Survival, mortality, competing risks and expected lifetime

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

EDEG 2025 / Umeå University, 17 May 2025

http://bendixcarstensen.com/AdvCoh/courses/Um-2025/

From C:\Bendix\teach\AdvCoh\courses\Um-2025\slides/slides.tex Friday 16th May, 2025, 07:42

Survival and rate data

Rates and Survival

Survival, mortality,
competing risks and
expected lifetime
EDEG 2025 / Umeå University,17 May 2025

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

all of these have a starting point ("since")

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

all of these have a starting point ("since")

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

all of these have a starting point ("since")

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

all of these have a starting point ("since")

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

all of these have a starting point ("since")

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

all of these have a starting point (“since”)

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

all of these have a starting point ("since")

Examples of time-to-event measurements

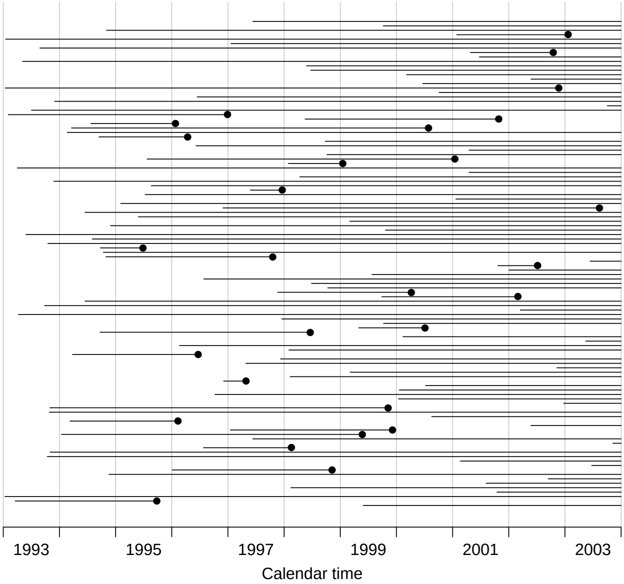
- Time from diagnosis of cancer to death.
- Time from randomisation to death in a cancer clinical trial
- Time from HIV infection to AIDS.
- Time from marriage to 1st child birth.
- Time from marriage to divorce.
- Time to re-offending after being released from jail

all of these have a starting point ("since")

Each line a person

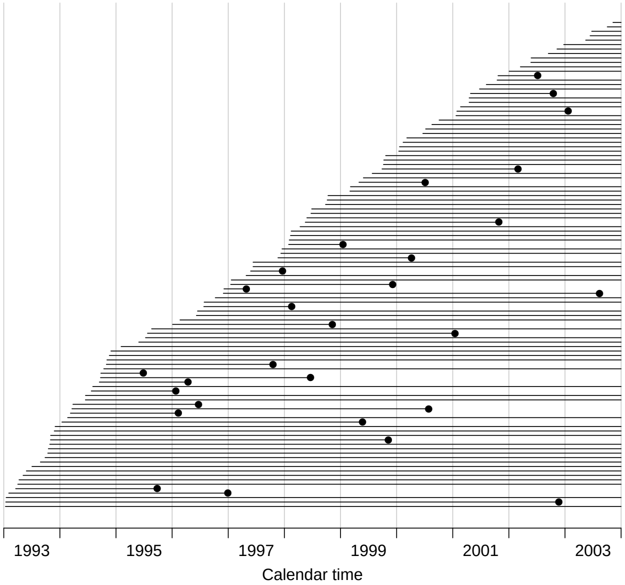
Each blob a death

Study ended at 31 Dec. 2003

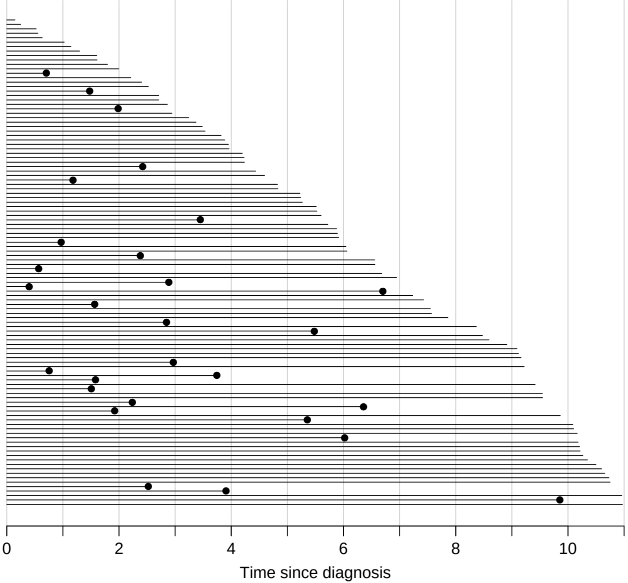


Ordered by date of entry

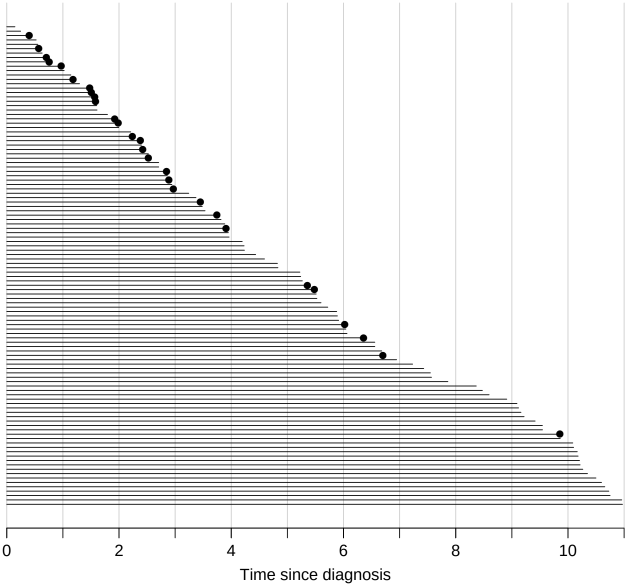
Most likely the order in your database.



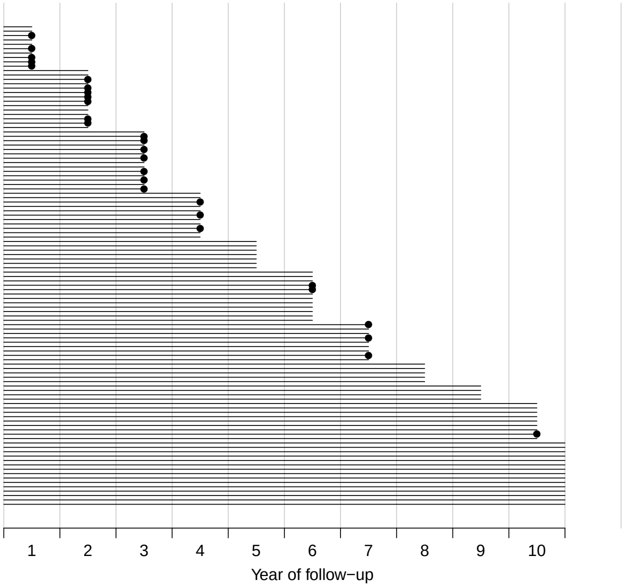
Timescale changed to "Time since diagnosis".



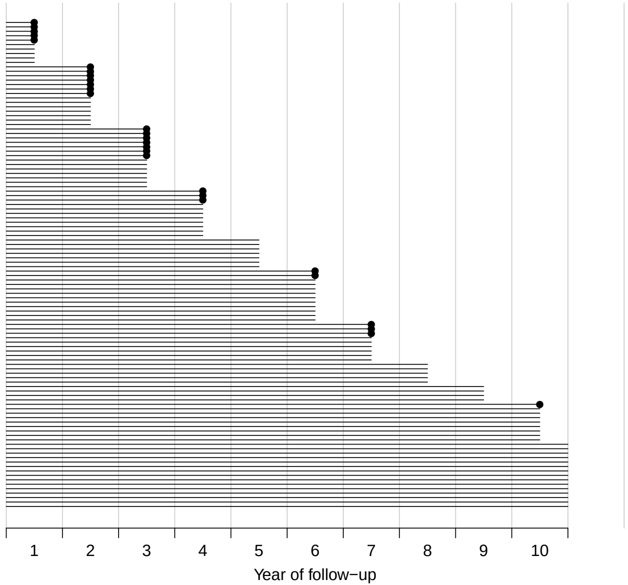
Patients ordered by survival time.



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

Life-table estimator of death probability: $D/(N - L/2)$

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

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

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

Estimated risk in year 1 for Stage I women is $5/107.5 = 0.0465$
Estimated risk in year 2 for Stage I women is $7/96.5 = 0.0725$
Estimated risk in year 3 for Stage I women is $7/82.5 = 0.0848$

Estimated 1 year survival is $1 - 0.0465 = 0.9535$
Estimated 2 year survival is $0.9535 \times (1 - 0.0725) = 0.8843$
Estimated 3 year survival is $0.8843 \times (1 - 0.0848) = 0.8093$

This is the **life-table estimator** of the survival curve.

- no need to use 1 year intervals: 1 day intervals could be used
- very small intervals will leave at most 1 censoring or 1 death in each
- interval with 1 death and n_i persons at risk
  $\mathrm{P}\{\text{Death}\} = 1/n_i$
- corresponding survival probability $1 - 1/n_i = (n_i - 1)/n_i$
- interval with 0 deaths has survival probability 1
- multiply these over times with event to get survival function

$$S(t) = \prod_{\tau < t \text{ with event}} (n_\tau - 1)/n_\tau$$

… you have the Kaplan-Meier estimator

- no need to use 1 year intervals: 1 day intervals could be used
- very small intervals will leave at most 1 censoring or 1 death in each
- interval with 1 death and n_i persons at risk: $P\{\text{Death}\} = 1/n_i$
- corresponding survival probability $1 - 1/n_i = (n_i - 1)/n_i$
- interval with 0 deaths has survival probability 1
- multiply these over times with event to get survival function

$$S(t) = \prod_{\tau < t \text{ with event}} (n_\tau - 1)/n_\tau$$

... you have the Kaplan-Meier estimator

- no need to use 1 year intervals: 1 day intervals could be used
- very small intervals will leave at most 1 censoring or 1 death in each
- interval with 1 death and n_t persons at risk: $\mathrm{P}\{\text{Death}\} = 1/n_t$
- corresponding survival probability $1 - 1/n_t = (n_t - 1)/n_t$
- interval with 0 deaths has survival probability 1
- multiply these over times with event to get survival function

$$S(t) = \prod_{\tau < t \text{ with event}} (n_\tau - 1)/n_\tau$$

you have the Kaplan-Meier estimator

- no need to use 1 year intervals: 1 day intervals could be used
- very small intervals will leave at most 1 censoring or 1 death in each
- interval with 1 death and n_t persons at risk:
  $\mathrm{P}\left\{\text{Death}\right\} = 1/n_t$
- corresponding survival probability $1 - 1/n_t = (n_t - 1)/n_t$
- interval with 0 deaths has survival probability 1
- multiply these over times with event to get survival function

$$S(t) = \prod_{\tau < t \text{ with event}} (n_\tau - 1)/n_\tau$$

you have the Kaplan-Meier estimator

- no need to use 1 year intervals: 1 day intervals could be used
- very small intervals will leave at most 1 censoring or 1 death in each
- interval with 1 death and n_t persons at risk: $\mathrm{P}\left\{\text{Death}\right\} = 1/n_t$
- corresponding survival probability $1 - 1/n_t = (n_t - 1)/n_t$
- interval with 0 deaths has survival probability 1
- multiply these over times with event to get survival function

$$S(t) = \prod_{\tau < t \text{ with event}} (n_\tau - 1)/n_\tau$$

... you have the Kaplan-Meier estimator

- no need to use 1 year intervals: 1 day intervals could be used
- very small intervals will leave at most 1 censoring or 1 death in each
- interval with 1 death and n_t persons at risk: $\mathrm{P}\left\{\text{Death}\right\} = 1/n_t$
- corresponding survival probability $1 - 1/n_t = (n_t - 1)/n_t$
- interval with 0 deaths has survival probability 1
- multiply these over times with event to get survival function:

$$S(t) = \prod_{\tau < t \text{ with event}} (n_\tau - 1)/n_\tau$$

. . . you have the **Kaplan-Meier estimator**

Survival after diabetes

computations

Survival, mortality,
competing risks and
expected lifetime
EDEG 2025 / Umeå University,17 May 2025

The DMlate data set

Get data, define age as age at dodm, omit if dox=dodm

```
> data(DMlate)
> DM <- mutate(DMlate, age = dodm - dobth)
> DM <- subset(DM, dox > dodm)
> head(DM)
```

```
       sex    dobth     dodm    dodth    dooad doins      dox      age
50185    F 1940.256 1998.917       NA       NA    NA 2009.997 58.66119
307563   M 1939.218 2003.309       NA 2007.446    NA 2009.997 64.09035
294104   F 1918.301 2004.552       NA       NA    NA 2009.997 86.25051
336439   F 1965.225 2009.261       NA       NA    NA 2009.997 44.03559
245651   M 1932.877 2008.653       NA       NA    NA 2009.997 75.77550
216824   F 1927.870 2007.886 2009.923       NA    NA 2009.923 80.01643
```

```
> str(DM)
```

```
'data.frame':        9996 obs. of  8 variables:
 $ sex  : Factor w/ 2 levels "M","F": 2 1 2 2 1 2 1 1 2 1 ...
 $ dobth: num  1940 1939 1918 1965 1933 ...
 $ dodm : num  1999 2003 2005 2009 2009 ...
 $ dodth: num  NA NA NA NA NA ...
 $ dooad: num  NA 2007 NA NA NA ...
```

Survival function: KM

Use survfit to construct the Kaplan-Meier estimator of overall survival:

```
> ?Surv
> ?survfit
```

```
> km <- survfit(Surv(dox - dodm, !is.na(dodth)) ~ 1, data = DM)
> km
```

```
Call: survfit(formula = Surv(dox - dodm, !is.na(dodth)) ~ 1, data = DM)

        n events median 0.95LCL 0.95UCL
[1,] 9996   2499   14.5    14.2      NA
```

```
> # summary(km) # very long output
```

We can plot the survival curve
—this is the default plot for a `survfit` object:

```
> plot(km)
```

What is the median survival? What does it mean?

Explore if survival patterns between men and women are different

```
> kms <- survfit(Surv(dox - dodm, !is.na(dodth)) ~ sex, data = DM)
> kms
Call: survfit(formula = Surv(dox - dodm, !is.na(dodth)) ~ sex, data = DM)

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

Exercises 1, 2

We can plot the survival curve
—this is the default plot for a `survfit` object:

```
> plot(km)
```

What is the median survival? What does it mean?
Explore if survival patterns between men and women are different:

```
> kms <- survfit(Surv(dox - dodm, !is.na(dodth)) ~ sex, data = DM)
> kms
```

```
Call: survfit(formula = Surv(dox - dodm, !is.na(dodth)) ~ sex, data = DM)

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

Exercises 1, 2

Men have worse survival than women, and women are a bit older at dodm:

```
> with(DM, tapply(dodm - dobth, sex, mean))
```

```
       M        F 
60.28980 62.45266 
```

Significant difference in survival between men and women

```
> survdiff(Surv(dox - dodm, !is.na(dodth)) ~ sex, data = DM)
```

```
Call:
survdiff(formula = Surv(dox - dodm, !is.na(dodth)) ~ sex, data = DM)

         N Observed Expected (O-E)^2/E (O-E)^2/V
sex=M 5183     1343     1271      4.08      8.31
sex=F 4813     1156     1228      4.22      8.31

 Chisq= 8.3  on 1 degrees of freedom, p= 0.004 
```

What is the null hypothesis tested here?

Rates and rate-ratios

- Occurrence **rate:**

$$\lambda(t) = \lim_{h \rightarrow 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

—measured in probability per time: time^{-1}

- observation in a survival study: (exit status, time alive)
- empirical rate (d, y) = (deaths, time)

Rates and rate-ratios

- Occurrence **rate:**

$$\lambda(t) = \lim_{h \rightarrow 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

  —measured in probability per time: time^{-1}

- observation in a survival study: (exit status, time alive)
- empirical rate $(d, y) = $ (deaths, time)

Rates and rate-ratios

- Occurrence **rate:**

$$\lambda(t) = \lim_{h \to 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

—measured in probability per time: time^{-1}

- observation in a survival study: (exit status, time alive)
- empirical rate $(d, y) = (\text{deaths}, \text{time})$

Rates and rate-ratios: Simple Cox model

Now explore how sex and age (at diagnosis) influence the mortality—note that in a Cox-model we are addressing the mortality rate and not the survival:

```
> c0 <- coxph(Surv(dox - dodm, !is.na(dodth)) ~ sex      , data = DM)
> c1 <- coxph(Surv(dox - dodm, !is.na(dodth)) ~ sex + age, data = DM)
> summary(c1)
> ci.exp(c0)
> ci.exp(c1)
```

What variables from DM are we using?

```
> c0 <- coxph(Surv(dox - dodm, !is.na(dodth)) ~ sex      , data = DM)
> c1 <- coxph(Surv(dox - dodm, !is.na(dodth)) ~ sex + age, data = DM)
> summary(c1)
```

```
Call:
coxph(formula = Surv(dox - dodm, !is.na(dodth)) ~ sex + age,
    data = DM)

  n= 9996, number of events= 2499

         coef exp(coef)  se(coef)      z Pr(>|z|)
sexF -0.386126  0.679685  0.040757 -9.474   <2e-16 ***
age   0.079884  1.083161  0.001833 43.569   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     exp(coef) exp(-coef) lower .95 upper .95
sexF    0.6797     1.4713    0.6275    0.7362
age     1.0832     0.9232    1.0793    1.0871

Concordance= 0.762  (se = 0.005 )
Likelihood ratio test= 2391  on 2 df,   p=<2e-16
Wald test            = 1902  on 2 df,   p=<2e-16
Score (logrank) test = 1875  on 2 df,   p=<2e-16
```

```
> ci.exp(c0)

     exp(Est.)      2.5%     97.5%
sexF 0.8908372 0.8234534 0.9637351

> ci.exp(c1)

     exp(Est.)      2.5%     97.5%
sexF 0.6796851 0.6275025 0.7362072
age  1.0831613 1.0792759 1.0870608
```

What do these estimates mean?

$$\lambda(t,x) = \lambda_0(t)\mathrm{exp}(\beta_1 x_1 + \beta_2 x_2)$$

Where is β_1 ? Where is β_2 ? Where is $\lambda_0(t)$?

What is the mortality RR for a 10 year age difference?

If mortality is assumed constant $(\lambda(t) = \lambda)$, then the likelihood for the Cox-model is equivalent to a Poisson likelihood, which can be fitted using the poisreg family from the Epi package:

```
> ?poisreg

> p1 <- glm(cbind(!is.na(dodth), dox - dodm) ~ sex + age,
+           family = poisreg,
+             data = DM)
> ci.exp(p1) # Poisson
               exp(Est.)        2.5%        97.5%
(Intercept) 0.0003520559 0.000274337 0.0004517924
sexF        0.6911295663 0.638139016 0.7485204093
age         1.0794724027 1.075733792 1.0832240061

> ci.exp(c1) # Cox
     exp(Est.)      2.5%     97.5%
sexF 0.6796851 0.6275025 0.7362072
age  1.0831613 1.0792759 1.0870608
```

Is the sex-effect confounded by age?

Sex and age effects are quite close for the Poisson and the Cox models.

Poisson model has an intercept term, the estimate of the (assumed) constant underlying mortality.

The risk time part of the response (second argument in the `cbind`) was entered in units of years, so the `(Intercept)` (taken from the `ci.exp`) is a rate per 1 person-month.

What age and sex does the `(Intercept)` refer to?

```
> ci.exp(p1) # Poisson

              exp(Est.)        2.5%        97.5%
(Intercept) 0.0003520559 0.000274337 0.0004517924
sexF        0.6911295663 0.638139016 0.7485204093
age         1.0794724027 1.075733792 1.0832240061
```

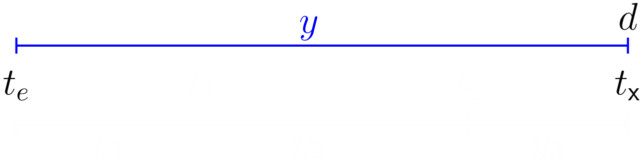
poisreg and poisson

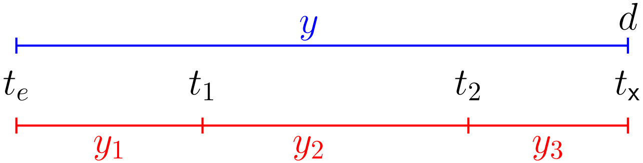
poisreg: cbind(d,y) ~ ...

```
> p1 <- glm(cbind(!is.na(dodth), dox - dodm) ~ sex + age,
+             family = poisreg,
+               data = DM)
```

poisson: d ~ ... + offset(log(y))

```
> px <- glm(!is.na(dodth) ~ sex + age + offset(log(dox - dodm)),
+             family = poisson,
+               data = lung)
> ## or:
> px <- glm(!is.na(dodth) ~ sex + age,
+             offset = log(dox - dodm),
+             family = poisson,
+               data = lung)
```


y
d
t_e
t_1
t_2
t_x
y_1
y_2
y_3


y
d
t_e
t_1
t_2
t_{x}
y_1
y_2
y_3

What is it that we see as outcome?

