Timescale changed to "Time since diagnosis".



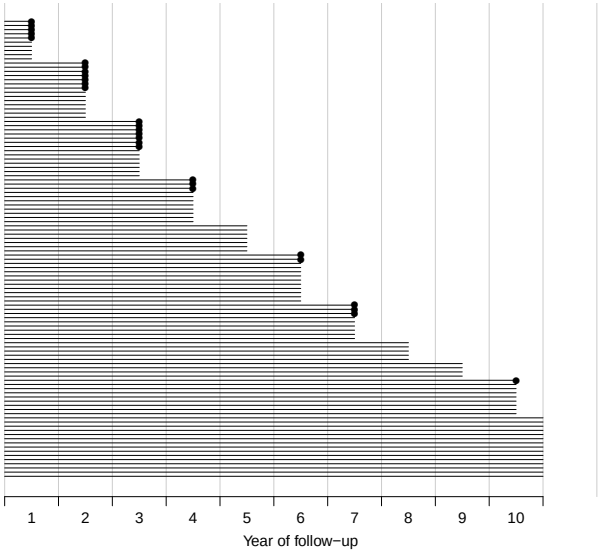
Patients ordered by survival time.



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:

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

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

Intensity / rate / hazard — same same

- The **intensity** or **hazard function**
- Probability of event in interval, relative to interval length:

$$\lambda(t) = \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

- Characterizes the distribution of survival times as does
  f (density) or
  F (cumulative distibution).
- Theoretical counterpart of a(n empirical) **rate**.

Survival and rate

Survival from rate — and vice versa;

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

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

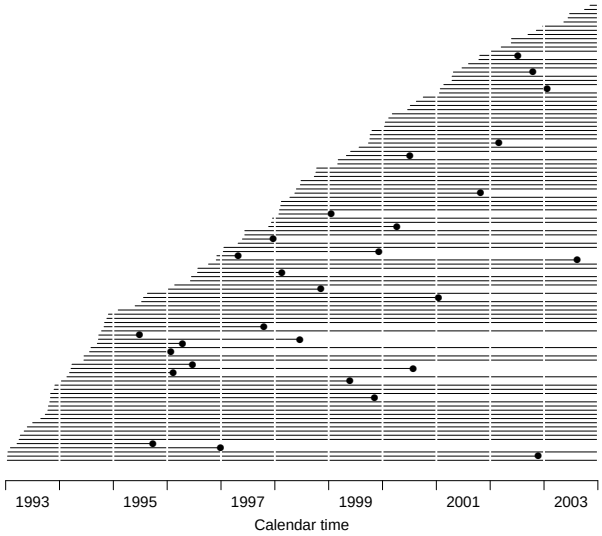
Note: A cumulative measure requires an origin!

. . . it is always survival **since** some timepoint — here 0

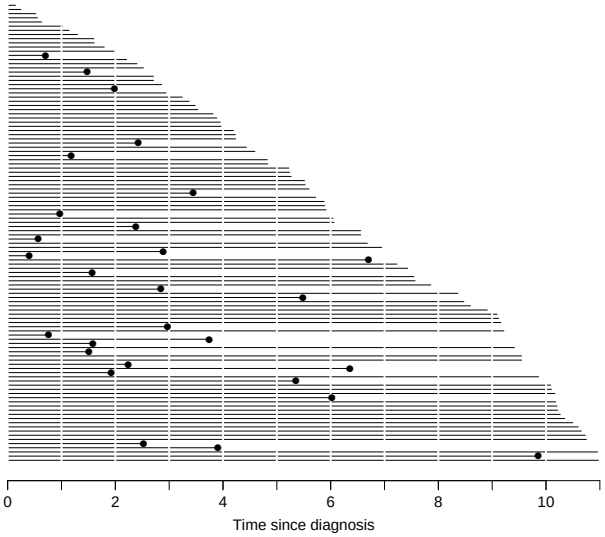
Empirical rates for individuals

- 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.

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:

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

- 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:

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

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

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

- $\log(\lambda)$ is modelled by covariates
- d is the response variable and
- $\log(y)$ is the offset variable, using the `poisson` family

Poisson likelihood

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

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

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

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

- $\log(\lambda)$ is modelled by covariates
- (d, y) is the response variable
- ... using the `poisreg` family

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
```

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
```

```
m2r <- glm( cbind(D,Y/1000) ~ gg, family=poisreg)
ci.exp( m2r )
```

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

Note the `family=poisreg`

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( cbind(D,Y/1000) ~ gg, family=poisreg )
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( cbind(D,Y/1000) ~ gg - 1, family=poisreg )
ci.exp( m3 )
```

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

You do it!

Lifetable estimators

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

`http://BendixCarstensen/Epi/Courses/IDEG2019` `ltab`

Survival analysis

- Response variable: Time to event, T
- Censoring time, Z
- 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, and therefore $Y = T - 0 = T$

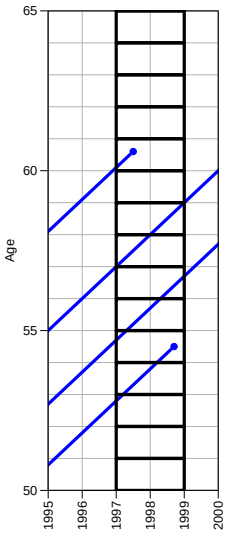
The life table method

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	$8/(59-1/2)=0.137$

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

Observations for the lifetable



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

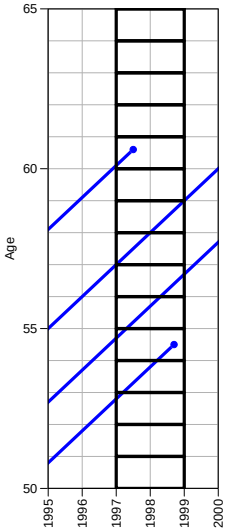
Age-specific rates — cross-sectional!

Survival function:

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

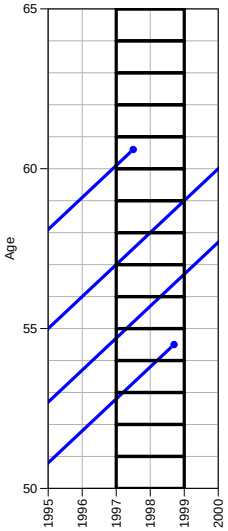
— assumes stability of rates to be interpretable for actual persons.

Observations for the lifetable

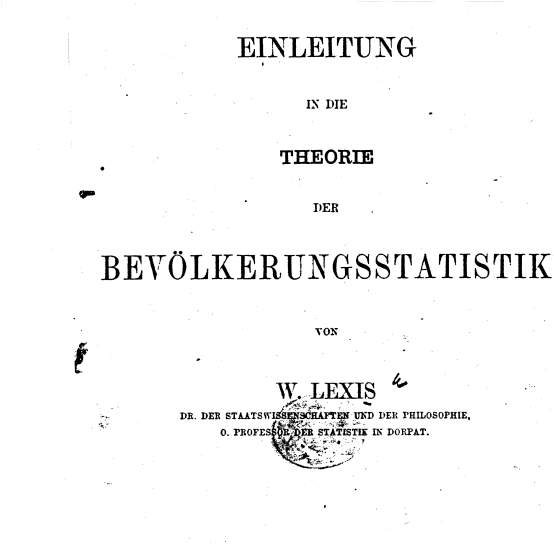


This is a **Lexis** diagram.

Observations for the lifetable



This is a **Lexis** diagram.



EINLEITUNG

IN DIE

THEORIE

DER

BEVÖLKERUNGSSTATISTIK

VON

W. LEXIS

DR. DER STAATSWISSENSCHAFTEN UND DER PHILOSOPHIE,
O. PROFESSOR DER STATISTIK IN DORPAT.

Life table approach

- 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.
- The **rates** are the basic building bocks — used for construction of:
  - RRs
  - cumulative measures (survival and risk)

Kaplan-Meier estimators

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

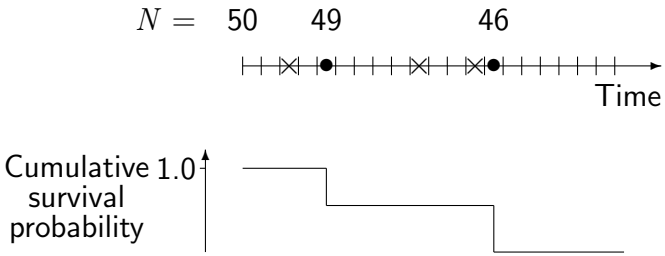
Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

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       0       90        90        NA      15
```

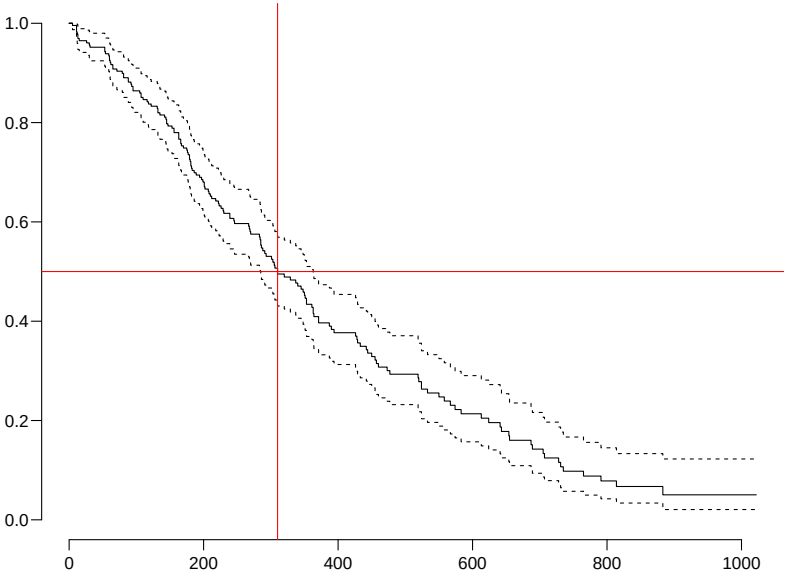
```
with( lung, Surv( time, status==2 ) )[1:10]
```

```
 [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" )
```


1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000

The Cox-model

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

The proportional hazards model

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

- The baseline hazard rate, $\lambda_0(t)$, is the hazard rate when all the covariates are 0
  — since then $\exp(x'\beta) = 1$
- The form of the above equation means that covariates act **multiplicatively** on the baseline hazard rate

The proportional hazards model

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

- Time (t) is a covariate (albeit modeled in a special way).
- The baseline hazard is a function of time and thus varies with time.
- No assumption about the shape of the underlying hazard function.
- — but you will never see the shape of the baseline hazard ...