(d, y) or: $(0, y_1)$, $(0, y_2)$, (d, y_3)

the amount of information is the same — or is it?

What we observe is occurrence rates

Statistical model — hazard, intensity, occurrence rate, λ:

$$\lambda(t) = \lim_{h \rightarrow 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

— measured in probability per time: time^{-1}

What are the measurement scales for t and λ?

What is it that we see as outcome?

(d, y) or: $(0, y_1)$, $(0, y_2)$, (d, y_3)

the amount of information is the same — or is it?

What we observe is **occurrence rates**

Statistical model — hazard, intensity, occurrence rate, λ

$$\lambda(t) = \lim_{h \to 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

—measured in probability per time: time^{-1}

What are the measurement scales for t and h?

What is it that we see as outcome?

(d, y) or: $(0, y_1)$, $(0, y_2)$, (d, y_3)

the amount of information is the same — or is it?

What we observe is **occurrence rates**

Statistical model — hazard, intensity, occurrence rate, λ:

$$\lambda(t) = \lim_{h \rightarrow 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h$$

—measured in probability per time: time^{-1}

What are the measurement scales for t and h?

Likelihood

- Likelihood is the **probability** of data as a function of parameters, **assuming** the model is correct

$$L(\lambda) = \mathrm{P}(d \text{ at } t_{\mathsf{x}} | \text{entry } t_e \text{ \& correct model})$$

—this is a quantity that depends on λ (model parameters)

- Maximum likelihood estimation is choosing the value of λ that makes $L(\lambda)$ as large a possible
- Normally we maximize log-likelihood, $\ell(\lambda) = \log(L(\lambda))$, m.l.e. called $\hat{\lambda}$
- The second derivative of $\ell(\lambda)$ evaluated at $\hat{\lambda}$ contains information about the uncertainty of $\hat{\lambda}$

Likelihood

- Likelihood is the **probability** of data as a function of parameters, **assuming** the model is correct

$$L(\lambda) = \mathrm{P}(d \text{ at } t_{\mathsf{x}} | \text{entry } t_e \text{ \& correct model})$$

  —this is a quantity that depends on λ (model parameters)

- Maximum likelihood estimation is choosing the value of λ that makes $L(\lambda)$ as large a possible
- Normally we maximize log-likelihood, $\ell(\lambda) = \log\big(L(\lambda)\big)$, m.l.e. called $\hat{\lambda}$
- The second derivative of $\ell(\lambda)$ evaluated at $\hat{\lambda}$ contains information about the uncertainty of $\hat{\lambda}$

Likelihood

- Likelihood is the **probability** of data as a function of parameters, **assuming** the model is correct

$$L(\lambda) = \mathrm{P}(d \text{ at } t_{\mathsf{x}} | \text{entry } t_e \text{ \& correct model})$$

  —this is a quantity that depends on λ (model parameters)

- Maximum likelihood estimation is choosing the value of λ that makes $L(\lambda)$ as large a possible
- Normally we maximize log-likelihood, $\ell(\lambda) = \log\big(L(\lambda)\big)$, m.l.e. called $\hat{\lambda}$
- The second derivative of $\ell(\lambda)$ evaluated at $\hat{\lambda}$ contains information about the uncertainty of $\hat{\lambda}$

Likelihood

- Likelihood is the **probability** of data as a function of parameters, **assuming** the model is correct

$$L(\lambda) = \mathrm{P}(d \text{ at } t_{\mathsf{x}} | \text{entry } t_e \text{ \& correct model})$$

  —this is a quantity that depends on λ (model parameters)
- Maximum likelihood estimation is choosing the value of λ that makes $L(\lambda)$ as large a possible
- Normally we maximize log-likelihood, $\ell(\lambda) = \log\big(L(\lambda)\big)$, m.l.e. called $\hat{\lambda}$
- The second derivative of $\ell(\lambda)$ evaluated at $\hat{\lambda}$ contains information about the uncertainty of $\hat{\lambda}$

∗ Likelihood and records

- Suppose a person is alive from t_e (entry) to t_x (exit) and
- that the person's status at t_x is d
  where $d = 0$ means alive and $d = 1$ means dead.
- If we choose, say, two time points t_1, t_2 between t_e and t_x,
- standard use of conditional probability
  (formally, repeated use of Bayes' formula) gives:

$$
\begin{aligned}
\mathrm{P}\{d \text{ at } t_x \mid \text{entry at } t_e\} = & \ \mathrm{P}\{\text{survive } (t_e, t_1] \mid \text{alive at } t_e\} \times \\
& \ \mathrm{P}\{\text{survive } (t_1, t_2] \mid \text{alive at } t_1\} \times \\
& \ \mathrm{P}\{\text{survive } (t_2, t_x) \mid \text{alive at } t_2\} \times \\
& \ \mathrm{P}\{d \text{ at } t_x \mid \text{alive just before } t_x\}
\end{aligned}
$$

∗ Likelihood and records

- Suppose a person is alive from t_e (entry) to t_x (exit) and
- that the person's status at t_x is d,
  where $d = 0$ means alive and $d = 1$ means dead.
- If we choose, say, two time points t_1, t_2 between t_e and t_x,
- standard use of conditional probability
  (formally, repeated use of Bayes' formula) gives

$$
\begin{aligned}
\mathrm{P}\{d \text{ at } t_x \mid \text{entry at } t_e\} = & \, \mathrm{P}\{\text{survive } (t_e, t_1] \mid \text{alive at } t_e\} \times \\
& \, \mathrm{P}\{\text{survive } (t_1, t_2] \mid \text{alive at } t_1\} \times \\
& \, \mathrm{P}\{\text{survive } (t_2, t_x) \mid \text{alive at } t_2\} \times \\
& \, \mathrm{P}\{d \text{ at } t_x \mid \text{alive just before } t_x\}
\end{aligned}
$$

∗ Likelihood and records

- Suppose a person is alive from t_e (entry) to t_x (exit) and
- that the person's status at t_x is d,
  where $d = 0$ means alive and $d = 1$ means dead.
- If we choose, say, two time points, t_1, t_2 between t_e and t_x,
- standard use of conditional probability
  (formally, repeated use of Bayes' formula) gives:

$$\begin{aligned} \mathrm{P}\left\{d \text{ at } t_x \mid \text{entry at } t_e\right\} = {} & \mathrm{P}\left\{\text{survive } (t_e, t_1] \mid \text{alive at } t_e\right\} \times \\ & \mathrm{P}\left\{\text{survive } (t_1, t_2] \mid \text{alive at } t_1\right\} \times \\ & \mathrm{P}\left\{\text{survive } (t_2, t_x] \mid \text{alive at } t_2\right\} \times \\ & \mathrm{P}\left\{d \text{ at } t_x \mid \text{alive just before } t_x\right\} \end{aligned}$$

* Likelihood and records

- Suppose a person is alive from t_e (entry) to t_x (exit) and
- that the person's status at t_x is d,
  where $d = 0$ means alive and $d = 1$ means dead.
- If we choose, say, two time points, t_1, t_2 between t_e and t_x,
- standard use of conditional probability
  (formally, repeated use of Bayes' formula) gives:

$$
\begin{aligned}
\mathrm{P}\{d \text{ at } t_x \mid \text{entry at } t_e\} = {} & \mathrm{P}\{\text{survive } (t_e, t_1] \mid \text{alive at } t_e\} \times \\
& \mathrm{P}\{\text{survive } (t_1, t_2] \mid \text{alive at } t_1\} \times \\
& \mathrm{P}\{\text{survive } (t_2, t_x) \mid \text{alive at } t_2\} \times \\
& \mathrm{P}\{d \text{ at } t_x \mid \text{alive just before } t_x\}
\end{aligned}
$$

* Likelihood and records

- Suppose a person is alive from t_e (entry) to t_x (exit) and
- that the person's status at t_x is d,
  where $d = 0$ means alive and $d = 1$ means dead.
- If we choose, say, two time points, t_1, t_2 between t_e and t_x,
- standard use of conditional probability
  (formally, repeated use of Bayes' formula) gives:

$$\begin{aligned} P\{d \text{ at } t_x \mid \text{entry at } t_e\} = {} & P\{\text{survive } (t_e, t_1] \mid \text{alive at } t_e\} \times \\ & P\{\text{survive } (t_1, t_2] \mid \text{alive at } t_1\} \times \\ & P\{\text{survive } (t_2, t_x] \mid \text{alive at } t_2\} \times \\ & P\{d \text{ at } t_x \mid \text{alive just before } t_x\} \end{aligned}$$

∗ Likelihood and records

- Suppose a person is alive from t_e (entry) to t_x (exit) and
- that the person's status at t_x is d,
  where $d = 0$ means alive and $d = 1$ means dead.
- If we choose, say, two time points, t_1, t_2 between t_e and t_x,
- standard use of conditional probability
  (formally, repeated use of Bayes' formula) gives:

$$\begin{aligned}
\mathrm{P}\left\{d \text{ at } t_x \,|\, \text{entry at } t_e\right\} = {} & \mathrm{P}\left\{\text{survive } (t_e, t_1] \,|\, \text{alive at } t_e\right\} \times \\
& \mathrm{P}\left\{\text{survive } (t_1, t_2] \,|\, \text{alive at } t_1\right\} \times \\
& \mathrm{P}\left\{\text{survive } (t_2, t_x] \,|\, \text{alive at } t_2\right\} \times \\
& \mathrm{P}\left\{d \text{ at } t_x \,|\, \text{alive just before } t_x\right\}
\end{aligned}$$

* Rates and likelihood

For a start assume that the mortality is constant over time $\lambda(t) = \lambda$:

$$\begin{aligned}\mathrm{P}\left\{\text{death during } (t, t+h] | \text{alive at } t\right\} &\approx \lambda h \\ \Rightarrow \mathrm{P}\left\{\text{survive } (t, t+h] | \text{alive at } t\right\} &\approx 1 - \lambda h\end{aligned} \tag{1}$$

where the approximation gets better the smaller h is.

∗ Dividing follow-up time

- Survival for a time span: $y = t_x - t_e$
- Subdivided in N intervals, each of length $h = y/N$
- The rate is assumed constant: $\lambda(t) = \lambda$
- Survival probability for the entire span from t_e to t_x is the product of probabilities of surviving each of the small intervals conditional on being alive at the beginning each interval

$$\mathrm{P}\left\{\text{survive } t_e \text{ to } t_x\right\} \approx (1 - \lambda h)^N = \left(1 - \frac{\lambda y}{N}\right)^N$$

* Dividing follow-up time

- Survival for a time span: $y = t_x - t_e$
- Subdivided in N intervals, each of length $h = y/N$
- The rate is assumed constant: $\lambda(t) = \lambda$
- Survival probability for the entire span from t_e to t_x is the product of probabilities of surviving each of the small intervals conditional on being alive at the beginning each interval

$$\mathrm{P}\left\{\text{survive } t_e \text{ to } t_x\right\} \approx (1 - \lambda h)^N = \left(1 - \frac{\lambda y}{N}\right)^N$$

∗ Dividing follow-up time

- Survival for a time span: $y = t_x - t_e$
- Subdivided in N intervals, each of length $h = y/N$
- The rate is assumed constant: $\lambda(t) = \lambda$
- Survival probability for the entire span from t_e to t_x is the product of probabilities of surviving each of the small intervals, conditional on being alive at the beginning each interval:

$$\mathrm{P}\left\{\text{survive } t_e \text{ to } t_x\right\} \approx (1 - \lambda h)^N = \left(1 - \frac{\lambda y}{N}\right)^N$$

∗ Dividing follow-up time

- Survival for a time span: $y = t_x - t_e$
- Subdivided in N intervals, each of length $h = y/N$
- The rate is assumed constant: $\lambda(t) = \lambda$
- Survival probability for the entire span from t_e to t_x is the **product** of probabilities of surviving each of the small intervals, conditional on being alive at the beginning each interval:

$$\mathrm{P}\left\{\text{survive } t_e \text{ to } t_x\right\} \approx (1 - \lambda h)^N = \left(1 - \frac{\lambda y}{N}\right)^N$$

∗ Dividing follow-up time in small pieces

- From mathematics it is known that $(1+x/n)^n \to \exp(x)$ as $n \to \infty$ (some define $\exp(x)$ this way).
- So if we divide the time span y in small pieces we will have that as $N \to \infty$

$$\mathrm{P}\{\text{survive } t_0 \text{ to } t_y\} \approx \left(1 - \frac{\lambda y}{N}\right)^N \to \exp(-\lambda y) \qquad (2)$$

- The contribution to the likelihood from a person observed for a time span of length y is $\exp(-\lambda y)$, and the contribution to the log-likelihood is therefore $-\lambda y$

∗ Dividing follow-up time in small pieces

- From mathematics it is known that $(1 + x/n)^n \to \exp(x)$ as $n \to \infty$ (some define $\exp(x)$ this way).
- So if we divide the time span y in small pieces we will have that as $N \to \infty$:

$$\mathrm{P}\left\{\text{survive } t_e \text{ to } t_x\right\} \approx \left(1 - \frac{\lambda y}{N}\right)^N \to \exp(-\lambda y) \qquad (2)$$

- The contribution to the likelihood from a person observed for a time span of length y is $\exp(-\lambda y)$, and the contribution to the log-likelihood is therefore $-\lambda y$

∗ Dividing follow-up time in small pieces

- From mathematics it is known that $(1 + x/n)^n \rightarrow \exp(x)$ as $n \rightarrow \infty$ (some define $\exp(x)$ this way).
- So if we divide the time span y in small pieces we will have that as $N \rightarrow \infty$:

$$\mathrm{P}\left\{\text{survive } t_e \text{ to } t_x\right\} \approx \left(1 - \frac{\lambda y}{N}\right)^N \rightarrow \exp(-\lambda y) \qquad (2)$$

- The contribution to the likelihood from a person observed for a time span of length y is $\exp(-\lambda y)$, and the contribution to the log-likelihood is therefore $-\lambda y$.

* Dividing follow-up time: death at the end

- A person dying at the end of the last interval, the contribution to the likelihood from the last interval will be
- the probability surviving till just before the end of the interval,
- multiplied by
- the probability of dying in the last tiny instant (of length ϵ) of the interval
- The probability of dying in this tiny instant is $\lambda\epsilon$
- log-likelihood contribution from this last instant is $\log(\lambda\epsilon) = \log(\lambda) + \log(\epsilon)$

* Dividing follow-up time: death at the end

- A person dying at the end of the last interval, the contribution to the likelihood from the last interval will be
- the probability surviving till just before the end of the interval,
- multiplied by
- the probability of dying in the last tiny instant (of length ϵ) of the interval
- The probability of dying in this tiny instant is $\lambda\epsilon$
- log-likelihood contribution from this last instant is $\log(\lambda\epsilon) = \log(\lambda) + \log(\epsilon)$

∗ Dividing follow-up time: death at the end

- A person dying at the end of the last interval, the contribution to the likelihood from the last interval will be
- the probability surviving till just before the end of the interval,
- **multiplied** by
- the probability of dying in the last tiny instant (of length ϵ) of the interval
- The probability of dying in this tiny instant is $\lambda\epsilon$
- log-likelihood contribution from this last instant is $\log(\lambda\epsilon) = \log(\lambda) + \log(\epsilon)$.

* Dividing follow-up time: death at the end

- A person dying at the end of the last interval, the contribution to the likelihood from the last interval will be
- the probability surviving till just before the end of the interval,
- **multiplied** by
- the probability of dying in the last tiny instant (of length ϵ) of the interval
- The probability of dying in this tiny instant is $\lambda\epsilon$
- log-likelihood contribution from this last instant is $\log(\lambda\epsilon) = \log(\lambda) + \log(\epsilon)$

∗ Dividing follow-up time: death at the end

- A person dying at the end of the last interval, the contribution to the likelihood from the last interval will be
- the probability surviving till just before the end of the interval,
- **multiplied** by
- the probability of dying in the last tiny instant (of length ϵ) of the interval
- The probability of dying in this tiny instant is $\lambda\epsilon$
- log-likelihood contribution from this last instant is $\log(\lambda\epsilon) = \log(\lambda) + \log(\epsilon)$.

$*$ Dividing follow-up time: death at the end

- A person dying at the end of the last interval, the contribution to the likelihood from the last interval will be
- the probability surviving till just before the end of the interval,
- **multiplied** by
- the probability of dying in the last tiny instant (of length ϵ) of the interval
- The probability of dying in this tiny instant is $\lambda\epsilon$
- log-likelihood contribution from this last instant is $\log(\lambda\epsilon) = \log(\lambda) + \log(\epsilon)$.

∗ Total likelihood

The total likelihood for one person is the product of all these terms from the follow-up intervals (i) for the person; and the log-likelihood (ℓ) is therefore the sum of the log-likelihood terms:

$$\begin{aligned}\ell(\lambda) &= \sum_i (-\lambda y_i + d_i \log(\lambda) + d_i \log(\epsilon)) \\ &= \sum_i \left(d_i \log(\lambda) - \lambda y_i\right) + \sum_i d_i \log(\epsilon)\end{aligned}$$

The last term does not depend on λ, so it can be ignored.

∗ Total likelihood

The total likelihood for one person is the product of all these terms from the follow-up intervals (i) for the person; and the log-likelihood (ℓ) is therefore the sum of the log-likelihood terms:

$$\begin{aligned} \ell(\lambda) &= \sum_i (-\lambda y_i + d_i \log(\lambda) + d_i \log(\epsilon)) \\ &= \sum_i \left(d_i \log(\lambda) - \lambda y_i \right) + \sum_i d_i \log(\epsilon) \end{aligned}$$

The last term does not depend on λ, so it can be ignored

∗ Total log-likelihood

- …for the follow up of **one** person is (the **rate** likelihood):

$$\sum_i \left(d_i \log(\lambda) - \lambda y_i \right)$$

- this is also the likelihood for independent Poisson variates d_i with means λy_i
- even though the d_is are neither Poisson nor independent
- Different models can have the same (log)likelihood

∗ Total log-likelihood

- . . . for the follow up of **one** person is (the **rate** likelihood):

$$\sum_i \left(d_i \log(\lambda) - \lambda y_i \right)$$

- this is also the likelihood for independent Poisson variates d_i with means λy_i.
- even though the d_is are neither Poisson nor independent
- Different models can have the same (log)likelihood

* Total log-likelihood

- ...for the follow up of **one** person is (the **rate** likelihood):

$$\sum_i \left(d_i \log(\lambda) - \lambda y_i\right)$$

- this is also the likelihood for independent Poisson variates d_i with means λy_i.
- even though the d_is are neither Poisson nor independent
- Different models can have the same (log)likelihood

∗ Total log-likelihood

- . . . for the follow up of **one** person is (the **rate** likelihood):

$$\sum_i \left(d_i \log(\lambda) - \lambda y_i\right)$$

- this is also the likelihood for independent Poisson variates d_i with means λy_i.
- even though the d_is are neither Poisson nor independent
- Different models can have the same (log)likelihood:
    - model for follow-up of a person (d_i, y_i), constant rate λ
    - model for independent Poisson variates (d_i), mean λy_i

∗ Total log-likelihood

- …for the follow up of **one** person is (the **rate** likelihood):

$$\sum_i \left(d_i \log(\lambda) - \lambda y_i\right)$$

- this is also the likelihood for independent Poisson variates d_i with means λy_i.
- even though the d_is are neither Poisson nor independent
- Different models can have the same (log)likelihood:
  - model for follow-up of a person (d_i, y_i), constant rate λ
  - model for independent Poisson variates (d_i), means λy_i

* Total log-likelihood

- ...for the follow up of **one** person is (the **rate** likelihood):

$$\sum_i \left(d_i \log(\lambda) - \lambda y_i\right)$$

- this is also the likelihood for independent Poisson variates d_i with means λy_i.
- even though the d_is are neither Poisson nor independent
- Different models can have the same (log)likelihood:
  - model for follow-up of a person (d_i, y_i), constant rate λ
  - model for independent Poisson variates (d_i), mean λy_i

What did we do?