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 between persons who **only** differ with respect to covariate x_j
- . . . assuming that the effect of x_j is the same across all other covariate values

Fitting a Cox- model in R

```
library( survival )
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 = bladder)

          coef exp(coef) se(coef)     z       p
number 0.20491   1.22742  0.07036 2.912 0.00359
size   0.06135   1.06327  0.10328 0.594 0.55254

Likelihood ratio test=7.04  on 2 df, p=0.02963
n= 85, number of events= 47
```

What is the meaning of the two regression parameters?

Plotting the base survival in R

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

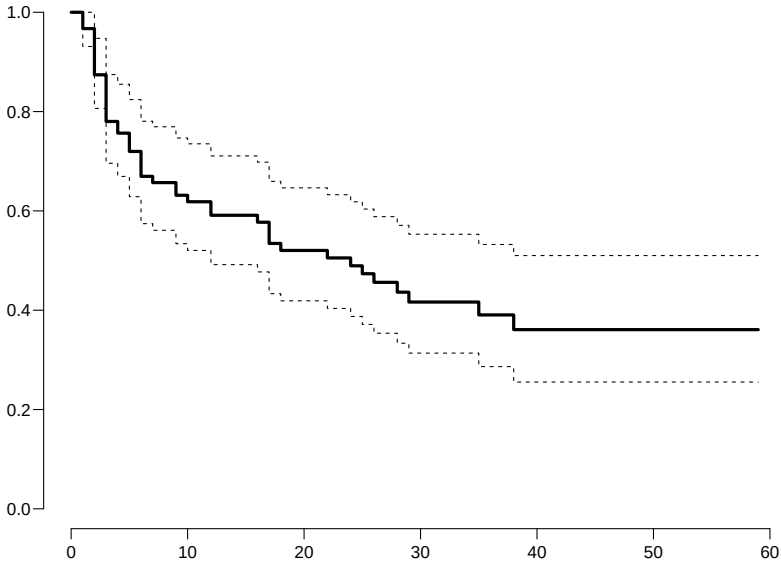
The `plot.coxph` plots the survival curve for a person with an average covariate value

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

— and not necessarily meaningful

```
c( mean(bladder$number), mean(bladder$size) )
```

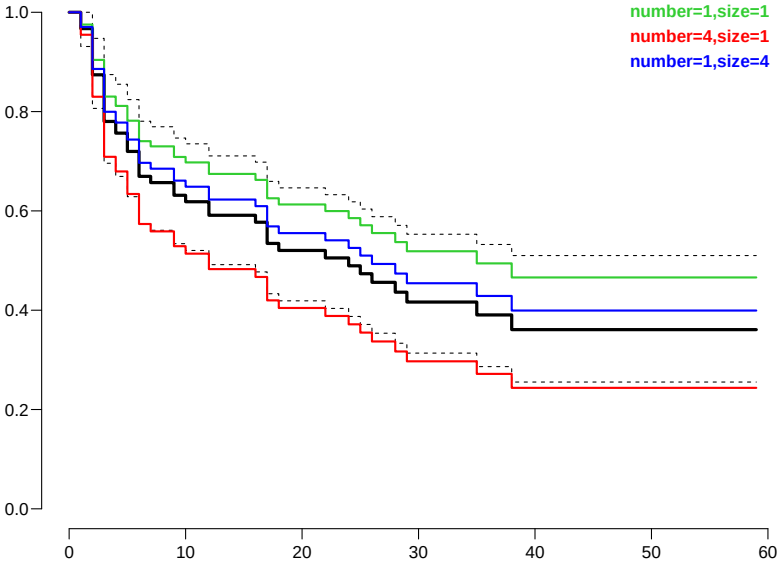
```
[1] 2.105882 2.011765
```


1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60

Plotting the base survival in R

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

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


number=1,size=1
number=4,size=1
number=1,size=4
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60

Who needs the Cox-model anyway?

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

http://BendixCarstensen/Epi/Courses/IDEG2019 WntCma

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.

The Cox-likelihood as profile likelihood

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

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

- Profile likelihood:
  - Derive estimates of α_t as function of data and βs — assuming constant rate between death times
  - Insert in likelihood, now only a function of data and βs
  - Turns out to be Cox's partial likelihood
  - The full likelihood is that of a Poisson model

Implications

- The Cox-model is a special case of a Poisson model
- . . . a model with one parameter per time (censoring or death) — typically hundreds of parameters
- A more sensible model would be one with a smooth effect of time.
- `bendixcarstensen.com/WntCma.pdf` gives a complete account
- . . . but here is a quick tour of how-to

```
library(Epi)
library(popEpi)
library(mgcv)
library(survival)
data(lung)
lung <- transform( lung, sex=factor(sex,labels=c("M","F")),
                         time=time+runif(nrow(lung)) )
```

Set up a Lexis object (outcome as a factor), and split time in small intervals (at all times):

```
Lx <- 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.
```

Split the follow-up in small intervals

```
sL <- splitMulti( Lx, tfe=c(0,sort(unique(Lx$lex.dur))) )
summary( Lx )
```

```
Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive    63  165       228      165   69703.91       228
```

```
summary( sL )
```

```
Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive 25941  165     26106      165   69703.91       228
```

The Cox model and the identical Poisson model on the `Lexis` data frames:

```
c0 <- coxph( Surv(tfe,tfe+lex.dur,lex.Xst=="Dead") ~ sex + age, data=Lx )
cx <- coxph.Lexis( Lx,           tfe  ~ sex + age )
```

```
survival::coxph analysis of Lexis object Lx:
Rates for the transition Alive->Dead
Baseline timescale: tfe
```

```
px <-   glm.Lexis( sL, ~ factor(tfe) + sex + age )
```

```
stats::glm Poisson analysis of Lexis object sL with log link:
Rates for the transition: Alive->Dead
```

```
length( coef(px) )
```

```
[1] 230
```

Fit smooth parametric model for baseline:

```
ps <- gam.Lexis( sL, formula= ~ s(tfe) + sex + age )
```

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

Compare estimates:

```
Ests <-cbind( rbind( ci.exp(cx,subset="sex"),
                     ci.exp(px,subset="sex"),
                     ci.exp(ps,subset="sex") ),
              rbind( ci.exp(cx,subset="age"),
                     ci.exp(px,subset="age"),
                     ci.exp(ps,subset="age") ) )
rownames(Ests) <- c("Cox","Pois-F","Pois-S")
colnames(Ests)[c(1,4)] <- c("sex","age")
round( Ests, 7 )
```

```
             sex      2.5%     97.5%      age      2.5%    97.5%
Cox    0.5989669 0.4313805 0.8316587 1.017154 0.9989336 1.035708
Pois-F 0.5989669 0.4313805 0.8316587 1.017154 0.9989336 1.035708
Pois-S 0.6017620 0.4335052 0.8353245 1.016415 0.9982477 1.034912
```

Prediction data frame for rates and survival — at what times do you want the rates and the survival shown for a 65 year old man, using the Poisson model with smooth effects:

```
ps <- gam.Lexis( sL, formula= ~ s(tfe) + sex + age )
```

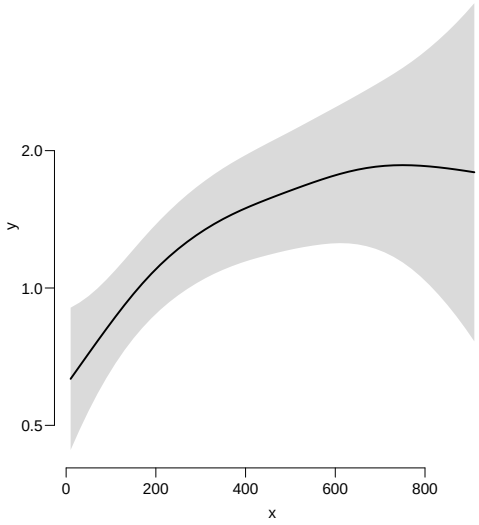
```
mgcv::gam Poisson analysis of Lexis object sL with log link:
Rates for the transition: Alive->Dead
```

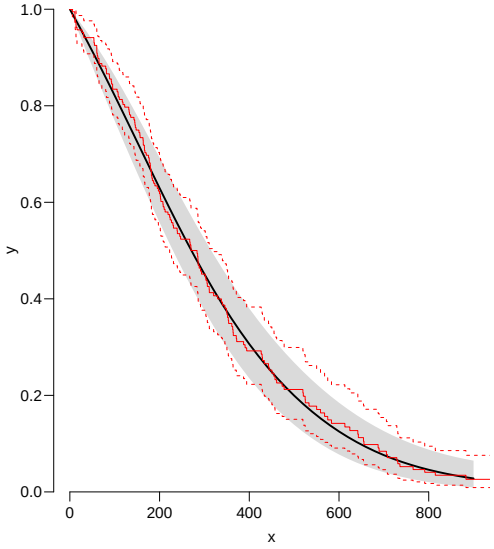
```
nd <- data.frame( tfe=seq(0,900,20)+10, sex="M", age=65 )
rate <- ci.pred( ps, nd )*365.25 # per year, not per day
surv <- ci.surv( ps, nd, int=20 ) # int is interval between times in nd
```

Plot the rates and the survival function for 65 year old man

```
par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
matshade( nd$tfe, rate, lwd=2, log="y", plot=TRUE )
matshade( nd$tfe-10, surv, lwd=2, yaxs="i", ylim=c(0,1), plot=TRUE )
lines( survfit( cx, newdata=nd[1,] ), col='red' )
```

Rates and survival, 65 year old man





Multiple time scales

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

Timescales

Mortality rates as a function of

- current age, a
- duration of diabetes, d
- age at diagnosis, $e = a - d$ (not a timescale!)
- $\Rightarrow a - d - e = 0$
  — this relation must be kept in any dataset

Model for mortality depending on current age and age at entry:

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

Two variables: age and age at diagnosis

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

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

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

We can claim any duration effect we like!

All three variables

Remember: $a - d - e = 0$

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

I makes no sense to show (any) one of the effects:

We can choose any slope for one of the effects, as long as we adjust the slopes of the two others.