- Divide follow-up time in small pieces for the sake of mathematical approximations
- . . . leading to an expression of the log-likelihood contribution from a single person's follow-up
- . . . as a sum of many small contributions with small FU
- . . . explains why the rate likelihood is the same as a Poisson likelihood (although the model is not a Poisson model)
- Unrelated to this, next we will subdivide follow-up for the sake of modeling the rate λ as a function of covariates that varies over time, within each person

What did we do?

- Divide follow-up time in small pieces for the sake of mathematical approximations
- . . . leading to an expression of the log-likelihood contribution from a single person's follow-up
- . . . as a sum of many small contributions with small FU
- . . . explains why the rate likelihood is the same as a Poisson likelihood (although the model is not a Poisson model)
- Unrelated to this, next we will subdivide follow-up for the sake of modeling the rate λ as a function of covariates that varies over time, within each person

What did we do?

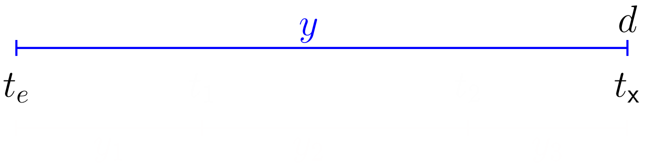
- Divide follow-up time in small pieces for the sake of mathematical approximations
- . . . leading to an expression of the log-likelihood contribution from a single person's follow-up
- . . . as a sum of many small contributions with small FU
- . . . explains why the rate likelihood is the same as a Poisson likelihood (although the model is not a Poisson model)
- Unrelated to this, next we will subdivide follow-up for the sake of modeling the rate λ as a function of covariates that varies over time, within each person

What did we do?

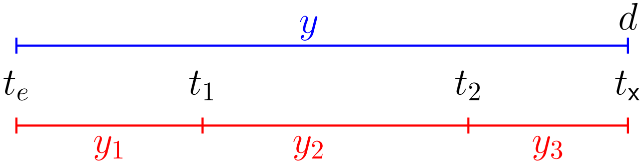
- Divide follow-up time in small pieces for the sake of mathematical approximations
- . . . leading to an expression of the log-likelihood contribution from a single person's follow-up
- . . . as a sum of many small contributions with small FU
- . . . explains why the rate likelihood is the same as a Poisson likelihood (although the model is not a Poisson model)
- Unrelated to this, next we will subdivide follow-up for the sake of modeling the rate λ as a function of covariates that varies over time, within each person

What did we do?

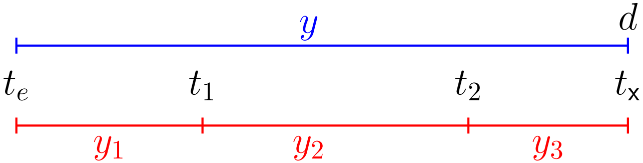
- Divide follow-up time in small pieces for the sake of mathematical approximations
- . . . leading to an expression of the log-likelihood contribution from a single person's follow-up
- . . . as a sum of many small contributions with small FU
- . . . explains why the rate likelihood is the same as a Poisson likelihood (although the model is not a Poisson model)
- **Unrelated** to this, next we will subdivide follow-up for the sake of **modeling** the rate λ as a function of covariates that varies over time, **within** each person



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_x \mid \text{entry } t_e)$	$d \log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_e \to t_1 \mid \text{entry } t_e)$	$= 0 \log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+\, 0 \log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(d \text{ at } t_x \mid \text{entry } t_2)$	$+\, d \log(\lambda) - \lambda y_3$



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_\mathsf{x} \mid \text{entry } t_e)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_e \to t_1 \mid \text{entry } t_e)$	$0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(d \text{ at } t_\mathsf{x} \mid \text{entry } t_2)$	$+d\log(\lambda) - \lambda y_3$



Probability

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

$= P(\text{surv } t_e \to t_1 | \text{entry } t_e)$

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

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

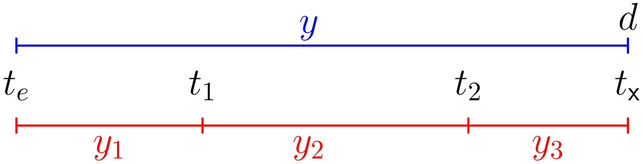
log-Likelihood

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

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

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

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



Probability

$\mathrm{P}(d \text{ at } t_\mathsf{x} | \text{entry } t_e)$

$= \mathrm{P}(\text{surv } t_e \to t_1 | \text{entry } t_e)$

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

$\times \mathrm{P}(d \text{ at } t_\mathsf{x} | \text{entry } t_2)$

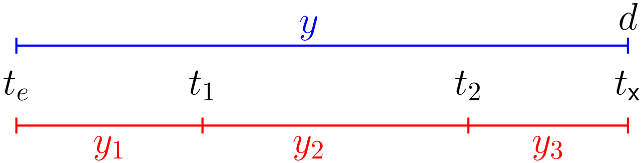
log-Likelihood

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

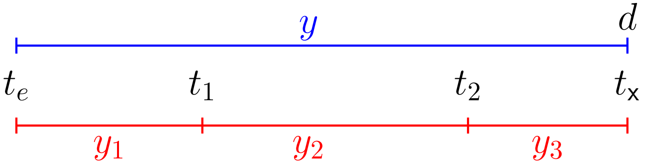
$= 0 \log(\lambda) - \lambda y_1$

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

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



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_{\mathsf{x}} \mid \text{entry } t_e)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_e \to t_1 \mid \text{entry } t_e)$	$= 0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \to t_2 \mid \text{entry } t_1)$	$+ 0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(d \text{ at } t_{\mathsf{x}} \mid \text{entry } t_2)$	$+ d\log(\lambda) - \lambda y_3$



Probability

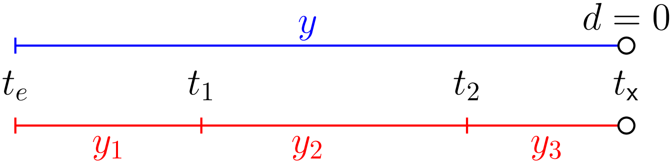
$$P(d \text{ at } t_{\mathsf{x}} | \text{entry } t_e)$$

$$= P(\text{surv } t_e \to t_1 | \text{entry } t_e)$$
$$\times P(\text{surv } t_1 \to t_2 | \text{entry } t_1)$$
$$\times P(d \text{ at } t_{\mathsf{x}} | \text{entry } t_2)$$

log-Likelihood

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

$$= 0 \log(\lambda) - \lambda y_1$$
$$+ 0 \log(\lambda) - \lambda y_2$$
$$+ d \log(\lambda) - \lambda y_3$$



Probability

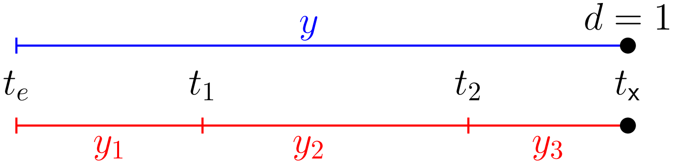
$$\mathrm{P}(\text{surv } t_e \rightarrow t_\mathsf{x} | \text{entry } t_e)$$

$$= \mathrm{P}(\text{surv } t_e \rightarrow t_1 | \text{entry } t_e)$$
$$\times \, \mathrm{P}(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1)$$
$$\times \, \mathrm{P}(\text{surv } t_2 \rightarrow t_\mathsf{x} | \text{entry } t_2)$$

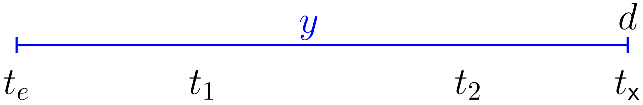
log-Likelihood

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

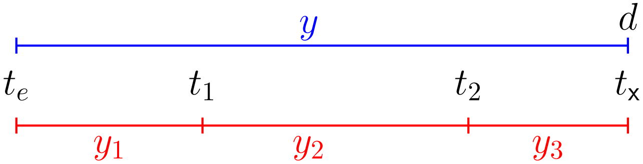
$$= 0 \log(\lambda) - \lambda y_1$$
$$+ \, 0 \log(\lambda) - \lambda y_2$$
$$+ \, 0 \log(\lambda) - \lambda y_3$$



Probability	log-Likelihood
$\mathrm{P}(\text{event at } t_x \mid \text{entry } t_e)$	$1 \log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_e \rightarrow t_1 \mid \text{entry } t_e)$	$= 0 \log(\lambda) - \lambda y_1$
$\times \mathrm{P}(\text{surv } t_1 \rightarrow t_2 \mid \text{entry } t_1)$	$+ 0 \log(\lambda) - \lambda y_2$
$\times \mathrm{P}(\text{event at } t_x \mid \text{entry } t_2)$	$+ 1 \log(\lambda) - \lambda y_3$



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_\mathsf{x} \mid \text{entry } t_e)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_e \rightarrow t_1 \mid \text{entry } t_e)$	$= 0\log(\lambda) - \lambda y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \rightarrow t_2 \mid \text{entry } t_1)$	$+\, 0\log(\lambda) - \lambda y_2$
$\times\, \mathrm{P}(d \text{ at } t_\mathsf{x} \mid \text{entry } t_2)$	$+\, d\log(\lambda) - \lambda y_3$



Probability

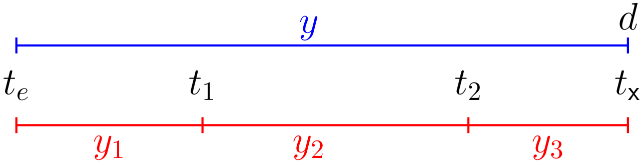
$$\mathrm{P}(d \text{ at } t_\mathsf{x} | \text{entry } t_e)$$

$$= \mathrm{P}(\text{surv } t_e \rightarrow t_1 | \text{entry } t_e)$$
$$\times\, \mathrm{P}(\text{surv } t_1 \rightarrow t_2 | \text{entry } t_1)$$
$$\times\, \mathrm{P}(d \text{ at } t_\mathsf{x} | \text{entry } t_2)$$

log-Likelihood

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

$$= 0 \log(\lambda) - \lambda y_1$$
$$+\, 0 \log(\lambda) - \lambda y_2$$
$$+\, d \log(\lambda) - \lambda y_3$$



Probability

$$P(d \text{ at } t_{\mathsf{x}} | \text{entry } t_e)$$

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

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

$$\times P(d \text{ at } t_{\mathsf{x}} | \text{entry } t_2)$$

log-Likelihood

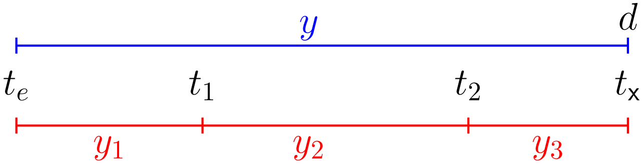
$$d \log(\lambda) - \lambda y$$

$$= 0 \log(\lambda_1) - \lambda_1 y_1$$

$$+ 0 \log(\lambda_2) - \lambda_2 y_2$$

$$+ d \log(\lambda_3) - \lambda_3 y_3$$

— allows different rates (λ_i) in each interval



Probability	log-Likelihood
$\mathrm{P}(d \text{ at } t_{\mathsf{x}} \mid \text{entry } t_e)$	$d\log(\lambda) - \lambda y$
$= \mathrm{P}(\text{surv } t_e \rightarrow t_1 \mid \text{entry } t_e)$	$= 0\log(\lambda_1) - \lambda_1 y_1$
$\times\, \mathrm{P}(\text{surv } t_1 \rightarrow t_2 \mid \text{entry } t_1)$	$+\, 0\log(\lambda_2) - \lambda_2 y_2$
$\times\, \mathrm{P}(d \text{ at } t_{\mathsf{x}} \mid \text{entry } t_2)$	$+\, d\log(\lambda_3) - \lambda_3 y_3$

— allows different rates (λ_i) in each interval

Maximum likelihood estimation of a rate

- One person (p) followed over many intervals contributes:

$$\ell_p(\lambda) = \sum_i \left(d_{pi}\log(\lambda) - \lambda y_{pi}\right)$$

- all persons followed over many intervals contributes:

$$\sum_p \ell_p(\lambda) = \sum_{p,i} \left(d_{pi}\log(\lambda) - \lambda y_{pi}\right) = D\log(\lambda) - \lambda Y$$

where D is total no. of deaths and Y is total risk time

- This is maximal for $\hat{\lambda} = D/Y$
- λ can depend on many parameters, so maximization is multidimensional

Maximum likelihood estimation of a rate

- One person (p) followed over many intervals contributes:

$$\ell_p(\lambda) = \sum_i \big(d_{pi} \log(\lambda) - \lambda y_{pi}\big)$$

- all persons followed over many intervals contributes:

$$\sum_p \ell_p(\lambda) = \sum_{p,i} \big(d_{pi} \log(\lambda) - \lambda y_{pi}\big) = D \log(\lambda) - \lambda Y$$

  where D is total no. of deaths and Y is total risk time

- This is maximal for $\hat{\lambda} = D/Y$
- λ can depend on many parameters, so maximization is multidimensional

Maximum likelihood estimation of a rate

- One person (p) followed over many intervals contributes:

$$\ell_p(\lambda) = \sum_i \big(d_{pi}\log(\lambda) - \lambda y_{pi}\big)$$

- all persons followed over many intervals contributes:

$$\sum_p \ell_p(\lambda) = \sum_{p,i} \big(d_{pi}\log(\lambda) - \lambda y_{pi}\big) = D\log(\lambda) - \lambda Y$$

  where D is total no. of deaths and Y is total risk time
- This is maximal for $\hat{\lambda} = D/Y$
- λ can depend on many parameters, so maximization is multidimensional . . .

Maximum likelihood estimation of a rate

- One person (p) followed over many intervals contributes:

$$\ell_p(\lambda) = \sum_i \big(d_{pi}\log(\lambda) - \lambda y_{pi}\big)$$

- all persons followed over many intervals contributes:

$$\sum_p \ell_p(\lambda) = \sum_{p,i} \big(d_{pi}\log(\lambda) - \lambda y_{pi}\big) = D\log(\lambda) - \lambda Y$$

  where D is total no. of deaths and Y is total risk time

- This is maximal for $\hat{\lambda} = D/Y$
- λ can depend on many parameters, so maximization is multidimensional...

Representation of follow-up: Lexis object

```
> L1 <- Lexis(entry = list(per = dodm, # "per"iod = calendar time of entry
+                          tfd = 0),   # "t"ime "f"rom "d"iabetes
+                exit = list(per = dox), #  calendar time of exit
+         exit.status = factor(!is.na(dodth),
+                              labels = c("DM","Dead")), # status at exit time
+                data = DM)
NOTE: entry.status has been set to "DM" for all.
> head(L1)
 lex.id     per tfd lex.dur lex.Cst lex.Xst sex   dobth    dodm   dodth   dooad
      1 1998.92   0   11.08      DM      DM   F 1940.26 1998.92      NA      NA
      2 2003.31   0    6.69      DM      DM   M 1939.22 2003.31      NA 2007.45
      3 2004.55   0    5.45      DM      DM   F 1918.30 2004.55      NA      NA
      4 2009.26   0    0.74      DM      DM   F 1965.23 2009.26      NA      NA
      5 2008.65   0    1.34      DM      DM   M 1932.88 2008.65      NA      NA
      6 2007.89   0    2.04      DM    Dead   F 1927.87 2007.89 2009.92      NA
 doins     dox   age
    NA 2010.00 58.66
    NA 2010.00 64.09
    NA 2010.00 86.25
    NA 2010.00 44.04
```

New variables in a `Lexis` object

`tfd`: time from diabetes diagnosis **at the time of entry**, therefore it is 0 for all persons; the entry time is 0 from the date of diabetes. Defines a **timescale** with name `tfd`.

`per` calendar time at the time of entry. Defines a **timescale** with name `per`

`lex.dur` the **length** of time a person is in state `lex.Cst`, here measured in years because all dates are

`lex.Cst` `Current state`, the state in which the `lex.dur` time is spent

`lex.Xst` `eXit state`, the state to which the person moves after the `lex.dur` time in `lex.Cst`

`lex.id` an id of each record in the source dataset. Can be explicitly set by `id=`

New variables in a Lexis object

tfd: time from diabetes diagnosis **at the time of entry**, therefore it is 0 for all persons; the entry time is 0 from the date of diabetes. Defines a **timescale** with name tfd.

per: calendar time at the time of entry. Defines a **timescale** with name per.

lex.dur the length of time a person is in state lex.Cst, here measured in years because all dates are

lex.Cst Current state, the state in which the lex.dur time is spent

lex.Xst eXit state, the state to which the person moves after the lex.dur time in lex.Cst

lex.id an id of each record in the source dataset. Can be explicitly set by id=

New variables in a `Lexis` object

`tfd`: time from diabetes diagnosis **at the time of entry**, therefore it is 0 for all persons; the entry time is 0 from the date of diabetes. Defines a **timescale** with name `tfd`.

`per`: calendar time at the time of entry. Defines a **timescale** with name `per`.

`lex.dur`: the **length** of time a person is in state `lex.Cst`, here measured in years because all dates are.

`lex.Cst`: Current state, the state in which the `lex.dur` time is spent.

`lex.Xst`: eXit state, the state to which the person moves after the `lex.dur` time in `lex.Cst`

`lex.id`: an id of each record in the source dataset. Can be explicitly set by `id=`

New variables in a `Lexis` object

`tfd`: time from diabetes diagnosis **at the time of entry**, therefore it is 0 for all persons; the entry time is 0 from the date of diabetes. Defines a **timescale** with name `tfd`.

`per`: calendar time at the time of entry. Defines a **timescale** with name `per`.

`lex.dur`: the **length** of time a person is in state `lex.Cst`, here measured in years because all dates are.

`lex.Cst`: Current state, the state in which the `lex.dur` time is spent.

`lex.Xst`: eXit state, the state to which the person moves after the `lex.dur` time in `lex.Cst`

`lex.id`: an id of each record in the source dataset. Can be explicitly set by `id=`

New variables in a Lexis object

tfd: time from diabetes diagnosis **at the time of entry**, therefore it is 0 for all persons; the entry time is 0 from the date of diabetes. Defines a **timescale** with name tfd.

per: calendar time at the time of entry. Defines a **timescale** with name per.

lex.dur: the **length** of time a person is in state lex.Cst, here measured in years because all dates are.

lex.Cst: Current state, the state in which the lex.dur time is spent.

lex.Xst: eXit state, the state to which the person moves after the lex.dur time in lex.Cst.

lex.id: an id of each record in the source dataset. Can be explicitly set by id=.

New variables in a `Lexis` object

- `tfd`: time from diabetes diagnosis **at the time of entry**, therefore it is 0 for all persons; the entry time is 0 from the date of diabetes. Defines a **timescale** with name `tfd`.
- `per`: calendar time at the time of entry. Defines a **timescale** with name `per`.
- `lex.dur`: the **length** of time a person is in state `lex.Cst`, here measured in years because all dates are.
- `lex.Cst`: Current `st`ate, the state in which the `lex.dur` time is spent.
- `lex.Xst`: eXit `st`ate, the state to which the person moves after the `lex.dur` time in `lex.Cst`.
- `lex.id`: an id of each record in the source dataset. Can be explicitly set by `id=`.

Lexis object: Overview of follow-up

Overkill?

The point is that the machinery generalizes to multistate data.

```
> summary(L1)

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

What is the average follow-up time for persons?

Lexis object: Overview of follow-up

Overkill?

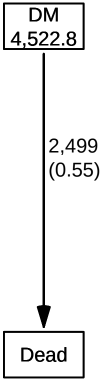
The point is that the machinery generalizes to multistate data.

```
> summary(L1)
```

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

What is the average follow-up time for persons?