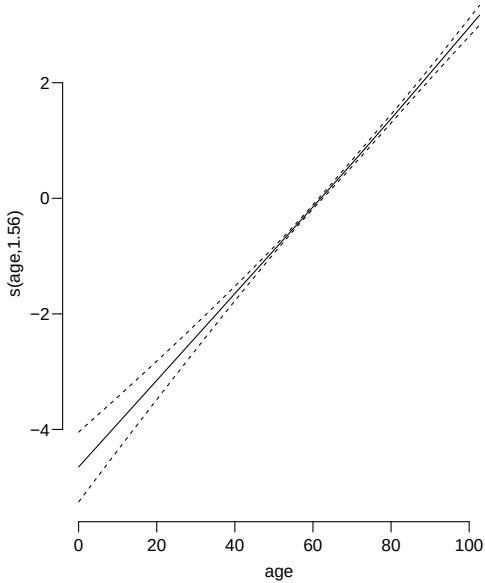
Predicted mortality

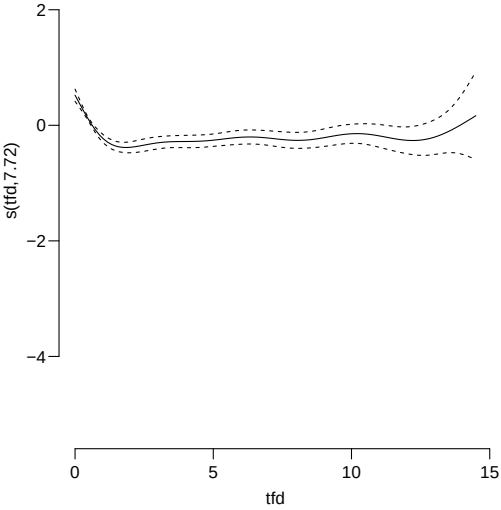
`age`: current age; `tfd`: duration; `ain`: age at DX:

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


anova( made, mad, test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(age) +
    s(tfd) + s(ain)
Model 2: cbind(trt(Lx$lex.Cst, Lx$lex.Xst) %in% trnam, Lx$lex.dur) ~ s(age) +
    s(tfd)
  Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
1    280378      24000
2    280378      24000 0.28932  0.42647   0.1664
```

. . . no **non-linear** effect of age at diagnosis—use model `mad`.





Predicted mortality

```
nd <- data.frame( expand.grid( tfd=c(NA,seq(0,14,.1)),
                               ain=c(3:7*10) ) )[-1,]
nd$age = nd$ain + nd$tfd
head( nd )
```

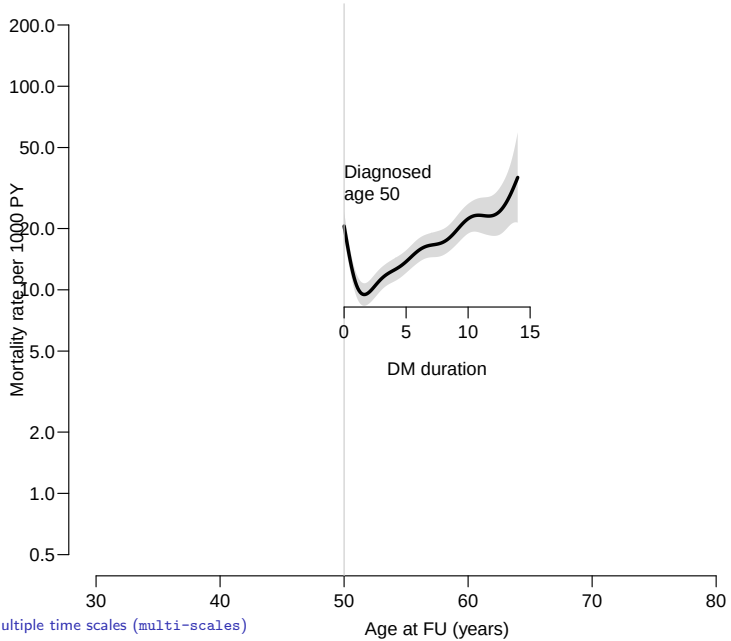
```
  tfd ain  age
2 0.0  30 30.0
3 0.1  30 30.1
4 0.2  30 30.2
5 0.3  30 30.3
6 0.4  30 30.4
7 0.5  30 30.5
```

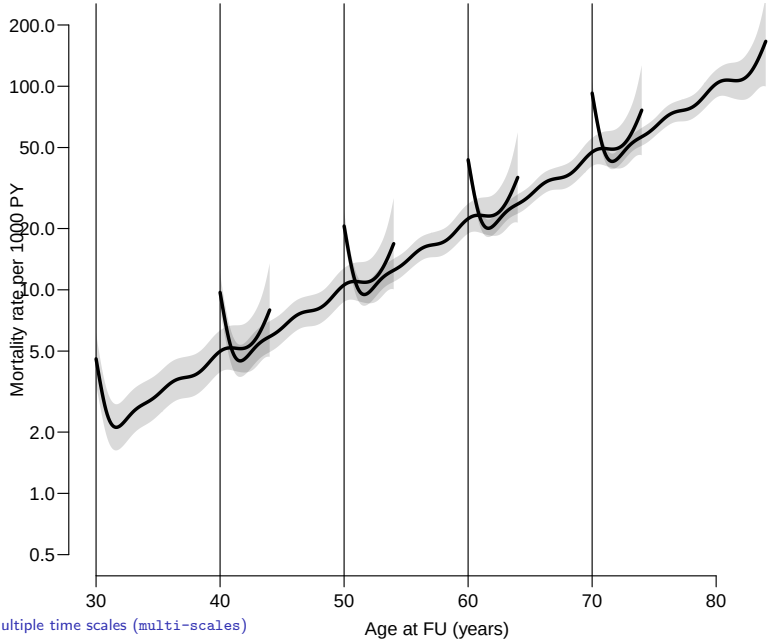
Predictions of mortality for these values of:
`age`: current age; `tdf`: duration and `ain`: age at DX.

Mortality rates, not effects

Predict mortality rates for Danish diabetes patients by age and duration of diabetes for persons diagnosed at ages 30, 40 etc.

```
matshade( nd$age, ci.pred( mad, nd )*1000, plot=TRUE,
          lwd=3, lty=1, log="y", las=1,
          xlim=c(30,85), ylim=c(1/2,200),
          xlab="Age at FU (years)",
          ylab="Mortality rate per 1000 PY" )
abline( v=3:7*10 )
```





Analysis by sex

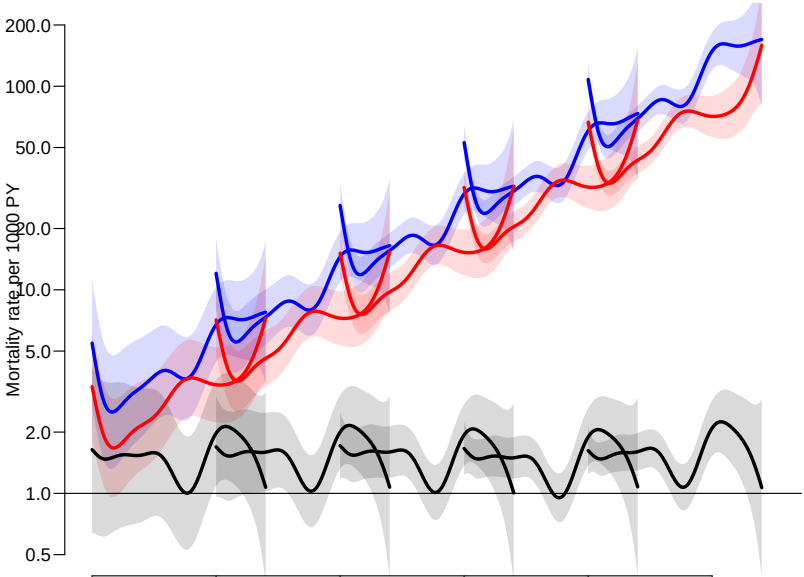
```
mm <- gam.Lexis( subset( Sdm, sex=="M" ), ~ s(age) + s(tfd) )
```

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

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

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

```
matshade( nd$age, cbind( ci.pred( mm, nd )*1000,
                         ci.pred( mw, nd )*1000,
                ci.ratio( ci.pred( mm, nd ),
                         ci.pred( mw, nd ) ) ), plot=TRUE,
          lwd=3, lty=1, log="y", las=1, col=c("blue","red","black"),
          xlim=c(30,85), ylim=c(1/2,200),
          xlab="Age at FU (years)",
          ylab="Mortality rate per 1000 PY" )
abline( h=1 )
```



. . . for you

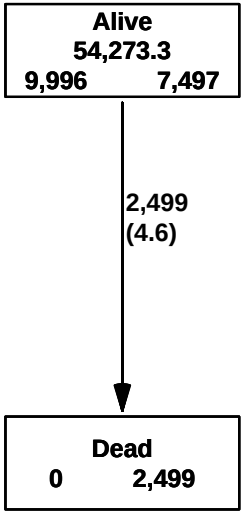
- What is is your conclusion for the effect of duration and age at diagnosis on the mortality rates?
- What is the effect of age **at** diagnosis?
- Your turn — do the analysis on your own computer.

Competing risks

Bendix Carstensen
Senior Statistician, Steno Diabetes Center Copenhagen

Survival
Multiple timescales
Competing risks
IDEG 2019 training day, Seoul,
29 November 2019

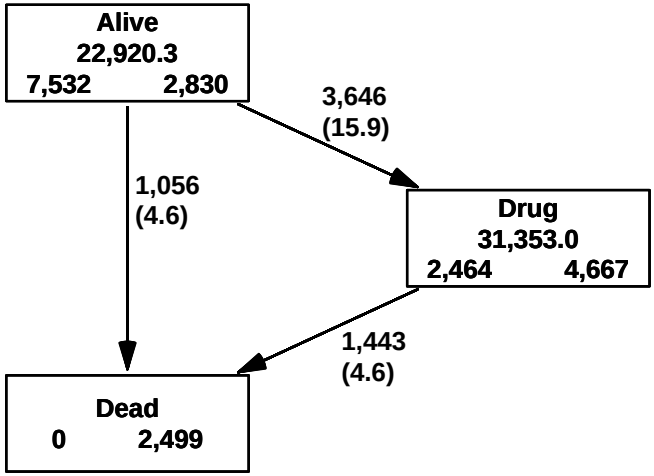
Survival analysis



One rate (the arrow)
One probability — $\mathrm{P}\{\text{alive at } t\}$

Some patients begin pharmaceutical treatment, they have follow-up
before `Drug` treatment and
after beginning `Drug` treatment

Three states, three transitions



Cut follow-up at beginning of drug therapy

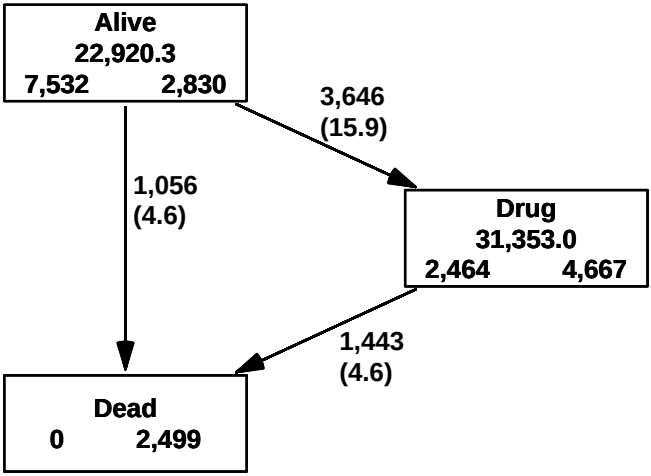
```
summary( Sdm )
```

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

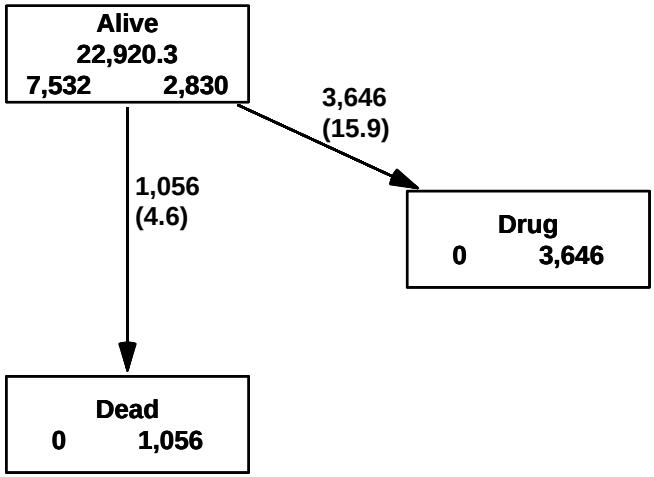
```
Sdm$dodr <- pmin(Sdm$dooad,Sdm$doins,na.rm=TRUE)
S3 <- cutLexis( data = Sdm,
                cut = Sdm$dodr,
          timescale = "per",
          new.state = "Drug",
    precursor.states = "Alive" )
summary( S3 )
```

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

Three states, three transitions



Three states, two (competing) transitions



Competing risk analysis

`lex.Xst` is factor with three levels:

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

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

. . . use it as response (event) variable in `Surv`:

```
m3 <- survfit( Surv( tfd, tfd+lex.dur, lex.Xst ) ~ 1,
               data = subset(S3,lex.Cst=="Alive"), id=lex.id )
```

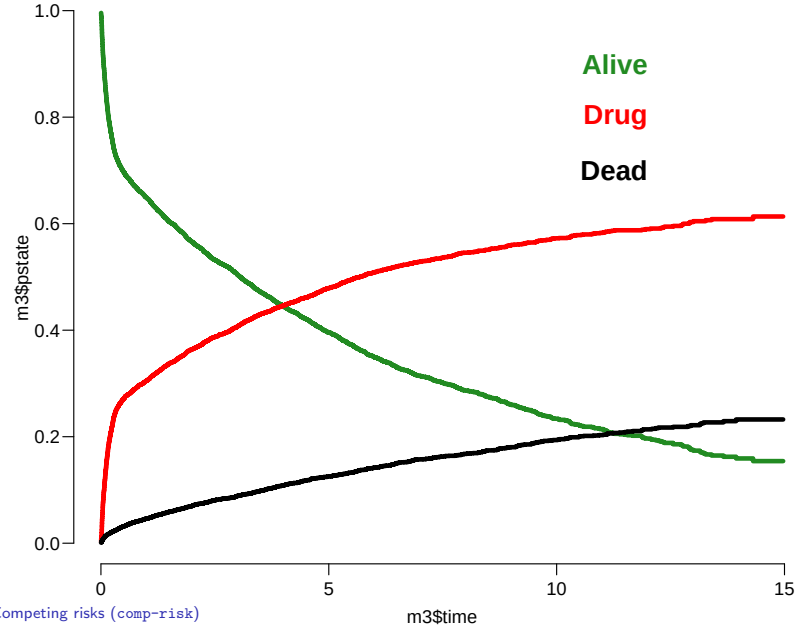
Computes the Aalen-Johansen estimator of state-probabilities — probability of being in each of the states assumed by `lex.Xst`

Competing risk analysis

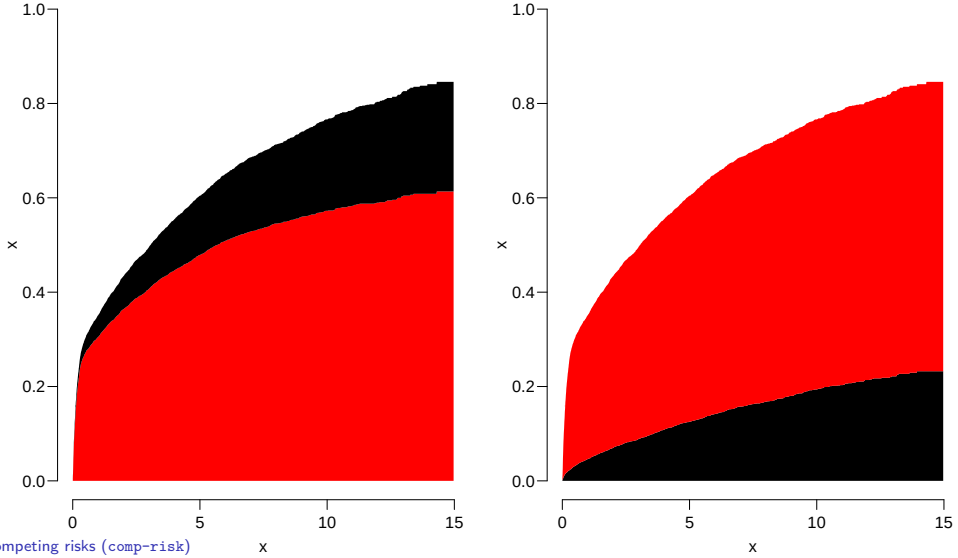
```
m3 <- survfit( Surv( tfd, tfd+lex.dur, lex.Xst ) ~ 1,
               data = subset(S3,lex.Cst=="Alive"), id=lex.id )
head( cbind(time=m3$time,m3$pstate), 7 )
```