```
> boxes(L1, boxpos = TRUE, scale.Y = 12, digits.R = 2)
```



Exercise 3

Cox model using the Lexis-specific variables:

```
> cl <- coxph(Surv(tfd,
+                  tfd + lex.dur,
+                  lex.Xst == "Dead") ~ sex + age,
+             data = L1)
```

`Surv(from-time, to-time, event indicator)`

Using the Lexis features:

```
> cL <- coxph.Lexis(L1, tfd ~ sex + age)
```

```
survival::coxph analysis of Lexis object L1:
Rates for the transition:
DM->Dead
Baseline timescale: tfd
```

```
> round(cbind(ci.exp(cL),
+             ci.exp(cl)), 3)
```

```
     exp(Est.)  2.5% 97.5% exp(Est.)  2.5% 97.5%
sexF     0.680 0.628 0.736     0.680 0.628 0.736
age      1.083 1.079 1.087     1.083 1.079 1.087
```

The crude Poisson model:

```
> pc <- glm(cbind(lex.Xst == "Dead", lex.dur) ~ sex + age,
+             family = poisreg,
+               data = L1)
```

or even simpler, by using the Lexis features:

```
> pL <- glm.Lexis(L1, ~ sex + age)

stats::glm Poisson analysis of Lexis object L1 with log link:
Rates for the transition:
DM->Dead

> round(cbind(ci.exp(pL),
+               ci.exp(pc)), 3)

            exp(Est.)  2.5% 97.5% exp(Est.)  2.5% 97.5%
(Intercept)     0.000 0.000 0.000     0.000 0.000 0.000
sexF            0.691 0.638 0.749     0.691 0.638 0.749
age             1.079 1.076 1.083     1.079 1.076 1.083
```

Poisson and Cox model

The crude Poisson model is a Cox-model with the (quite brutal) assumption that baseline rate is constant over time.

But results are similar:

```
> round(cbind(ci.exp(cL),
+              ci.exp(pL)[-1,]), 3)
     exp(Est.)  2.5% 97.5% exp(Est.)  2.5% 97.5%
sexF     0.680 0.628 0.736     0.691 0.638 0.749
age      1.083 1.079 1.087     1.079 1.076 1.083
```

Poisson and Cox model

The crude Poisson model is a Cox-model with the (quite brutal) assumption that baseline rate is constant over time.

But results are similar:

```
> round(cbind(ci.exp(cL),
+             ci.exp(pL)[-1,]), 3)
     exp(Est.)  2.5% 97.5% exp(Est.)  2.5% 97.5%
sexF     0.680 0.628 0.736     0.691 0.638 0.749
age      1.083 1.079 1.087     1.079 1.076 1.083
```

Baseline hazard: splitting time

```
> S1 <- splitMulti(L1, tfd = seq(0, 15, 0.5))
> summary(L1)

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

> summary(S1)

Transitions:
     To
From     DM Dead  Records:  Events: Risk time:  Persons:
  DM 111178 2499    113677     2499   54273.27      9996
```

What happended to no. records?

What happended to amount of risk time?

What happended to no. events?

```
> wh <- names(L1)[1:10] # names of variables in some order
> subset(L1, lex.id == 6)[,wh]
```

```
 lex.id     per tfd lex.dur lex.Cst lex.Xst sex   dobth    dodm   dodth
      6 2007.89   0    2.04      DM    Dead   F 1927.87 2007.89 2009.92
```

```
> subset(S1, lex.id == 6)[,wh]
```

```
 lex.id     per tfd lex.dur lex.Cst lex.Xst sex   dobth    dodm   dodth
      6 2007.89 0.0    0.50      DM      DM   F 1927.87 2007.89 2009.92
      6 2008.39 0.5    0.50      DM      DM   F 1927.87 2007.89 2009.92
      6 2008.89 1.0    0.50      DM      DM   F 1927.87 2007.89 2009.92
      6 2009.39 1.5    0.50      DM      DM   F 1927.87 2007.89 2009.92
      6 2009.89 2.0    0.04      DM    Dead   F 1927.87 2007.89 2009.92
```

In `S1` each record now represents a small interval (0.5 year) of follow-up for a person, so each person has many records.

Natural splines for baseline hazard

```
> ps <- glm(cbind(lex.Xst == "Dead", lex.dur)
+             ~ Ns(tfd, knots = seq(0, 15, 5)) + sex + age,
+           family = poisreg,
+             data = S1)
```

or even simpler:

```
> ps <- glm.Lexis(S1, ~ Ns(tfd, knots = seq(0, 15, 5)) + sex + age)

stats::glm Poisson analysis of Lexis object S1 with log link:
Rates for the transition:
DM->Dead

> ci.exp(ps)

                                   exp(Est.)          2.5%       97.5%
(Intercept)                     0.0002647664 0.0002005196 0.000349598
Ns(tfd, knots = seq(0, 15, 5))1 2.4823273077 1.9470986530 3.164682413
Ns(tfd, knots = seq(0, 15, 5))2 1.6172454509 1.0715875536 2.440755158
Ns(tfd, knots = seq(0, 15, 5))3 2.2067211974 1.3528945106 3.599407349
sexF                            0.6798768856 0.6276865380 0.736406712
age                             1.0832396476 1.0793524197 1.087140875
```

Comparing with estimates from the Cox-model and from the model with constant baseline:

```
> round(cbind(ci.exp(cl),
+             ci.exp(ps, subset = c("sex","age")),
+             ci.exp(pc, subset = c("sex","age"))), 4)
```

```
     exp(Est.)   2.5%  97.5% exp(Est.)   2.5%  97.5% exp(Est.)   2.5%  97.5%
sexF    0.6797 0.6275 0.7362    0.6799 0.6277 0.7364    0.6911 0.6381 0.7485
age     1.0832 1.0793 1.0871    1.0832 1.0794 1.0871    1.0795 1.0757 1.0832
```

But where is the baseline hazard?

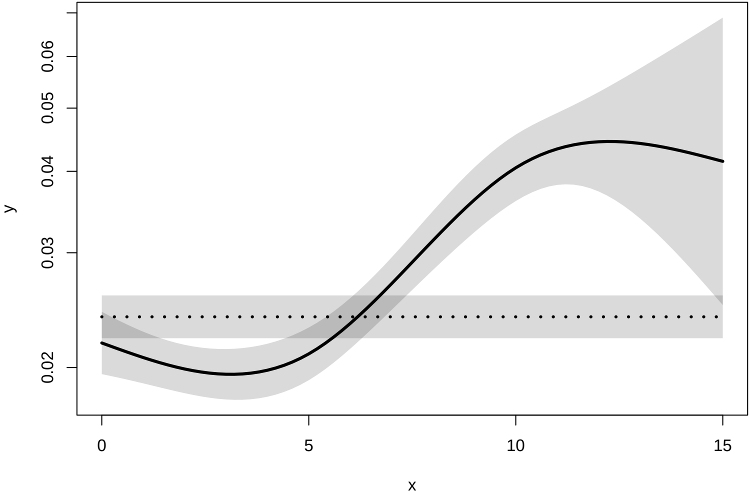
ps is a model for the hazard so we can predict the baseline hazard at defined values for given sets of covariates in the model:

```
> prf <- data.frame(tfd = seq(0, 15, 0.2),
+                   sex = "F",
+                   age = 60)
```

We can over-plot with the predicted rates from the model where mortality rates are constant, the only change is the model (pc instead of ps):

```
> matshade(prf$tfd, ci.pred(ps, prf),
+          plot = TRUE, log = "y", lwd = 3)
> matshade(prf$tfd, ci.pred(pc, prf), lty = 3, lwd = 3)
```

Here is the baseline hazard!



What are the units on the y-axis? Describe the mortality rates as a function of tfd

Survival function and hazard function

$$S(t) = \exp(-\int_0^t \lambda(u)\,\mathrm{d}u)$$

Simple, but the CI for $S(t)$ not so simple

Implemented in the `ci.surv` function

Arguments: 1 model, 2 prediction data frame, 3 equidistance

Prediction data frame must correspond to a sequence of equidistant time points:

```
> matshade(prf$cfd, ci.surv(pc, prf, intl = 0.2),
+              plot = TRUE, lwd = 3, ylim = c(0.5, 1))
>     lines(prf$cfd, ci.surv(pc, prf, intl = 0.2)[,1], col="blue")
> lines(survfit(c1, newdata = data.frame(sex = "M", age = 60)),
+           lwd = 2, lty = 1, col = "limegreen")
```

Survival function and hazard function

$$S(t) = \exp(-\int_0^t \lambda(u)\,\mathrm{d}u)$$

Simple, but the CI for $S(t)$ not so simple. . .

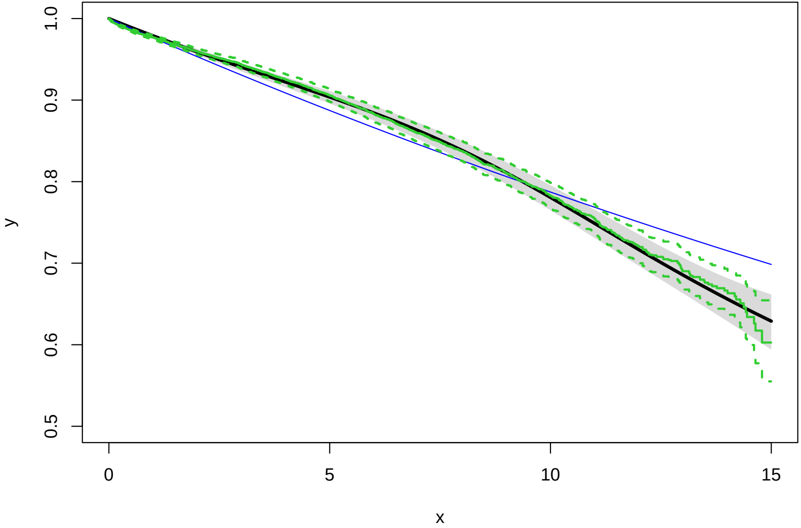
Implemented in the `ci.surv` function

Arguments: 1:model, 2:prediction data frame, 3:equidistance

Prediction data frame must correspond to a sequence of equidistant time points:

```
> matshade(prf$tfd, ci.surv(ps, prf, intl = 0.2),
+          plot = TRUE, lwd = 3, ylim = c(0.5, 1))
>    lines(prf$tfd, ci.surv(pc, prf, intl = 0.2)[,1], col="blue")
> lines(survfit(c1, newdata = data.frame(sex = "F", age = 60)),
+        lwd = 2, lty = 1, col = "limegreen")
```

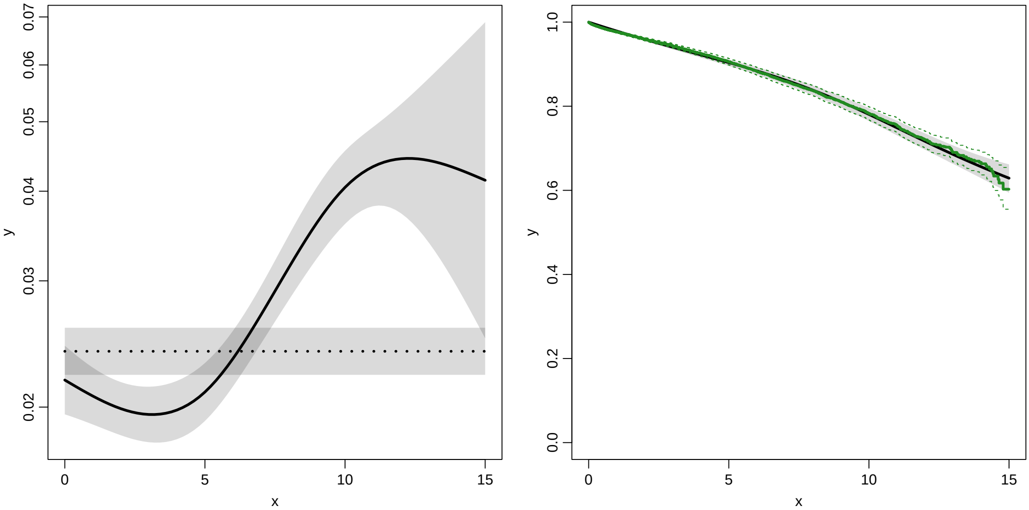
Survival functions



Hazard and survival functions

```
> par(mfrow = c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6)
> #
> # hazard scale
> matshade(prf$tfd, ci.pred(ps, prf),
+          plot = TRUE, log = "y", lwd = 3)
> matshade(prf$tfd, ci.pred(pc, prf), lty = 3, lwd = 3)
> #
> # survival
> matshade(prf$tfd, ci.surv(ps, prf, intl = 0.2),
+          plot = TRUE, ylim = 0:1, lwd = 3)
> lines(survfit(c1, newdata = data.frame(sex = "F", age = 60)),
+       col = "forestgreen", lwd = 3, conf.int = FALSE)
> lines(survfit(c1, newdata = data.frame(sex = "F", age = 60)),
+       col = "forestgreen", lwd = 1, lty = 1)
```

Hazard and survival functions

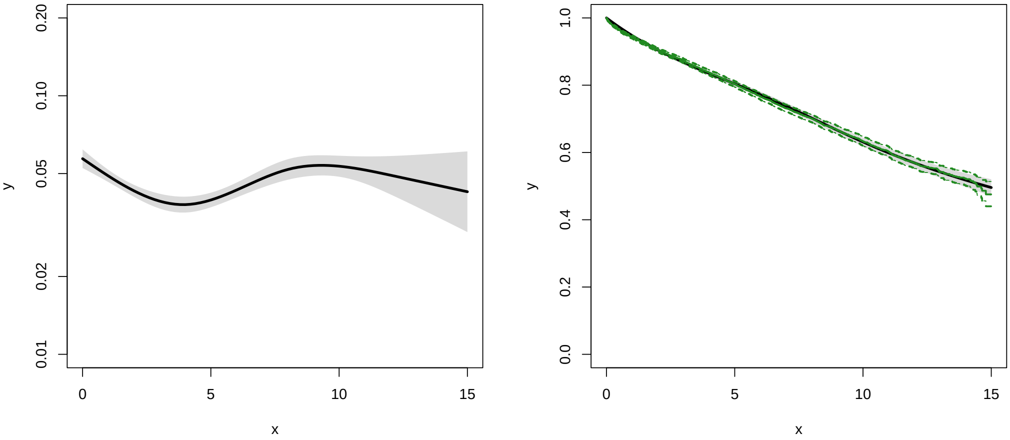


K-M estimator and smooth Poisson model

Kaplan-Meier estimator compared to survival from corresponding Poisson-model, which is the model with time from diabetes (tfd) as the only covariate:

```
> par(mfrow=c(1,2))
> pk <- glm(cbind(lex.Xst == "Dead",
+                 lex.dur) ~ Ns(tfd, knots = seq(0, 12, 4)),
+           family = poisreg,
+             data = S1)
> # hazard
> matshade(prf$tfd, ci.pred(pk, prf),
+          plot = TRUE, log = "y", lwd = 3, ylim = c(0.01,0.2))
> # survival from smooth model
> matshade(prf$tfd, ci.surv(pk, prf, intl = 0.2) ,
+          plot = TRUE, lwd = 3, ylim = 0:1)
> # K-M estimator
> lines(km, lwd = 1, col = "forestgreen")
> lines(km, lwd = 2, col = "forestgreen", confint = FALSE)
```

K-M estimator and smooth Poisson model

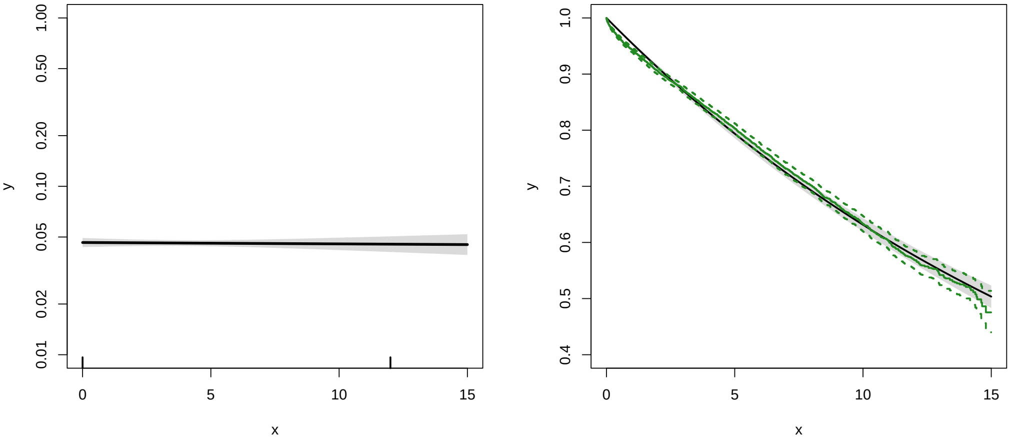


K-M estimator and smooth Poisson model

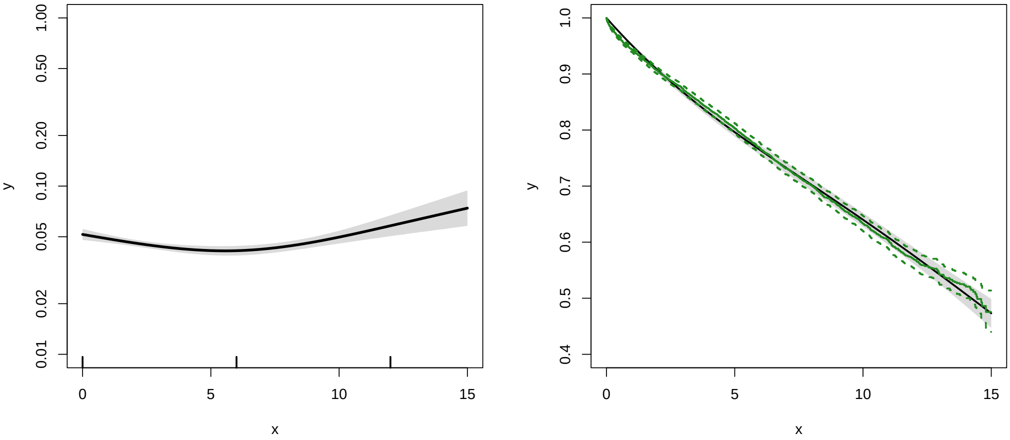
We can explore how the tightness of the knots in the smooth model influence the underlying hazard and the resulting survival function:

```
> zz <- function(dk) # distance between knots
+ {
+ par(mfrow=c(1,2))
+ kn <- seq(0, 12, dk)
+ pk <- glm(cbind(lex.Xst == "Dead",
+                  lex.dur) ~ Ns(tfd, knots = kn),
+            family = poisreg,
+              data = S1)
+ matshade(prf$tfd, ci.pred(pk, prf),
+           plot = TRUE, log = "y", lwd = 3, ylim = c(0.01,1))
+ rug(kn, lwd=2)
+
+ matshade(prf$tfd, ci.surv(pk, prf, intl = 0.2) ,
+           plot = TRUE, lwd = 2, ylim = c(0.4, 1))
+ lines(km, lwd = 2, col = "forestgreen")
+ }
> zz(12)
```

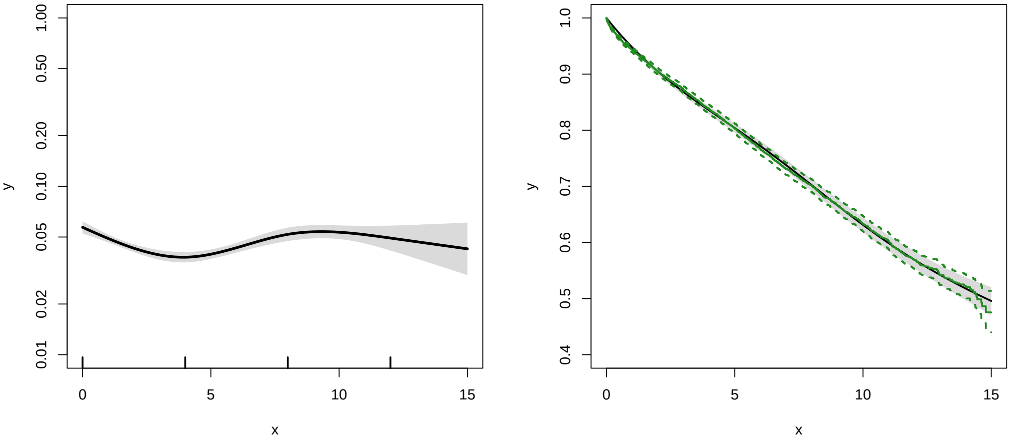
K-M estimator and smooth Poisson model



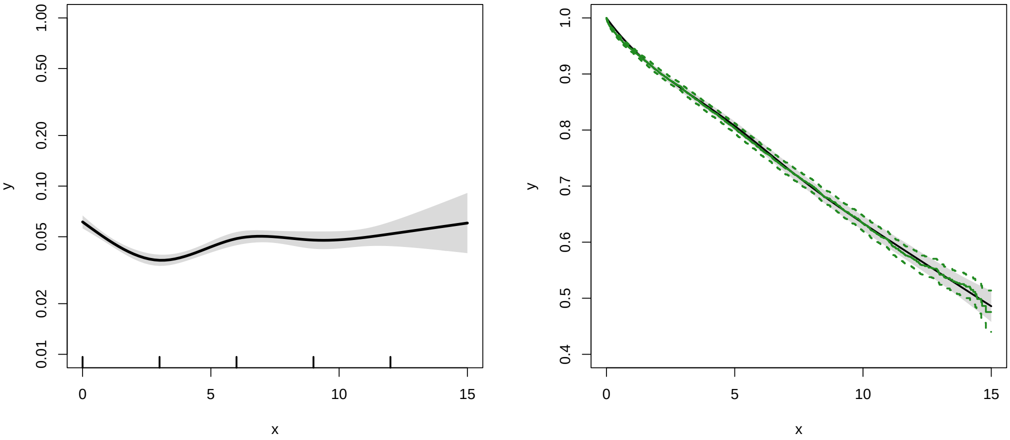
K-M estimator and smooth Poisson model



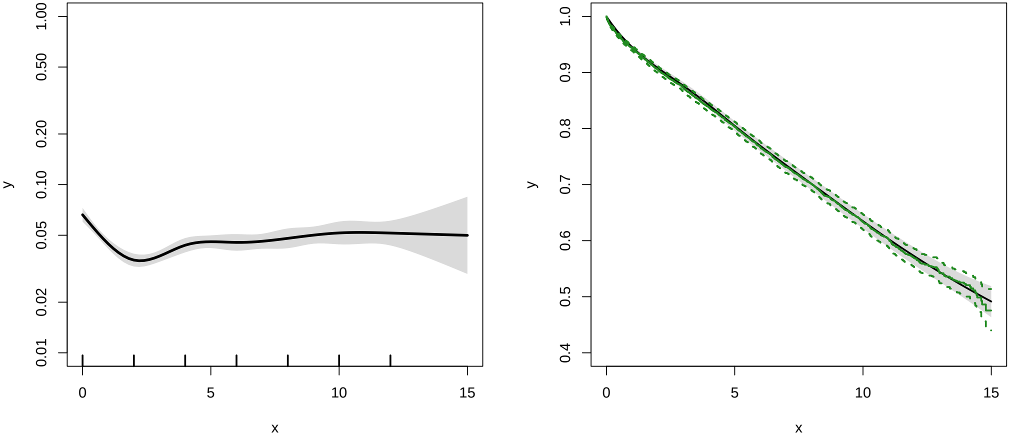
K-M estimator and smooth Poisson model



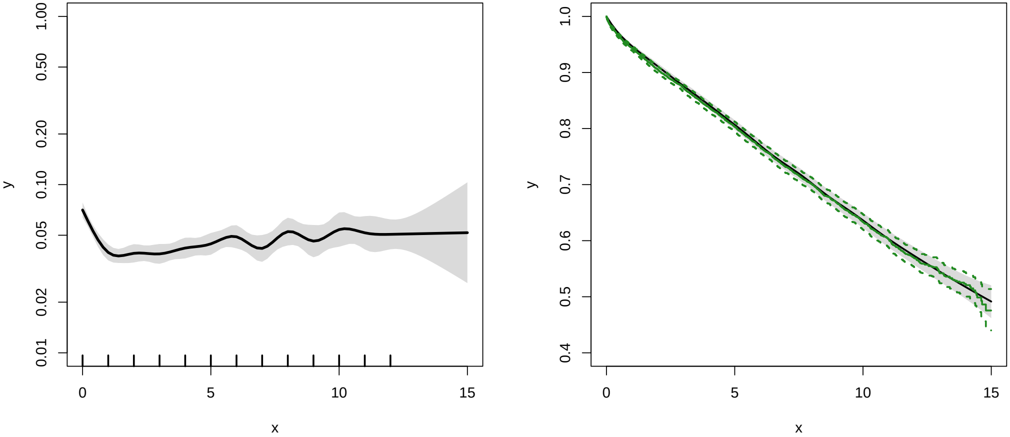
K-M estimator and smooth Poisson model



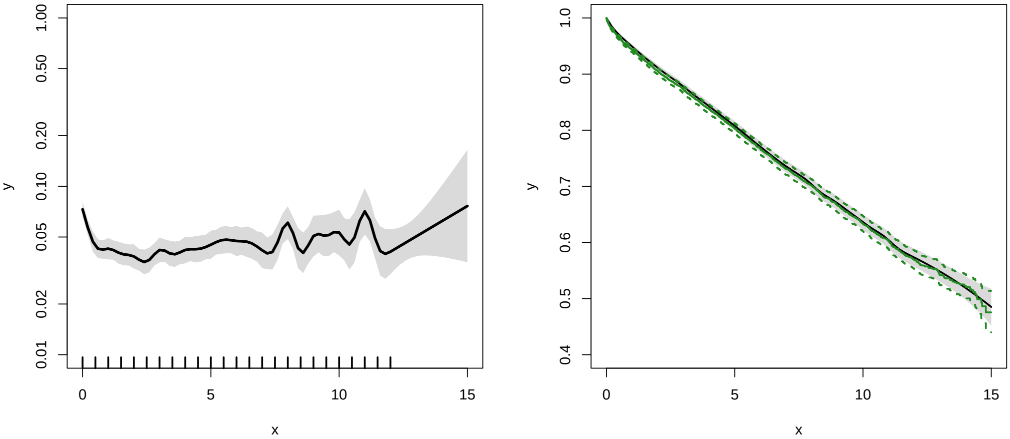
K-M estimator and smooth Poisson model



K-M estimator and smooth Poisson model



K-M estimator and smooth Poisson model



Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

Survival analysis summary

- 1 to 1 correspondence between
  - hazard function + starting point
  - survival function
- K-M and Cox use a very detailed baseline hazard (and omits it)
- Smooth parametric hazard function more credible:
  - Define `Lexis` object
  - Split along time
  - Fit Poisson model: smooth effect of time
  - Define prediction data frame
  - `ci.pred` to get baseline rates
  - `ci.surv` to get baseline survival

```
> data(DMlate)
> DMlate <- mutate(subset(DMlate, dodm < dox), age = dodm - dobth)
> Lx <- Lexis(exit = list(tfd = dox - dodm), # tfd at exit
+        exit.status = factor(!is.na(dodth)), # status at exit time
+               data = DMlate)
> sL <- splitMulti(Lx, tfd = seq(0, 15, 1/12))
```

Smooth parametric hazard function

```
> m0 <- glm.Lexis(sL, ~ Ns(tfd, knots = seq(0, 14, , 5)) + sex + age)
```

Prediction data frame

```
> nd <- data.frame(tfd = seq(0, 15, 1/10), sex = "M", age = 65)
```

Predicted rates and survival

```
> rate <- ci.pred(m0, nd) # rates per year
> surv <- ci.surv(m0, nd, int = 1/10)
```

Plot the rates and the survival function

```
> matshade(nd$tfd, rate, log = "y", plot = TRUE)
> matshade(nd$tfd, surv, ylim = c(0, 1), plot = TRUE)
```

Exercises 4, 5

Competing risks

estimation

Survival, mortality,
competing risks and
expected lifetime
EDEG 2025 / Umeå University,17 May 2025

Lexis object from DM to Death

```
> data(DMlate)
> dl <- mutate(DMlate, dofin = pmin(dodth, doins, dox, na.rm = TRUE),
+                        xstat = factor(case_when(dofin == dodth ~ "Dead",
+                                                 dofin == doins ~ "Ins",
+                                                           TRUE ~ "DM"),
+                                       levels = c("DM", "Ins", "Dead")))
> Ldm <- Lexis(exit = list(tfd = dofin - dodm),
+       exit.status = xstat,
+              data = dl)
```

```
NOTE: entry.status has been set to "DM" for all.
NOTE: entry is assumed to be 0 on the tfd timescale.
NOTE: Dropping  101  rows with duration of follow up < tol
```

```
> summary(Ldm)
```

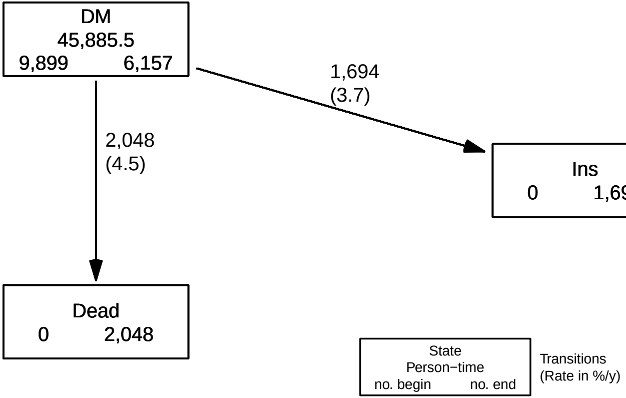
```
Transitions:
     To
From    DM  Ins Dead  Records:  Events: Risk time:  Persons:
  DM 6157 1694 2048      9899     3742   45885.49      9899
```

Produce graphical overview of FU

```
> boxes(Ldm, boxpos = TRUE, scale.R = 100, show.BE = TRUE)
> legendbox(70, 10, rates = "\n(Rate in %/y)")
> args(legendbox)
```

```
function (x, y, state = "State", py = "Person-time", begin = "no. begin",
    end = "no. end", trans = "Transitions", rates = "\n(Rate)",
    font = 1, right = !left, left = !right, ...)
NULL
```

Transitions: competing rates



Exercise 6

Survival function?

$$
\begin{aligned}
S(t) &= \exp\left(-\int_0^t \lambda_{\mathsf{Ins}}(u) + \mu(u)\,\mathrm{d}u\right)\\
S(t) &= \exp\left(-\int_0^t \lambda_{\mathsf{Ins}}(u)\,\mathrm{d}u\right)\\
S(t) &= \exp\left(-\int_0^t \mu(u)\,\mathrm{d}u\right)
\end{aligned}
$$

Survival function and Cumulative risk function

`survfit` does the trick; the requirements are:

1. (start, stop, event) arguments to `Surv`
2. the third argument to the `Surv` function is a `factor`
3. an `id` argument is given, pointing to an id variable that links together records belonging to the same person
4. the initial state (`DM`) must be the first level of the factor (for a `Lexis` object, `lex.Cst`)

Survival function and Cumulative risk function

survfit does the trick; the requirements are:

1. (start, stop, event) arguments to Surv
2. the third argument to the Surv function is a factor
3. an id argument is given, pointing to an id variable that links together records belonging to the same person
4. the initial state (DM) must be the first level of the factor (in a Lexis object, lex.Cst)

Survival function and Cumulative risk function

survfit does the trick; the requirements are:

1. (start, stop, event) arguments to Surv
2. the third argument to the Surv function is a factor
3. an id argument is given, pointing to an id variable that links together records belonging to the same person.
4. the initial state (DM) must be the first level of the factor (for a Lexis object, lex.Cst)

Survival function and Cumulative risk function

`survfit` does the trick; the requirements are:

1. (start, stop, event) arguments to `Surv`
2. the third argument to the `Surv` function is a `factor`
3. an `id` argument is given, pointing to an id variable that links together records belonging to the same person.
4. the initial state (`DM`) must be the first level of the factor (in a `Lexis` object, `lex.Cst`)

Survival function and Cumulative risk function

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

[1] "DM"   "Ins"  "Dead"

> m3 <- survfit(Surv(tfd, tfd + lex.dur, lex.Xst) ~ 1,
+                id = lex.id,
+              data = Ldm)
> m3$states

[1] "(s0)" "Ins"  "Dead"

> head(cbind(time = m3$time, m3$pstate))

            time      (s0)          Ins         Dead
[1,] 0.002737851 0.9988888 0.0003030609 0.0008081624
[2,] 0.005475702 0.9982825 0.0005051424 0.0012123254
[3,] 0.008213552 0.9972721 0.0011113869 0.0016164884
[4,] 0.010951403 0.9955543 0.0024250496 0.0020206923
[5,] 0.013689254 0.9939374 0.0038397633 0.0022227943
[6,] 0.016427105 0.9916133 0.0057597319 0.0026269982
```

—this is called the Aalen-Johansen estimator of state probabilities

Survival function and Cumulative risk function

the Aalen-Johansen estimator of state probabilities is obtained easily from a Lexis object

```
> aaj <- AaJ.Lexis(Ldm)
NOTE: Timescale is tfd
> head(cbind(time = aaj$time, aaj$pstate))
            time        DM         Dead          Ins
[1,] 0.002737851 0.9988888 0.0008081624 0.0003030609
[2,] 0.005475702 0.9982825 0.0012123254 0.0005051424
[3,] 0.008213552 0.9972721 0.0016164884 0.0011113869
[4,] 0.010951403 0.9955543 0.0020206923 0.0024250496
[5,] 0.013689254 0.9939374 0.0022227943 0.0038397633
[6,] 0.016427105 0.9916133 0.0026269982 0.0057597319
```

Survival function and cumulative risks

$$\begin{aligned}
S(t) &= \exp\Big(-\int_0^t \lambda(u) + \mu(u)\,\mathrm{d}u\Big) \\
R_{\texttt{Dead}}(t) &= \int_0^t \mu(u)S(u)\,\mathrm{d}u \\
R_{\texttt{Ins}}(t) &= \int_0^t \lambda(u)S(u)\,\mathrm{d}u) \\
&= \int_0^t \lambda(u)\exp\Big(-\int_0^u \lambda(s) + \mu(s)\,\mathrm{d}s\Big)\,\mathrm{d}u
\end{aligned}$$

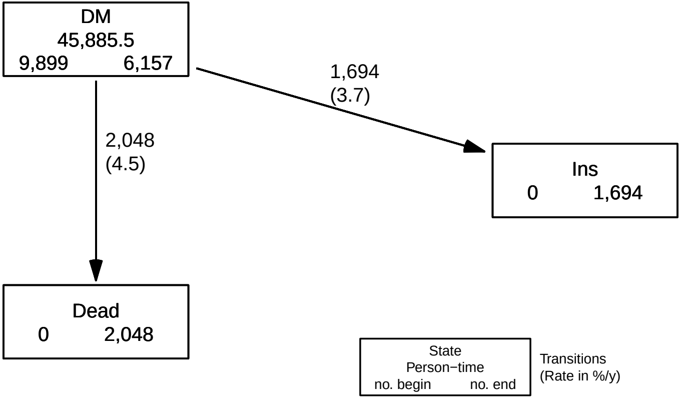
$$S(t) + R_{\texttt{Ins}}(t) + R_{\texttt{Dead}}(t) = 1, \quad \forall t$$

Survival function and cumulative risks

$$
\begin{aligned}
S(t) &= \exp\left(-\int_0^t \lambda(u) + \mu(u)\,\mathrm{d}u\right) \\
R_{\texttt{Dead}}(t) &= \int_0^t \mu(u) S(u)\,\mathrm{d}u \\
R_{\texttt{Ins}}(t) &= \int_0^t \lambda(u) S(u)\,\mathrm{d}u) \\
&= \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s) + \mu(s)\,\mathrm{d}s\right)\mathrm{d}u
\end{aligned}
$$

$$
S(t) + R_{\texttt{Ins}}(t) + R_{\texttt{Dead}}(t) = 1, \quad \forall t
$$

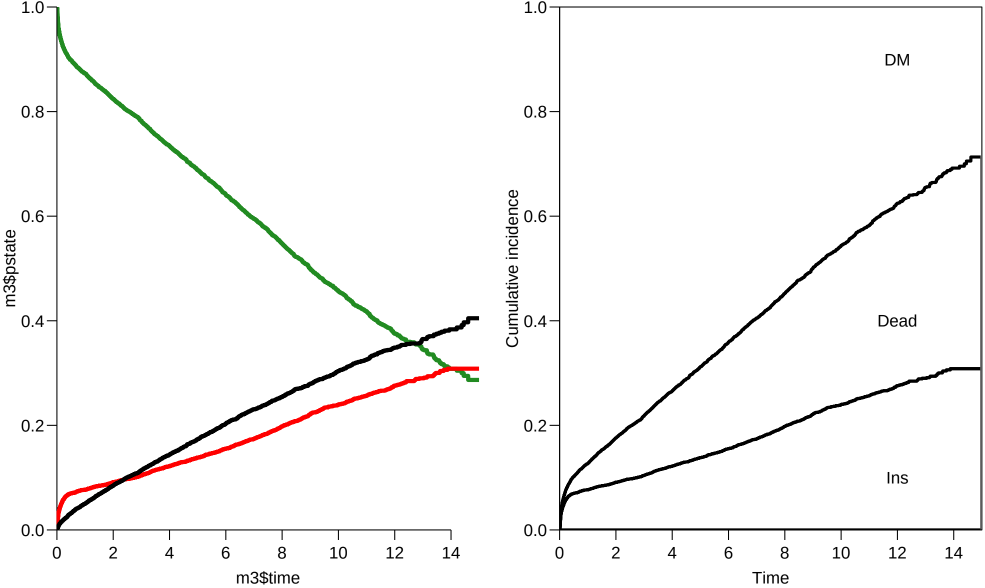
Transitions: competing rates



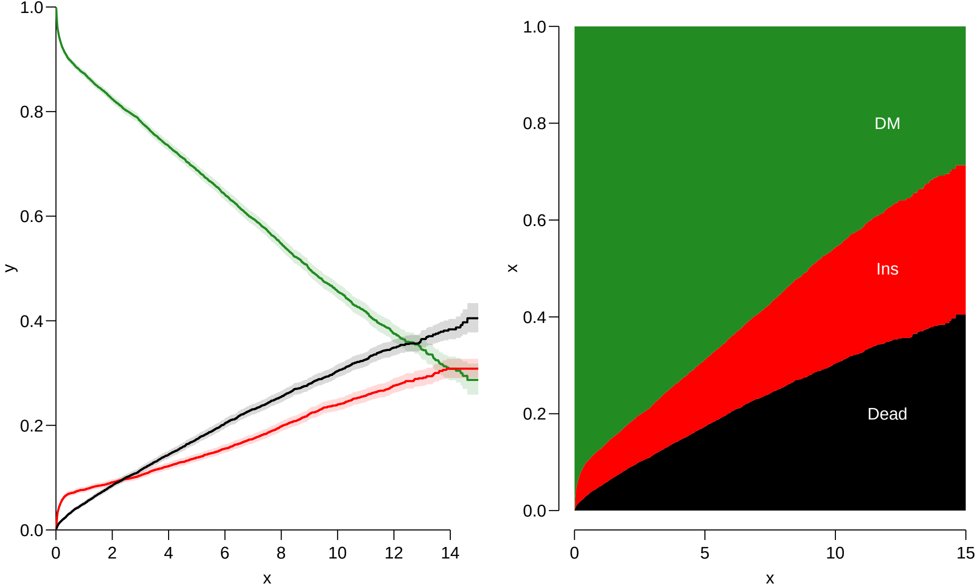
Survival function and cumulative risks

```
> par( mfrow=c(1,2) )
> matplot(m3$time, m3$pstate,
+         type="s", lty=1, lwd=4,
+         col=c("ForestGreen","red","black"),
+         xlim=c(0,15), xaxs="i",
+         ylim=c(0,1), yaxs="i" )
> stackedCIF(m3, lwd = 3, xlim = c(0,15), xaxs = "i", yaxs = "i" )
> text(rep(12,3), c(0.9,0.1,0.4), levels(Ldm))
> box(bty="o")

> par(mfrow = c(1, 2))
> matshade(m3$time, cbind(m3$pstate,
+                         m3$lower,
+                         m3$upper)[, c(1, 4, 7, 2, 5, 8, 3, 6, 9)],
+          plot = TRUE, lty = 1, lwd = 2,
+          col = clr <- c("ForestGreen","red","black"),
+          xlim=c(0,15), xaxs="i",
+          ylim = c(0,1), yaxs = "i")
> mat2pol(m3$pstate, perm = 3:1, x = m3$time, col = clr[3:1])
> text(rep(12, 3), c(0.8, 0.5, 0.2), levels(Ldm), col = "white")
```

Survival and cumulative risk functions



Survival and cumulative risk functions



Survival function and cumulative risks: don't

$$
\begin{aligned}
R_{\texttt{Ins}}(t) &= \int_0^t \lambda(u) S(u)\,\mathrm{d}u) \\
&= \int_0^t \lambda(u) \exp\Big(-\int_0^u \lambda(s) + \mu(s)\,\mathrm{d}s\Big)\,\mathrm{d}u \\
&\neq \int_0^t \lambda(u) \exp\Big(-\int_0^u \lambda(s)\,\mathrm{d}s\Big)\,\mathrm{d}u \\
&= 1 - \exp\Big(-\int_0^t \lambda(s)\,\mathrm{d}s\Big)
\end{aligned}
$$

nice formula, but wrong!

Probability of Ins assuming Dead does not exist and rate of Ins unchanged!

$\exp(-\int_0^t \lambda(s)\,\mathrm{d}s)$ known as "net survival" or "cause specific survival"

Survival function and cumulative risks: don't

$$
\begin{aligned}
R_{\texttt{Ins}}(t) &= \int_0^t \lambda(u) S(u)\,\mathrm{d}u) \\
&= \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s) + \mu(s)\,\mathrm{d}s\right)\mathrm{d}u \\
&\neq \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s)\,\mathrm{d}s\right)\mathrm{d}u \\
&= 1 - \exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right) \text{ — nice formula, but wrong!}
\end{aligned}
$$

Probability of Ins assuming Dead does not exist and rate of Ins unchanged!

$\exp(-\int_0^t \lambda(s)\,\mathrm{d}s)$ known as "net survival" or "cause specific survival" …

Survival function and cumulative risks: don't

$$
\begin{aligned}
R_{\texttt{Ins}}(t) &= \int_0^t \lambda(u) S(u)\,\mathrm{d}u) \\
&= \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s) + \mu(s)\,\mathrm{d}s\right)\mathrm{d}u \\
&\neq \int_0^t \lambda(u) \exp\left(-\int_0^u \lambda(s)\,\mathrm{d}s\right)\mathrm{d}u \\
&= 1 - \exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right) \text{ — nice formula, but wrong!}
\end{aligned}
$$

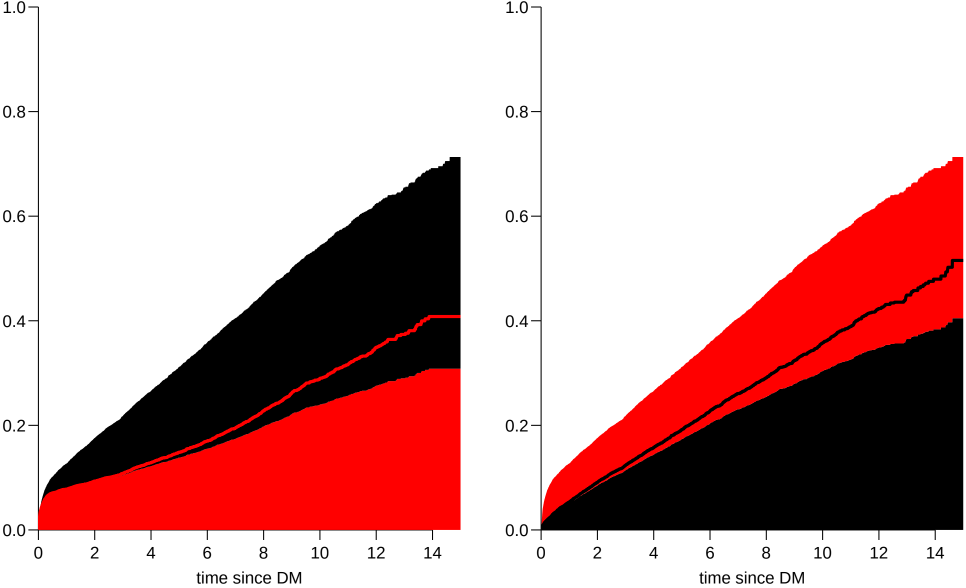
Probability of `Ins` **assuming** `Dead` does not exist **and** rate of `Ins` unchanged!

$\exp\left(-\int_0^t \lambda(s)\,\mathrm{d}s\right)$ known as “net survival” or ”cause specific survival”…

Survival function and cumulative risks—don’t

```
> m2 <- survfit(Surv(tfd,
+                    tfd + lex.dur,
+                    lex.Xst == "Ins" ) ~ 1,
+               data = Ldm)
> M2 <- survfit(Surv(tfd,
+                    tfd + lex.dur,
+                    lex.Xst == "Dead") ~ 1,
+               data = Ldm)
> par(mfrow = c(1,2))
> mat2pol(m3$pstate, c(2,3,1), x = m3$time,
+         col = c("red", "black", "transparent"),
+         xlim=c(0,15), xaxs="i",
+         yaxs = "i", xlab = "time since DM", ylab = "" )
>   lines(m2$time, 1 - m2$surv, lwd = 3, col = "red" )
> mat2pol(m3$pstate, c(3,2,1), x = m3$time, yaxs = "i",
+         col = c("black","red","transparent"),
+         xlim=c(0,15), xaxs="i",
+         yaxs = "i", xlab = "time since DM", ylab = "" )
>   lines(M2$time, 1 - M2$surv, lwd = 3, col = "black" )
```

Survival and cumulative risk functions



Cause-specific rates

- There is nothing wrong with modeling the cause-specific event-rates, the problem lies in how you transform them into probabilities.
- The relevant model for a competing risks situation normally consists of separate models for each of the cause-specific rates.
- These models have no common parameters (effects of time or other covariates are not constrained to be the same).
- … not for statistical reasons, but for substantial reasons: it is unlikely that rates of different types of event (insulin initiation and death, say) depend on time in the same way.

Cause-specific rates

- There is nothing wrong with modeling the cause-specific event-rates, the problem lies in how you transform them into probabilities.
- The relevant model for a competing risks situation normally consists of separate models for each of the cause-specific rates.
- These models have no common parameters (effects of time or other covariates are not constrained to be the same).
- … not for statistical reasons, but for substantial reasons: it is unlikely that rates of different types of event (Insulin initiation and death, say) depend on time in the same way.

Cause-specific rates

- There is nothing wrong with modeling the cause-specific event-rates, the problem lies in how you transform them into probabilities.
- The relevant model for a competing risks situation normally consists of separate models for each of the cause-specific rates.
- These models have no common parameters (effects of time or other covariates are not constrained to be the same).
- … not for statistical reasons, but for substantial reasons: it is unlikely that rates of different types of event (insulin initiation and death, say) depend on time in the same way.

Cause-specific rates

- There is nothing wrong with modeling the cause-specific event-rates, the problem lies in how you transform them into probabilities.
- The relevant model for a competing risks situation normally consists of separate models for each of the cause-specific rates.
- These models have no common parameters (effects of time or other covariates are not constrained to be the same).
- . . . not for statistical reasons, but for **substantial** reasons: it is unlikely that rates of different types of event (Insulin initiation and death, say) depend on time in the same way.

Cause-specific rates

```
> Sdm <- splitMulti(Ldm, tfd = seq(0, 20, 0.1))
> summary(Ldm)

Transitions:
     To
From   DM  Ins Dead  Records:  Events: Risk time:  Persons:
  DM 6157 1694 2048      9899     3742   45885.49      9899

> summary(Sdm)

Transitions:
     To
From     DM  Ins Dead  Records:  Events: Risk time:  Persons:
  DM 460054 1694 2048    463796     3742   45885.49      9899
```

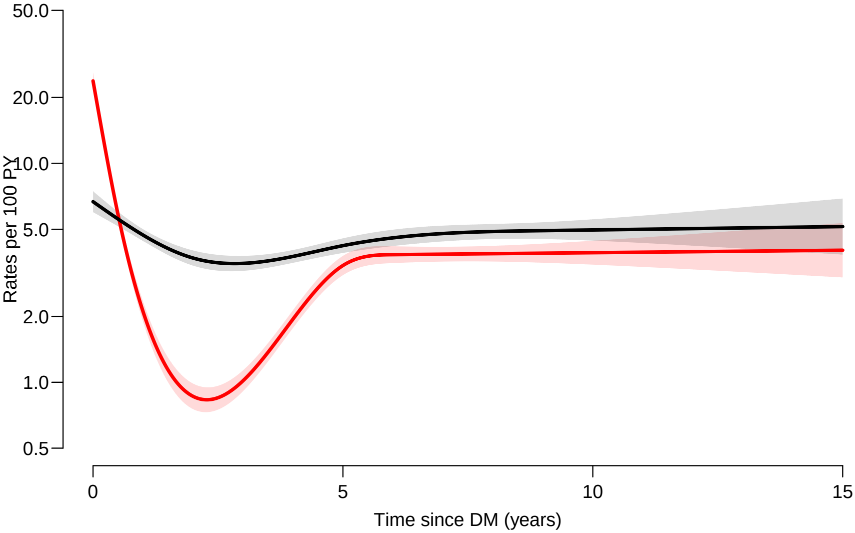
Cause-specific rates

```
> round(cbind(
+ with(subset(Sdm, lex.Xst == "Ins" ), quantile(tfd + lex.dur, 0:4/4)),
+ with(subset(Sdm, lex.Xst == "Dead"), quantile(tfd + lex.dur, 0:4/4))), 2)

      [,1]  [,2]
0%    0.00  0.00
25%   0.11  1.10
50%   1.82  3.08
75%   5.77  5.83
100% 13.88 14.61

> ikn <- c(0, 0.5, 3, 10)
> dkn <- c(0, 2.0, 5,  9)
>  Ins.glm <- glm.Lexis(Sdm, ~ Ns(tfd, knots = ikn), to = "Ins" )

stats::glm Poisson analysis of Lexis object Sdm with log link:
Rates for the transition:
DM->Ins

> Dead.glm <- glm.Lexis(Sdm, ~ Ns(tfd, knots = dkn), to = "Dead")

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

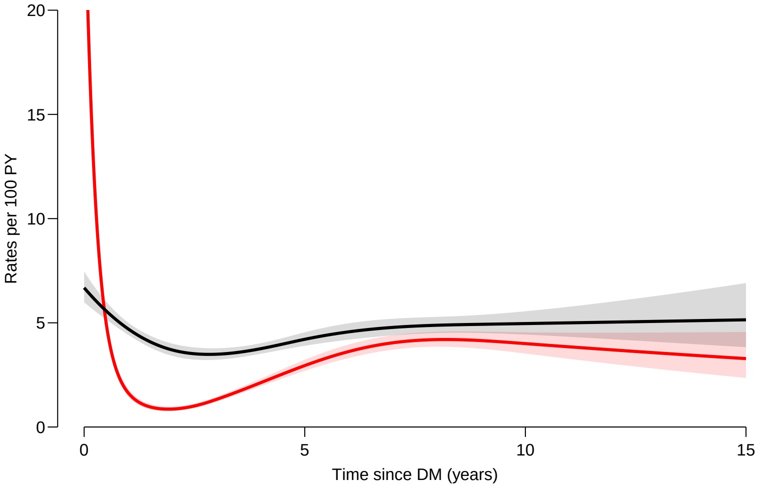
Cause-specific rates

```
> int <- 0.01
> nd <- data.frame(tfd = seq(0, 15, int))
> l.glm <- ci.pred( Ins.glm, nd)
> m.glm <- ci.pred(Dead.glm, nd)
> matshade(nd$tfd,
+          cbind(l.glm, m.glm) * 100,
+          plot = TRUE,
+          yaxs="i", ylim = c(0, 20),
+        # log = "y", ylim = c(2, 20),
+          col = rep(c("red","black"), 2), lwd = 3,
+          xlab = "Time since DM (years)",
+          ylab = "Rates per 100 PY")
```

Survival and cumulative risk functions



Survival and cumulative risk functions



Exercise 7, 8

∗ Integrals with R

- Integrals look scary to many people, but they are really just areas under curves.
- In R, a curve of the function $\mu(t)$ is a set of two vectors: one vector of ts and one vector $y = \mu(t)$s
- When we have a model such as the `glm` above that estimates the mortality as a function of time (`tfd`), we can get the mortality as a funtion of time by first choosing the timepoints, say from 0 to 15 years in steps of 0.01 year ($\approx$ 4 days)
- Using `ci.pred` on this gives the predicted rates
- Then use the formuale with all the integrals to get the state probabilities

* Integrals with R

- Integrals look scary to many people, but they are really just areas under curves.
- In R, a curve of the function $\mu(t)$ is a set of two vectors: one vector of ts and one vector $y = \mu(t)$s.
- When we have a model such as the `glm` above that estimates the mortality as a function of time (`tfd`), we can get the mortality as a funtion of time by first choosing the t timepoints say from 0 to 15 years in steps of 0.01 year ($\approx$ 4 days)
- Using `ci.pred` on this gives the predicted rates
- Then use the formuale with all the integrals to get the state probabilities

∗ Integrals with R

- Integrals look scary to many people, but they are really just areas under curves.
- In R, a curve of the function $\mu(t)$ is a set of two vectors: one vector of ts and one vector $y = \mu(t)$s.
- When we have a model such as the `glm` above that estimates the mortality as a function of time (`tfd`), we can get the mortality as a funtion of time by first choosing the timepoints, say from 0 to 15 years in steps of 0.01 year (≈ 4 days)
- Using `ci.pred` on this gives the predicted rates
- Then use the formuale with all the integrals to get the state probabilities

∗ Integrals with R

- Integrals look scary to many people, but they are really just areas under curves.
- In R, a curve of the function $\mu(t)$ is a set of two vectors: one vector of ts and one vector $y = \mu(t)$s.
- When we have a model such as the `glm` above that estimates the mortality as a function of time (`tfd`), we can get the mortality as a funtion of time by first choosing the timepoints, say from 0 to 15 years in steps of 0.01 year (≈ 4 days)
- Using `ci.pred` on this gives the predicted rates
- Then use the formuale with all the integrals to get the state probabilities

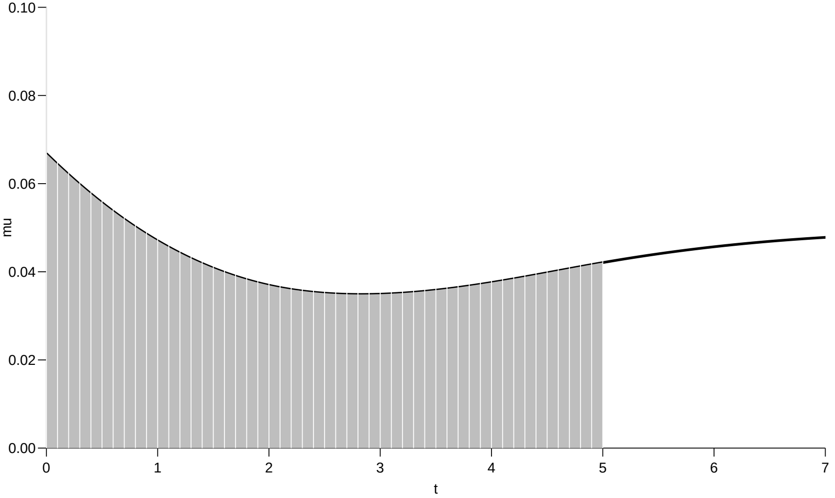
∗ Integrals with R

- Integrals look scary to many people, but they are really just areas under curves.
- In R, a curve of the function $\mu(t)$ is a set of two vectors: one vector of ts and one vector $y = \mu(t)$s.
- When we have a model such as the `glm` above that estimates the mortality as a function of time (`tfd`), we can get the mortality as a funtion of time by first choosing the timepoints, say from 0 to 15 years in steps of 0.01 year (≈ 4 days)
- Using `ci.pred` on this gives the predicted rates
- Then use the formuale with all the integrals to get the state probabilities.

∗ Integrals with R

```
> t <- seq(0, 15, 0.01)
> nd <- data.frame(tfd = t)
> mu <- ci.pred(Dead.glm, nd)[,1]
> head(cbind(t, mu))

     t         mu
1 0.00 0.06681677
2 0.01 0.06657067
3 0.02 0.06632549
4 0.03 0.06608123
5 0.04 0.06583789
6 0.05 0.06559547

> plot(t, mu, type="l", lwd = 3,
+      xlim = c(0, 7), xaxs = "i",
+      ylim = c(0, 0.1), yaxs = "i")
> polygon(t[c(1:501,501:1)], c(mu[1:501], rep(0, 501)),
+         col = "gray", border = "transparent")
> abline(v=0:50/10, col="white")
```

* Integrals with R



* Numerical integration with R

```
> mid <- function(x) x[-1] - diff(x) / 2
> (x <- c(1:5, 7, 10))
[1]  1  2  3  4  5  7 10
> mid(x)
[1] 1.5 2.5 3.5 4.5 6.0 8.5
```

mid(x) is a vector that is 1 shorter than the vector x, just as diff(x) is.

So if we want the integral over the period 0 to 5 years, we want the sum over the first 500 intervals, corresponding to the first 501 interval endpoints:

```
> cbind(diff(t), mid(mu))[1:5,]
  [,1]       [,2]
2 0.01 0.06669372
3 0.01 0.06644808
4 0.01 0.06620336
5 0.01 0.06595956
6 0.01 0.06571668
> sum(diff(t[1:501]) * mid(mu[1:501]))
```

∗ Numerical integration with R

In practice we will want the integral **function** of μ, so for every t we want $M(t) = \int_0^t \mu(s)\,\mathrm{d}(s)$. This is easily accomplished by the function `cumsum`:

```
> Mu <- c(0, cumsum(diff(t) * mid(mu)))
> head(cbind(t, Mu))
```

```
     t           Mu
  0.00 0.0000000000
2 0.01 0.0006669372
3 0.02 0.0013314180
4 0.03 0.0019934516
5 0.04 0.0026530472
6 0.05 0.0033102141
```

Note the first value which is the integral from 0 to 0, so by definition 0.

Cumulative risks from parametric models

If we have estimates of λ and μ as functions of time, we can derive the cumulative risks.

In practice this will be by numerical integration; compute the rates at closely spaced intervals and evaluate the integrals as sums. This is easy.

What is not so easy is to come up with confidence intervals for the cumulative risks.

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model.
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates.
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in `Epi`

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in Epi

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in Epi

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in Epi

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in Epi

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in Epi

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in Epi

Simulation of cumulative risks: ci.Crisk

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function ci.Crisk in Epi

Simulation of cumulative risks: `ci.Crisk`

1. a random vector from the multivariate normal distribution with
   - mean equal to the parameters of the model,
   - variance-covariance equal to the estimated variance-covariance of the parameter estimates
2. use this to generate a simulated set of rates $(\lambda(t), \mu(t))$, evaluated a closely spaced times
3. derive state probabilities at these times by numerical integration
4. repeat to obtain, say, 1000 sets of state probabilities at these times
5. derive confidence intervals for the state probabilities as the 2.5 and 97.5 percentiles of the state probabilities at each time

This machinery is implemented in the function `ci.Crisk` in `Epi`

Cumulative risks from parametric models

```
> cR <- ci.Crisk(mods = list(Ins =  Ins.glm,
+                           Dead = Dead.glm),
+                  nd = nd)
```

```
NOTE: Times are assumed to be in the column tfd at equal distances of 0.01
```

```
> str(cR)
```

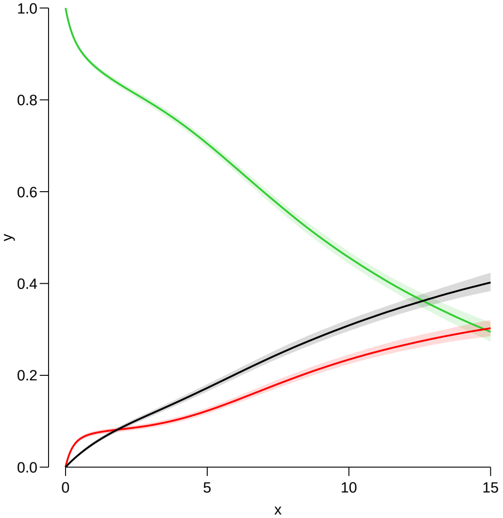
```
List of 4
 $ Crisk: num [1:1501, 1:3, 1:3] 1 0.997 0.993 0.99 0.987 ...
  ..- attr(*, "dimnames")=List of 3
  .. ..$ tfd  : chr [1:1501] "0" "0.01" "0.02" "0.03" ...
  .. ..$ cause: chr [1:3] "Surv" "Ins" "Dead"
  .. ..$      : chr [1:3] "50%" "2.5%" "97.5%"
 $ Srisk: num [1:1501, 1:2, 1:3] 0 0.000666 0.001328 0.001985 0.002637 ...
  ..- attr(*, "dimnames")=List of 3
  .. ..$ tfd  : chr [1:1501] "0" "0.01" "0.02" "0.03" ...
  .. ..$ cause: chr [1:2] "Dead" "Dead+Ins"
  .. ..$      : chr [1:3] "50%" "2.5%" "97.5%"
 $ Stime: num [1:1501, 1:3, 1:3] 0 0.00998 0.01993 0.02985 0.03974 ...
  ..- attr(*, "dimnames")=List of 3
  .. ..$ tfd  : chr [1:1501] "0" "0.01" "0.02" "0.03" ...
  .. ..$ cause: chr [1:3] "Surv" "Ins" "Dead"
```

Cumulative risks from parametric models

So now plot the cumulative **risks** of being in each of the states (the `Crisk` component):