```
          time
[1,] 0.002737851 0.9956187 0.003319172 0.001062135
[2,] 0.005475702 0.9901745 0.008232201 0.001593273
[3,] 0.008213552 0.9875188 0.010356754 0.002124411
[4,] 0.010951403 0.9847304 0.012614091 0.002655550
[5,] 0.013689254 0.9784895 0.018589397 0.002921119
[6,] 0.016427105 0.9727797 0.024033564 0.003186688
[7,] 0.019164955 0.9652100 0.031470515 0.003319491
```

```
matplot( m3$time, m3$pstate,
         type="s", lty=1, lwd=4,
         col=c("forestgreen","red","black") )
text( 12, 9:7/10, levels(S3$lex.Xst), adj=1, font=2, cex=1.5,
      col=c("forestgreen","red","black") )
```



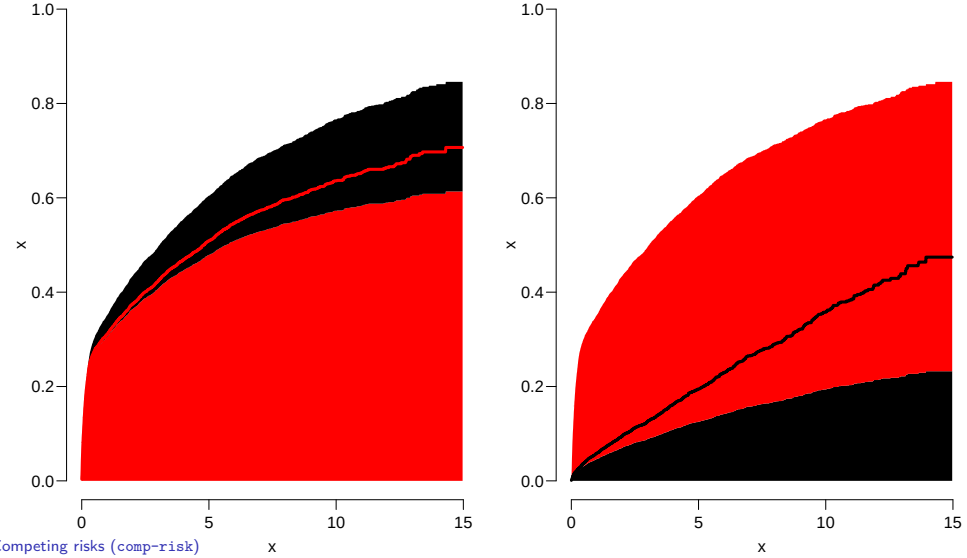
The stacked probabilities



Getting it wrong

- It is commonly seen that a traditional survival analyses are conducted where transition to `Drug` is taken as event and deaths just counted as censorings.
- This is wrong; it will overestimate the probability of going on drugs.
- But nothing wrong with the estimate of the **rate** of initiating drugs.
- Only the calculation of the cumulative **probability** is wrong — the probability of having initiated a drug depends on `both` the rate of drug initiation **and** the mortality rate.

The stacked probabilities + the wrong ones



What are the wrong probabilities?

Probability of `Drug` under the **assumptions:**

- `Dead` does not occur
- `Drug` occurs at the same rate as when `Dead` was a possibility
- hypothetical scenario about which there is no information in data
- . . . and about which no data can be collected

Getting the maths right

- rate of drug initiation (`Alive`→`Drug`): $\lambda(t)$
  mortality before drug initiation (`Alive`→`Dead`): $\mu(t)$
- $\Rightarrow$ probability of being alive without drug treatment at time t is:

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

- cumulative risk of `Drug` before time t is:

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

—and similarly for cumulative risk of `Dead`

Where is the error

- Error **only** in the calculations of the cumulative risk — the probability of transition to `Drug`.
- The "wrong" red line in the figure comes from omitting the green term $\mu(s)$ (the mortality rate) from the formula
- The temptations:
  - the mathematics becomes nicer if you compute the wrong thing
  - it is what comes out of standard programs when regarding `Drug` as the only type of event...
  - the hazard **ratios** are correct.
  - ...the program does not know there is a competing event if you don't tell
  - so the cumulative risks are wrong

Competing risks — practicalities

- Cause-specific **rates** can be modeled separately: cause-specific rates and HRs are perfectly valid
- Regression models for cause-specific rates translates to predicted probabilities for given covariates
- Fine-Gray models
  - the subdistribution hazard for cause c: $\frac{\partial}{\partial t}\log(1 - F_c(t))$
  - not a hazard, it's a mathematical transformation of the cumulative risk.
  - will not give probabilities that sum to 1 across causes

. . . not recommended

Competing risks summary

- No such thing as a competing risks analysis of event **rates**
- the competing risks aspect comes about only when you want to address **cumulative risk** of a particular event —in which case you probably want to look at cumulative risks of **all** types of events.