```
> matshade(as.numeric(dimnames(cR$Crisk)[[1]]),
+          cbind(cR$Crisk[,1,],
+                cR$Crisk[,2,],
+                cR$Crisk[,3,]), plot = TRUE,
+          lwd = 2, yaxs = "i", col = c("limegreen","red","black"))
```

Survival and cumulative risk functions



Stacked probabilities: (matrix 2 polygons)

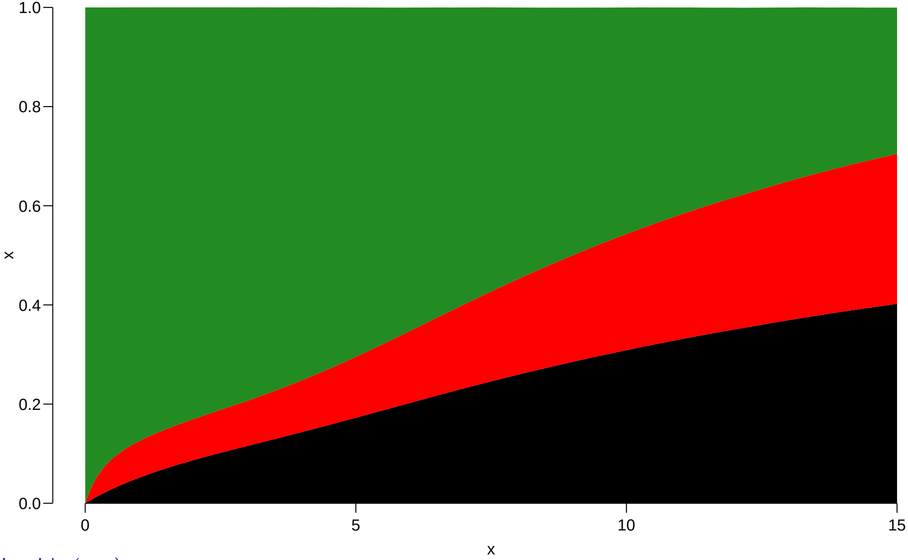
```
> mat2pol(cR$Crisk[,3:1,1], yaxs = "i",
+         col = c("forestgreen","red","black")[3:1])
```

1st argument to mat2pol must be a 2-dimensional matrix, with rows representing the x-axis of the plot, and columns states.

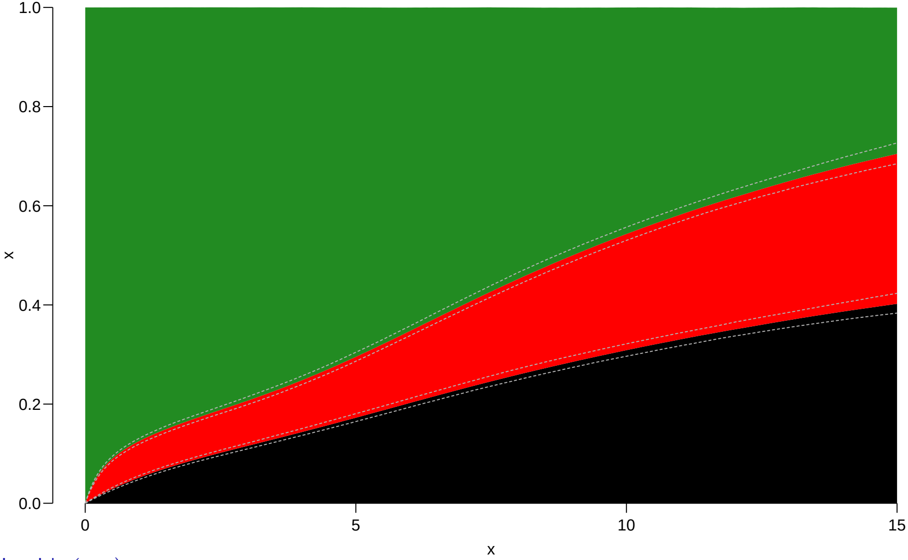
The component Srisk has the confidence limits of the stacked probabilities:

```
> mat2pol(cR$Crisk[,3:1,1], yaxs = "i",
+         col = c("forestgreen","red","black")[3:1])
> matlines(as.numeric(dimnames(cR$Srisk)[[1]]),
+          cbind(cR$Srisk[,"Dead"    ,2:3],
+                cR$Srisk[,"Dead+Ins",2:3]),
+          lty = "32", lwd = 1, col = gray(0.7))
```

Survival and cumulative risk functions



Survival and cumulative risk functions



Expected life time: using simulated objects

The areas between the lines (up to say 10 years) are **expected sojourn times**, that is:

- expected years alive without Ins
- expected years lost to death without Ins
- expected years after Ins, including years dead after Ins

Not all of direct relevance, actually only the first may be so

They are available (with simulation-based confidence intervals) in the component of `cR, Stime (Sojourn time)`.

Exercise 9

Expected life time: using simulated objects

The areas between the lines (up to say 10 years) are **expected sojourn times**, that is:

- expected years alive without Ins
- expected years lost to death without Ins
- expected years after Ins, including years dead after Ins

Not all of direct relevance, actually only the first may be so

They are available (with simulation-based confidence intervals) in the component of `cR, Stime` (`Sojourn time`).

Exercise 9

Expected life time: using simulated objects

The areas between the lines (up to say 10 years) are **expected sojourn times**, that is:

- expected years alive without Ins
- expected years lost to death without Ins
- expected years after Ins, including years dead after Ins

Not all of direct relevance; actually only the first may be so

They are available (with simulation based confidence intervals) in the component of `cR, Stime` (`Sojourn time`)

Exercise 9

Expected life time: using simulated objects

The areas between the lines (up to say 10 years) are **expected sojourn times**, that is:

- expected years alive without Ins
- expected years lost to death without Ins
- expected years after Ins, including years dead after Ins

Not all of direct relevance; actually only the first may be so

They are available (with simulation based confidence intervals) in the component of `cR`, `Stime (Sojourn time)`

Exercise 9

Expected life time: using simulated objects

The areas between the lines (up to say 10 years) are **expected sojourn times**, that is:

- expected years alive without Ins
- expected years lost to death without Ins
- expected years after Ins, including years dead after Ins

Not all of direct relevance; actually only the first may be so.

They are available (with simulation-based confidence intervals) in the component of `cR`, `Stime` (Sojourn `time`).

Exercise 9

Expected life time: using simulated objects

A relevant quantity would be the expected time alive without Ins during the first 5, 10 and 15 years:

```
> str(cR$Stime)
```

```
 num [1:1501, 1:3, 1:3] 0 0.00998 0.01993 0.02985 0.03974 ...
 - attr(*, "dimnames")=List of 3
  ..$ tfd  : chr [1:1501] "0" "0.01" "0.02" "0.03" ...
  ..$ cause: chr [1:3] "Surv" "Ins" "Dead"
  ..$      : chr [1:3] "50%" "2.5%" "97.5%"
```

```
> round(cR$Stime[c("5","10","15"),"Surv",], 1)
```

```
tfd  50% 2.5% 97.5%
  5  4.1  4.0   4.1
  10 7.0  6.9   7.0
  15 8.8  8.7   8.9
```

Exercise 10, 11 (and 12)

RMST

simulation

Survival, mortality,
competing risks and
expected lifetime
EDEG 2025 / Umeå University,17 May 2025

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term "sojourn time" is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term "sojourn time" is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term “sojourn time” is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term “sojourn time” is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes:
  - rate-ratio (M/W HR, typically a function of time)
  - 5 or 10 year survival
  - RMST during the next, say, 10 years for a given age, say, 60
  - Note that RMST refers to an interval, in this case age 60 to 60 + 10

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term “sojourn time” is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes:
  - rate-ratio (M/W HR, typically a function of time)
  - 5 or 10 year survival
  - RMST during the next, say, 10 years for a given age, say, 60
  - Note that RMST refers to an interval, in this case age 60 to 60 + 10

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term “sojourn time” is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes:
  - rate-ratio (M/W HR, typically a function of time)
  - 5 or 10 year survival
  - RMST during the next, say, 10 years for a given age, say, 60
  - Note that RMST refers to an interval, in this case age 60 to 60 + 10

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term “sojourn time” is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes:
  - rate-ratio (M/W HR, typically a function of time)
  - 5 or 10 year survival
  - RMST during the next, say, 10 years for a given age, say, 60
  - Note that RMST refers to an interval, in this case age 60 to 60+10

Comparisons

- RMST — Restricted Mean Survival Time
- a variant of expected lifetime, or more precisely expected residual lifetime as has been available in published life tables for eons
- The term “sojourn time” is also used for the time spent in a given state
- mortality rates among diabetes patients of the two different sexes:
  - rate-ratio (M/W HR, typically a function of time)
  - 5 or 10 year survival
  - RMST during the next, say, 10 years for a given age, say, 60
  - Note that RMST refers to an **interval**, in this case age 60 to $60 + 10$

```
> data(DMlate)
> set.seed(19540803)
> DMlate <- DMlate[sample(1:nrow(DMlate), 1000), ]
> Lx <- Lexis(entry = list(age = dodm - dobth,
+                          tfd = 0),
+               exit = list(tfd = dox - dodm),
+        exit.status = factor(!is.na(dodth), labels = c("DM", "Dead")),
+               data = DMlate)
```

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

```
> sL <- splitLexis(Lx, seq(0, 15, 0.5), "tfd")
> summary(Lx)
```

```
Transitions:
     To
From  DM Dead  Records:  Events: Risk time:  Persons:
  DM 769  231      1000      231    5398.05      1000
```

```
> summary(sL)
```

```
Transitions:
     To
From    DM Dead  Records:  Events: Risk time:  Persons:
  DM 11063  231     11294      231    5398.05      1000
```

proportional hazards model:

```
> m1 <- glmLexis(sL, ~ Ns(age, knots = c(30, 50, 70))
+                   + Ns(tfd, knots = c(0, 1, 4, 10))
+                   + sex)
```

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

```
> round(ci.exp(m1, subset = "sex"), 3)
```

```
     exp(Est.)  2.5% 97.5%
sexF     0.937 0.723 1.215
```

- Women have a mortality about 6% smaller that that of men
- What hazards are proportional here?

proportional hazards model:

```
> m1 <- glmLexis(sL, ~ Ns(age, knots = c(30, 50, 70))
+                    + Ns(tfd, knots = c(0, 1, 4, 10))
+                    + sex)
```

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

```
> round(ci.exp(m1, subset = "sex"), 3)
```

```
     exp(Est.)  2.5% 97.5%
sexF     0.937 0.723 1.215
```

- Women have a mortality about 6% smaller that that of men
- What hazards are proportional here?

Proportional hazards model:

Comparative measures on other possible outcome scales are:

- differences in survival probabilities at certain *times*
- differences in expected life times during certain *time intervals*
- need to specify times and the intervals of interest

Proportional hazards model:

Comparative measures on other possible outcome scales are:

- differences in survival probabilities at certain *times*
- differences in expected life times during certain *time intervals*
- need to specify times and the intervals of interest

Proportional hazards model:

Comparative measures on other possible outcome scales are:

- differences in survival probabilities at certain *times*
- differences in expected life times during certain *time intervals*
- need to specify times and the intervals of interest:
  - at what times since diagnosis do we want comparison of survival between men and women
  - from what time and to what time do we want the expected lifetime computed?
  - for what age (adx: age at diagnosis) do we want the comparison

Proportional hazards model:

Comparative measures on other possible outcome scales are:

- differences in survival probabilities at certain *times*
- differences in expected life times during certain *time intervals*
- need to specify times and the intervals of interest:
  - *at* what times since diagnosis do we want comparison of survival between men and women
  - *from* what time and to what time do we want the expected lifetime computed?
  - for what age (adx: age at diagnosis) do we want the comparison

Proportional hazards model:

Comparative measures on other possible outcome scales are:

- differences in survival probabilities at certain *times*
- differences in expected life times during certain *time intervals*
- need to specify times and the intervals of interest:
  - *at* what times since diagnosis do we want comparison of survival between men and women
  - *from* what time and *to* what time do we want the expected lifetime computed?
  - for what age (adx: age at diagnosis) do we want the comparison

Proportional hazards model:

Comparative measures on other possible outcome scales are:

- differences in survival probabilities at certain *times*
- differences in expected life times during certain *time intervals*
- need to specify times and the intervals of interest:
  - *at* what times since diagnosis do we want comparison of survival between men and women
  - *from* what time and *to* what time do we want the expected lifetime computed?
  - for what age (`adx`, age at diagnosis) do we want the comparison

- compare 5 and 10 year survival
- for men and women
- diagnosed with diabetes at ages 50, 60 and 70

6 survival curves at 150 times, with CI:

```
> surv.arr <- NArray(list(adx = c(50, 60, 70),
+                         sex = c("M", "F"),
+                         tfd = tfd <- seq(0, 15, .1),
+                        surv = c("surv", "lo", "up")))
> str(surv.arr)
 logi [1:3, 1:2, 1:151, 1:3] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ adx : chr [1:3] "50" "60" "70"
  ..$ sex : chr [1:2] "M" "F"
  ..$ tfd : chr [1:151] "0" "0.1" "0.2" "0.3" ...
  ..$ surv: chr [1:3] "surv" "lo" "up"
```

- compare 5 and 10 year survival
- for men and women
- diagnosed with diabetes at ages 50, 60 and 70

6 survival curves at 150 times, with CI:

```
> surv.arr <- NArray(list(adx = c(50, 60, 70),
+                         sex = c("M", "F"),
+                         tfd = tfd <- seq(0, 15, .1),
+                        surv = c("surv", "lo", "up")))
> str(surv.arr)
```

```
 logi [1:3, 1:2, 1:151, 1:3] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ adx : chr [1:3] "50" "60" "70"
  ..$ sex : chr [1:2] "M" "F"
  ..$ tfd : chr [1:151] "0" "0.1" "0.2" "0.3" ...
  ..$ surv: chr [1:3] "surv" "lo" "up"
```

- compare 5 and 10 year survival
- for men and women
- diagnosed with diabetes at ages 50, 60 and 70

6 survival curves at 150 times, with CI:

```
> surv.arr <- NArray(list(adx = c(50, 60, 70),
+                         sex = c("M", "F"),
+                         tfd = tfd <- seq(0, 15, .1),
+                        surv = c("surv", "lo", "up")))
> str(surv.arr)
 logi [1:3, 1:2, 1:151, 1:3] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ adx : chr [1:3] "50" "60" "70"
  ..$ sex : chr [1:2] "M" "F"
  ..$ tfd : chr [1:151] "0" "0.1" "0.2" "0.3" ...
  ..$ surv: chr [1:3] "surv" "lo" "up"
```

Survival at 5 and 10 years

```
> for(adx in c(50, 60, 70))
+ for( sx in c("M", "F"))
+    {
+    nd <- data.frame(tfd = tfd,
+                     age = adx + tfd,
+                     sex = sx)
+    surv.arr[paste(adx), sx, , ] <- ci.surv(m1, nd)
+    }
NOTE: interval length chosen from  as tfd[2] - tfd[1]
NOTE: interval length chosen from  as tfd[2] - tfd[1]
NOTE: interval length chosen from  as tfd[2] - tfd[1]
NOTE: interval length chosen from  as tfd[2] - tfd[1]
NOTE: interval length chosen from  as tfd[2] - tfd[1]
NOTE: interval length chosen from  as tfd[2] - tfd[1]
```

Survival at 5 and 10 years

```
> round(ftable(surv.arr[,,c("5","10"),] * 100, row.vars = c(1,3)), 1)
        sex     M              F
        surv surv   lo   up surv   lo   up
adx tfd
50  5        96.0 97.2 94.2 96.2 97.4 94.4
    10       90.8 93.3 87.4 91.3 93.8 87.9
60  5        89.7 92.1 86.7 90.3 92.7 87.2
    10       77.6 82.2 72.0 78.8 83.5 73.1
70  5        75.3 79.4 70.5 76.7 80.8 71.8
    10       51.5 58.2 44.3 53.7 60.4 46.5
> # round(ftable(surv.arr[,,c("5","10"),] * 100, row.vars = c(3,1,2)), 1)
```

Exercises 14 & 15

RMST

Use `ci.Crisk` to get estimates of RMST

```
> head(nd)
  tfd  age sex
1 0.0 70.0   F
2 0.1 70.1   F
3 0.2 70.2   F
4 0.3 70.3   F
5 0.4 70.4   F
6 0.5 70.5   F
> msM <- ci.Crisk(list(Mort = m1), mutate(nd, sex = "M"))$Stime
NOTE: Times are assumed to be in the column tfd at equal distances of 0.1
> msF <- ci.Crisk(list(Mort = m1), mutate(nd, sex = "F"))$Stime
NOTE: Times are assumed to be in the column tfd at equal distances of 0.1
> str(msF)
 num [1:151, 1:2, 1:3] 0 0.0997 0.199 0.2977 0.396 ...
 - attr(*, "dimnames")=List of 3
  ..$ tfd  : chr [1:151] "0" "0.1" "0.2" "0.3" ...
```

RMST confidence intervals

We can get confidence intervals from (parametric) bootstrap samples of the cumulative rates.

This is done by simulation from the distribution of the model parameters.

Again an array to store the simulated cumulative risks:

```
> nB <- 10000 # no of bootstrap samples
> ain <- 5:7 * 10 # baseline ages
> sex <- c("M", "F")
> simres <- NArray(list(adx = ain,
+                       sex = sex,
+                       tfd = nd$tfd,
+                       sim = 1:nB))
> str(simres)
 logi [1:3, 1:2, 1:151, 1:10000] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ adx: chr [1:3] "50" "60" "70"
```

RMST confidence intervals for differences

Comparing M and F requires
the same stream of simulated parameters for different predictions:
reset random seed inside loop

```
> for (adx in ain)
+ for ( sx in sex)
+     {
+ set.seed(20250503)
+ simres[paste(adx), sx, , ] <- ci.Crisk(list(Mort = m1),
+                                         nd = mutate(nd, sex = sx,
+                                                         age = adx + tfd),
+                                         nB = nB,
+                                         sim.res = "crisk")[, "Surv", ]
+     }
```

Exercises 16 & 17

RMST confidence intervals for differences

Comparing M and F requires
the same stream of simulated parameters for different predictions:
reset random seed inside loop

```
> for (adx in ain)
+ for ( sx in sex)
+     {
+ set.seed(20250503)
+ simres[paste(adx), sx, , ] <- ci.Crisk(list(Mort = m1),
+                                        nd = mutate(nd, sex = sx,
+                                                        age = adx + tfd),
+                                        nB = nB,
+                                        sim.res = "crisk")[, "Surv", ]
+     }
```

Exercises 16 & 17

Further exercises

- Exercise 18 Predicted mortality from PH model
- Exercise 19 Interaction model (non-PH)
- Exercise 20 M to F differences
- Exercise 21 Age differences in RMST
- Exercise 22 Overview of RMST

Further exercises

- Exercise 18 Predicted mortality from PH model
- Exercise 19 Interaction model (non-PH)
- Exercise 20 M to F differences
- Exercise 21 Age differences in RMST
- Exercise 22 Overview of RMST

Further exercises

- Exercise 18 Predicted mortality from PH model
- Exercise 19 Interaction model (non-PH)
- Exercise 20 M to F differences
- Exercise 21 Age differences in RMST
- Exercise 22 Overview of RMST

Further exercises

- Exercise 18 Predicted mortality from PH model
- Exercise 19 Interaction model (non-PH)
- Exercise 20 M to F differences
- Exercise 21 Age differences in RMST
- Exercise 22 Overview of RMST

Further exercises

- Exercise 18 Predicted mortality from PH model
- Exercise 19 Interaction model (non-PH)
- Exercise 20 M to F differences
- Exercise 21 Age differences in RMST
- Exercise 22 Overview of RMST

Multistate model

simulation

Survival, mortality,
competing risks and
expected lifetime
EDEG 2025 / Umeå University,17 May 2025

BAckground: Steno 2 trial

- Clinical trial for diabetes ptt. with kidney disease (micro-albuminuria)
- 80 ptt. randomised to either of
- 1993–2001
- follow-up till 2018

BAckground: Steno 2 trial

- Clinical trial for diabetes ptt. with kidney disease (micro-albuminuria)
- 80 ptt. randomised to either of
  - Conventional treatment
  - Intensified multifactorial treament
- 1993–2001
- follow-up till 2018

BAckground: Steno 2 trial

- Clinical trial for diabetes ptt. with kidney disease (micro-albuminuria)
- 80 ptt. randomised to either of
  - Conventional treatment
  - Intensified multifactorial treatment
- 1993–2001
- follow-up till 2018

BAckground: Steno 2 trial

- Clinical trial for diabetes ptt. with kidney disease (micro-albuminuria)
- 80 ptt. randomised to either of
  - Conventional treatment
  - Intensified multifactorial treament
- 1993–2001
- follow-up till 2018

BAckground: Steno 2 trial

- Clinical trial for diabetes ptt. with kidney disease (micro-albuminuria)
- 80 ptt. randomised to either of
  - Conventional treatment
  - Intensified multifactorial treament
- 1993–2001
- follow-up till 2018

BAckground: Steno 2 trial

- Clinical trial for diabetes ptt. with kidney disease (micro-albuminuria)
- 80 ptt. randomised to either of
  - Conventional treatment
  - Intensified multifactorial treament
- 1993–2001
- follow-up till 2018

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on
- Quantification of treatment effect

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on
- Quantification of treatment effect

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on
- Quantification of treatment effect

Steno 2 trial: goal

- Is there a treatment effect on:
    - CVD mortality
    - non-CVD mortality
- Does the treatment effect depend on:
    - Albuminuria state
- Quantification of treatment effect

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on:
  - Albuminuria state
- Quantification of treatment effect

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on:
  - Albuminuria state
- Quantification of treatment effect:
  - Rate-ratios
  - Life times
  - Changes in clinical parameters

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on:
  - Albuminuria state
- Quantification of treatment effect:
  - Rate-ratios
  - Life-times
  - Changes in clinical parameters

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on:
  - Albuminuria state
- Quantification of treatment effect:
  - Rate-ratios
  - Life times
  - Changes in clinical parameters

Steno 2 trial: goal

- Is there a treatment effect on:
  - CVD mortality
  - non-CVD mortality
- Does the treatment effect depend on:
  - Albuminuria state
- Quantification of treatment effect:
  - Rate-ratios
  - Life times
  - Changes in clinical parameters

```
> data(steno2)
> steno2 <- cal.yr(steno2)
> steno2 <- transform(steno2,
+                     doEnd = pmin(doDth, doEnd, na.rm = TRUE))
> str(steno2)
```

```
'data.frame':        160 obs. of  14 variables:
 $ id      : num  1 2 3 4 5 6 7 8 9 10 ...
 $ allo    : Factor w/ 2 levels "Int","Conv": 1 1 2 2 2 2 2 1 1 1 ...
 $ sex     : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 2 2 2 ...
 $ baseCVD : num  0 0 0 0 0 1 0 0 0 0 ...
 $ deathCVD: num  0 0 0 0 1 0 0 0 1 0 ...
 $ doBth   : 'cal.yr' num  1932 1947 1943 1945 1936 ...
 $ doDM    : 'cal.yr' num  1991 1982 1983 1977 1986 ...
 $ doBase  : 'cal.yr' num  1993 1993 1993 1993 1993 ...
 $ doCVD1  : 'cal.yr' num  2014 2009 2002 1995 1994 ...
 $ doCVD2  : 'cal.yr' num  NA 2009 NA 1997 1995 ...
 $ doCVD3  : 'cal.yr' num  NA 2010 NA 2003 1998 ...
 $ doESRD  : 'cal.yr' num  NaN NaN NaN NaN 1998 ...
 $ doEnd   : 'cal.yr' num  2015 2015 2002 2003 1998 ...
 $ doDth   : 'cal.yr' num  NA NA 2002 2003 1998 ...
```

A Lexis object

```
> L2 <- Lexis(entry = list(per = doBase,
+                          age = doBase - doBth,
+                          tfi = 0),
+               exit = list(per = doEnd),
+        exit.status = factor(deathCVD + !is.na(doDth),
+                             labels=c("Mic","D(oth)","D(CVD)")),
+                 id = id,
+               data = steno2)
NOTE: entry.status has been set to "Mic" for all.
```

Explain the coding of `exit.status`.

A Lexis object

```
> summary(L2, t = TRUE)
Transitions:
     To
From  Mic D(oth) D(CVD)  Records:  Events: Risk time:  Persons:
  Mic  67     55     38       160       93    2416.59       160

Timescales:
per age tfi
 ""  ""  ""
```

How many persons are there in the cohort?

How many deaths are there in the cohort?

How much follow-up time is there in the cohort?

How many states are there in the model (so far)?

Albuminuria status

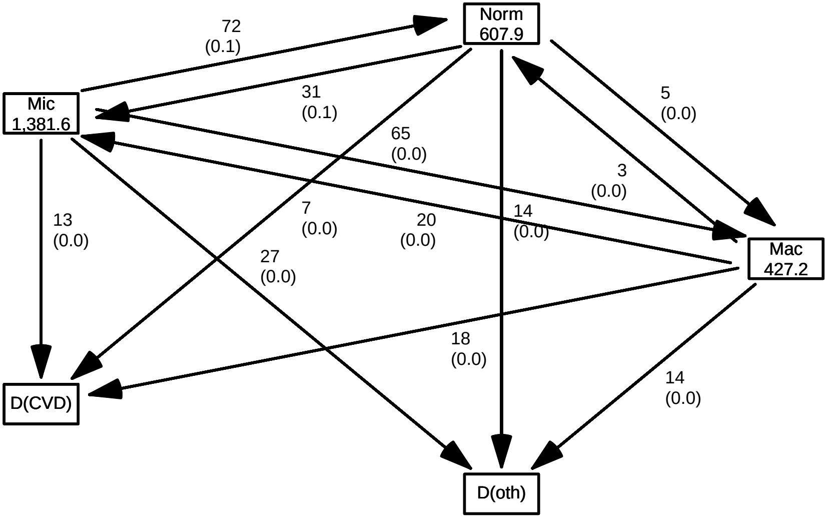
```
> data(st2alb) ; head(st2alb, 3)

  id       doTr state
1  1 1993-06-12   Mic
2  1 1995-05-13  Norm
3  1 2000-01-26   Mic

> cut2 <- rename(cal.yr(st2alb),
+                  lex.id = id,
+                     cut = doTr,
+               new.state = state)
> with(cut2, addmargins(table(table(lex.id))))

  1   2   3   4   5 Sum
  4  25  40  46  41 156
```

What does this table mean?

Albuminuria status as states

```
> L3 <- rcutLexis(L2, cut2, time = "per")
> summary(L3)

Transitions:
     To
From   Mic Norm Mac D(oth) D(CVD)  Records:  Events: Risk time:  Persons:
  Mic  299   72  65     27     13       476      177    1381.57       160
  Norm  31   90   5     14      7       147       57     607.86        69
  Mac   20    3  44     14     18        99       55     427.16        64
  Sum  350  165 114     55     38       722      289    2416.59       160

> boxes(L3, boxpos = TRUE, cex = 0.8)
```

What's wrong with this



What's in `jump`

```
> (jump <-
+ subset(L3, (lex.Cst == "Norm" & lex.Xst == "Mac") |
+            (lex.Xst == "Norm" & lex.Cst == "Mac"))[,
+         c("lex.id", "per", "lex.dur","lex.Cst", "lex.Xst")])
 lex.id     per lex.dur lex.Cst lex.Xst
     70 1999.49    2.67     Mac    Norm
     86 2001.76   12.82    Norm     Mac
    130 2000.91    1.88     Mac    Norm
    131 1997.76    4.24    Norm     Mac
    136 1997.21    0.47     Mac    Norm
    136 1997.69    4.24    Norm     Mac
    171 1996.39    5.34    Norm     Mac
    175 2004.58    9.88    Norm     Mac
```

—and what will you do about it?

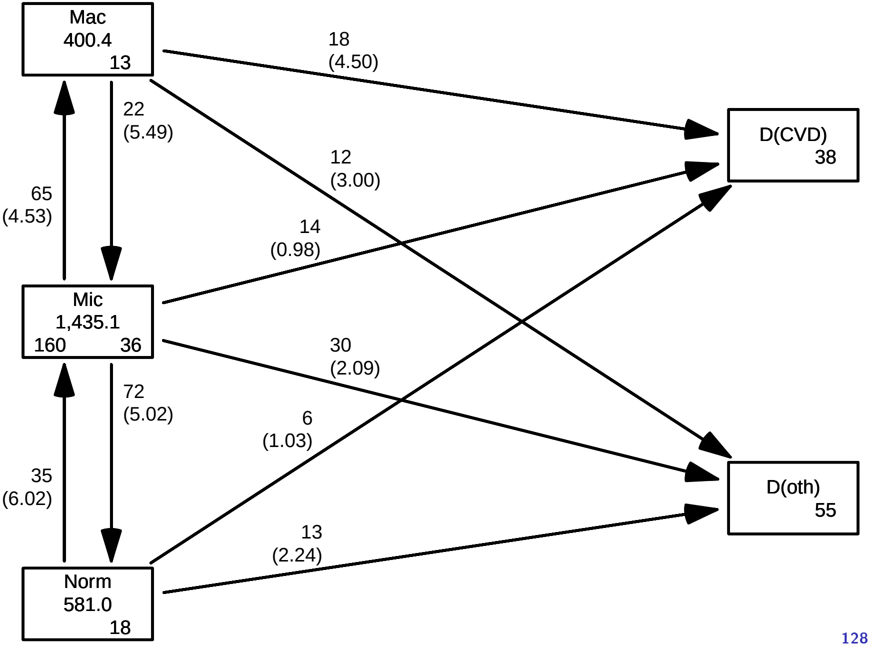
How to fix things

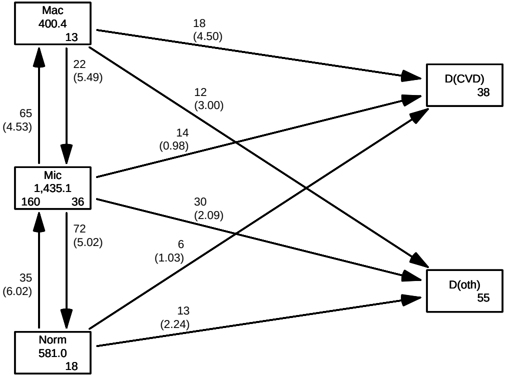
```
> set.seed(1952)
> xcut <- transform(jump,
+                    cut = per + lex.dur * runif(per, 0.1, 0.9),
+              new.state = "Mic")
> xcut <- select(xcut, c(lex.id, cut, new.state))
> L4 <- rcutLexis(L3, xcut)
> L4 <- Relevel(L4, c("Norm","Mic","Mac","D(CVD)","D(oth)"))
> summary(L4)
```

```
Transitions:
     To
From   Norm Mic Mac D(CVD) D(oth)  Records:  Events: Risk time:  Persons:
  Norm   90  35   0      6     13       144       54     581.04        66
  Mic    72 312  65     14     30       493      181    1435.14       160
  Mac     0  22  41     18     12        93       52     400.41        60
  Sum   162 369 106     38     55       730      287    2416.59       160
```

Plot the boxes

```
> boxes(L4, boxpos = list(x = c(20, 20, 20, 80, 80),
+                         y = c(10, 50, 90, 75, 25)),
+            show.BE = "nz",
+            scale.R = 100, digits.R = 2,
+            cex = 0.9, pos.arr = 0.3)
```


Mac
400.4
13
Mic
1,435.1
160
36
Norm
581.0
18
D(CVD)
38
D(oth)
55
18
(4.50)
22
(5.49)
12
(3.00)
65
(4.53)
14
(0.98)
30
(2.09)
72
(5.02)
6
(1.03)
35
(6.02)
13
(2.24)



Explain all the numbers in the graph.

Describe the overall effect of albuminuria on the two mortality rates.

Modeling transition rates

- A model with a smooth effect of timescales on the rates require follow-up in small bits
- Achieved by `splitLexis` (or `splitMulti` from `popEpi`)
- Compare the `Lexis` objects

Modeling transition rates

- A model with a smooth effect of timescales on the rates require follow-up in small bits
- Achieved by `splitLexis` (or `splitMulti` from `popEpi`)
- Compare the `Lexis` objects

Modeling transition rates

- A model with a smooth effect of timescales on the rates require follow-up in small bits
- Achieved by `splitLexis` (or `splitMulti` from `popEpi`)
- Compare the `Lexis` objects

```
> S4 <- splitMulti(L4, tfi = seq(0, 25, 1/2))
> summary(L4)

Transitions:
     To
From   Norm Mic Mac D(CVD) D(oth)  Records:  Events: Risk time:  Persons:
  Norm   90  35   0      6     13       144       54     581.04        66
  Mic    72 312  65     14     30       493      181    1435.14       160
  Mac     0  22  41     18     12        93       52     400.41        60
  Sum   162 369 106     38     55       730      287    2416.59       160

> summary(S4)

Transitions:
     To
From   Norm  Mic Mac D(CVD) D(oth)  Records:  Events: Risk time:  Persons:
  Norm 1252   35   0      6     13      1306       54     581.04        66
  Mic    72 3101  65     14     30      3282      181    1435.14       160
  Mac     0   22 844     18     12       896       52     400.41        60
  Sum  1324 3158 909     38     55      5484      287    2416.59       160
```

How the split works:

```
> subset(L4, lex.id == 96)[,1:7]
```

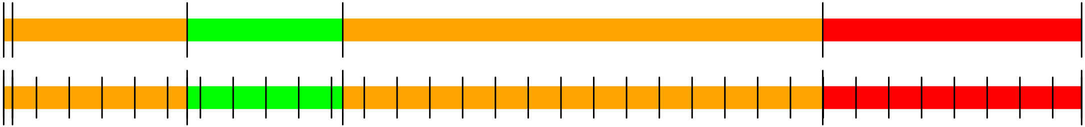
```
 lex.id     per   age  tfi lex.dur lex.Cst lex.Xst
     96 1993.65 51.53 0.00    0.45     Mic    Norm
     96 1994.10 51.99 0.45    2.58    Norm    Norm
     96 1996.68 54.57 3.03    1.90    Norm    Norm
     96 1998.59 56.47 4.94    2.90    Norm  D(CVD)
```

```
> s4 <- subset(S4, lex.id == 96)[,1:7]
> s4[c(1:4,NA,nrow(s4)+(-3:0)),]
```

```
 lex.id     per   age  tfi lex.dur lex.Cst lex.Xst
     96 1993.65 51.53 0.00    0.45     Mic    Norm
     96 1994.10 51.99 0.45    0.05    Norm    Norm
     96 1994.15 52.03 0.50    0.50    Norm    Norm
     96 1994.65 52.53 1.00    0.50    Norm    Norm
     NA      NA    NA   NA      NA    <NA>    <NA>
     96 1999.65 57.53 6.00    0.50    Norm    Norm
     96 2000.15 58.03 6.50    0.50    Norm    Norm
     96 2000.65 58.53 7.00    0.50    Norm    Norm
     96 2001.15 59.03 7.50    0.33    Norm  D(CVD)
```

```
> subset(L4, lex.id == 159)[,1:7]

 lex.id     per   age   tfi lex.dur lex.Cst lex.Xst
    159 1994.02 67.50  0.00    0.13     Mic     Mic
    159 1994.16 67.63  0.13    2.66     Mic    Norm
    159 1996.82 70.29  2.80    2.37    Norm     Mic
    159 1999.20 72.67  5.17    7.32     Mic     Mac
    159 2006.52 79.99 12.49    3.95     Mac  D(CVD)

> subset(S4, lex.id == 159)[c(1:2,NA,6:7,NA,12:13,NA,27:28,NA,36:37),1:7]

 lex.id     per   age   tfi lex.dur lex.Cst lex.Xst
    159 1994.02 67.50  0.00    0.13     Mic     Mic
    159 1994.16 67.63  0.13    0.37     Mic     Mic
     NA      NA    NA    NA      NA    <NA>    <NA>
    159 1996.02 69.50  2.00    0.50     Mic     Mic
    159 1996.52 70.00  2.50    0.30     Mic    Norm
     NA      NA    NA    NA      NA    <NA>    <NA>
    159 1998.52 72.00  4.50    0.50    Norm    Norm
    159 1999.02 72.50  5.00    0.17    Norm     Mic
     NA      NA    NA    NA      NA    <NA>    <NA>
    159 2005.52 79.00 11.50    0.50     Mic     Mic
    159 2006.02 79.50 12.00    0.49     Mic     Mac
     NA      NA    NA    NA      NA    <NA>    <NA>
    159 2009.52 83.00 15.50    0.50     Mac     Mac
    159 2010.02 83.50 16.00    0.44     Mac  D(CVD)
```

How the split works



Same amount of follow-up

Same transitions

More intervals (5, resp. 37)

Different value of time scales between intervals

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate
  from a given origin state (`lex.Cst`)
  to a given destination state (`lex.Xst`)
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate from a given origin state (`lex.Cst`) to a given destination state (`lex.Xst`)
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate
  from a given origin state (lex.Cst)
  to a given destination state (lex.Xst)
- looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate from a given origin state (`lex.Cst`) to a given destination state (`lex.Xst`)
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate from a given origin state (`lex.Cst`) to a given destination state (`lex.Xst`).
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate
  from a given origin state (`lex.Cst`)
  to a given destination state (`lex.Xst`)
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate from a given origin state (`lex.Cst`) to a given destination state (`lex.Xst`).
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate
  `from` a given origin state (`lex.Cst`)
  `to` a given destination state (`lex.Xst`).
- —looks as the likelihood for a single Poisson observation

Purpose of the split

- Assumption of constant rate in each interval
- All intervals are (shorter than) 0.5 years
- Magnitude of the rates depend on covariates:
  - fixed covariates
  - time scales
  - randomly varying covariates (not now)
- values of covariates differ between intervals
- each interval contributes to the (log-)likelihood for a specific rate `from` a given origin state (`lex.Cst`)
  `to` a given destination state (`lex.Xst`).
- —looks as the likelihood for a single Poisson observation

Modeling the rate: `Mic -> D(CVD)`

```
> mr <- glm(cbind(lex.Xst == "D(CVD)" & lex.Cst != lex.Xst,
+                 lex.dur)
+             ~ Ns(tfi, knots = seq( 0, 20, 5)) +
+               Ns(age, knots = seq(50, 80, 10)),
+             family = poisreg,
+               data = subset(S4, lex.Cst == "Mic"))
```

. . . the same as:

```
> mp <- glm((lex.Xst == "D(CVD)" & lex.Cst != lex.Xst)
+             ~ Ns(tfi, knots = seq( 0, 20, 5)) +
+               Ns(age, knots = seq(50, 80, 10)),
+             offset = log(lex.dur),
+             family = poisson,
+               data = subset(S4, lex.Cst == "Mic"))
> summary(coef(mr) - coef(mp))
```

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.296e-12 -2.295e-13 -2.509e-14 -1.521e-13 -6.745e-15  6.697e-13
```

Modeling the rate: Mic -> D(CVD)

A convenient wrapper for Lexis objects simplifies things substantially:

```
> mL <- glm.Lexis(S4, ~ Ns(tfi, knots = seq( 0, 20, 5)) +
+                       Ns(age, knots = seq(50, 80, 10)),
+                 from = "Mic",
+                   to = "D(CVD)")
```

```
stats::glm Poisson analysis of Lexis object S4 with log link:
Rates for the transition:
Mic->D(CVD)
```

```
> summary(coef(mr) - coef(mL))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0       0       0       0       0       0 
```

```
> summary(coef(mp) - coef(mL))
```

```
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
-6.697e-13  6.745e-15  2.509e-14  1.521e-13  2.295e-13  1.296e-12 
```

`glm.Lexis` by default models all transitions `to` absorbing states, `from` states preceding these

```
> mX <- glm.Lexis(S4, ~ Ns(tfi, knots = seq( 0, 20, 5)) +
+                       Ns(age, knots = seq(50, 80, 10)) +
+                       lex.Cst)
```

```
NOTE:
Multiple transitions *from* state ' Mac', 'Mic', 'Norm ' - are you sure?
The analysis requested is effectively merging outcome states.
You may want analyses using a *stacked* dataset - see ?stack.Lexis
stats::glm Poisson analysis of Lexis object S4 with log link:
Rates for transitions:
Norm->D(CVD)
Mic->D(CVD)
Mac->D(CVD)
Norm->D(oth)
Mic->D(oth)
Mac->D(oth)
```

Describe the model(s) in mX (look at the figure with the boxes)

- What rates are modeled ?
- How are they modeled (assumptions about shapes) ?
- What are the differences between the rates modeled?
- What would you rather do?

Describe the model(s) in mX (look at the figure with the boxes)

- What rates are modeled ?
- How are they modeled (assumptions about shapes) ?
- What are the differences between the rates modeled ?
- What would you conclude ?

Describe the model(s) in mX (look at the figure with the boxes)

- What rates are modeled ?
- How are they modeled (assumptions about shapes) ?
- What are the differences between the rates modeled?
- What would you rather do?

Describe the model(s) in `mX` (look at the figure with the boxes)

- What rates are modeled ?
- How are they modeled (assumptions about shapes) ?
- What are the differences between the rates modeled?
- What would you rather do?