Multistate models with `mstate`

STK4080 H16

1. Review with some R-examples
2. General event histories and

Aalen-Johansen estimator using the `mstate` library

Survival data

We have considered right censored and possibly left truncated survival data $(V_i, \tilde{T}_i, D_i$, in particular

- Nelson-Aalen estimator of cumulative hazard $A(t) = \int_0^t \alpha(s)ds$
- Kaplan-Meier estimator of survival function $S(t) = \exp(-A(t)) = \mathrm{P}(T_i > t)$
- Non-parametric tests for differences of hazards (distributions) $\alpha_1(s) = \alpha_2(s) = \cdots = \alpha_K(s)$ in groups $k = 1, \ldots, K$.

In the following slides these methods are applied to a data set on death from melanoma.

Melanoma data

Short description:
In the period 1962-77 a total of 205 patients with malignant melanoma (cancer of the skin) were operated at Odense University hospital in Denmark. A number of covariates were recorded at operation, and the patients were followed up until death or censoring at the end of the study at December 31, 1977.

Coding of variables:

1) status: 1=death from disease, 2=censored, 4=death from other cause
2) lifetime: life time from operation in years
3) ulcer: ulceration (1=present, 2=absent)
4) thickn: tumor thickness in mm
5) sex: 1=female, 2=male
6) age: age at operation in years
7) grthick: grouped tumor thickness (1: 0-1 mm, 2: 2-5 mm, 3: 5+ mm)
8) logthick: logarithm of tumor thickness

Reading the data i R

```
> library(survival)
> path="http://www.uio.no/studier/emner/matnat/math/STK4080/h14
                                                    /melanoma.txt"
> mel=read.table(path,header=T)
> dim(mel)
[1] 205   8
> summary(mel)
     status          lifetime          ulcer           thickn
 Min.   :1.000   Min.   : 0.0274   Min.   :1.000   Min.   : 0.10
 1st Qu.:1.000   1st Qu.: 4.1781   1st Qu.:1.000   1st Qu.: 0.97
 Median :2.000   Median : 5.4932   Median :2.000   Median : 1.94
 Mean   :1.859   Mean   : 5.8981   Mean   :1.561   Mean   : 2.92
 3rd Qu.:2.000   3rd Qu.: 8.3342   3rd Qu.:2.000   3rd Qu.: 3.56
 Max.   :4.000   Max.   :15.2466   Max.   :2.000   Max.   :17.42
      sex              age            grthick         logthick
 Min.   :1.000   Min.   : 4.00   Min.   :1.000   Min.   :-2.30259
 1st Qu.:1.000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:-0.03046
 Median :1.000   Median :54.00   Median :1.000   Median : 0.66269
 Mean   :1.385   Mean   :52.46   Mean   :1.624   Mean   : 0.61817
 3rd Qu.:2.000   3rd Qu.:65.00   3rd Qu.:2.000   3rd Qu.: 1.26976
 Max.   :2.000   Max.   :95.00   Max.   :3.000   Max.   : 2.85762
```

Nelson-Aalen estimator can be obtained

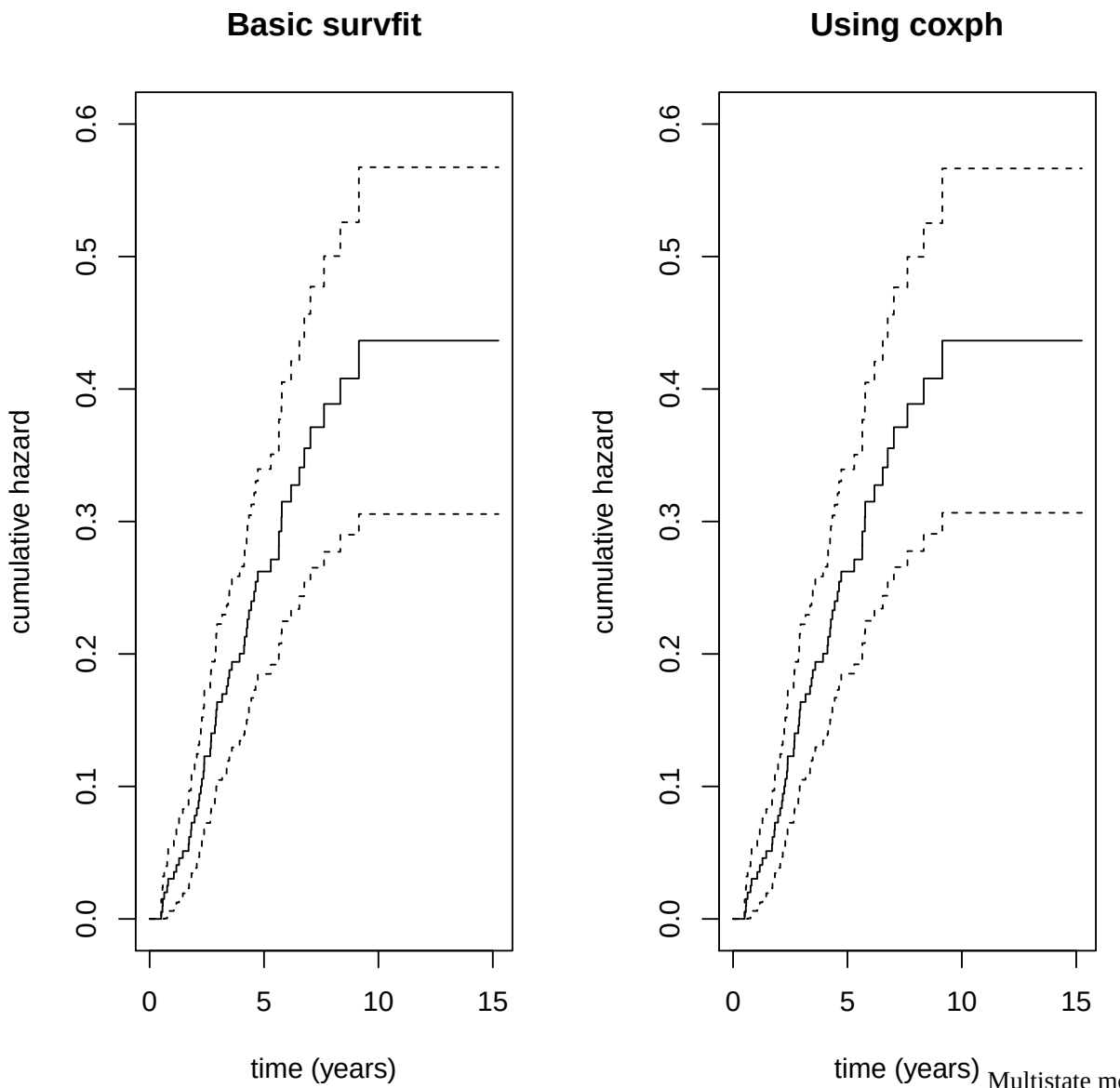
- using `survfit` directly
- going through a formal cox-regression

Result will be the same, R commands:

```
> par(mfrow=c(1,2))
> surv1=survfit(Surv(lifetime,status==1)~1,data=mel,type="fh")
> plot(surv1,fun="cumhaz",ylim=c(0,0.6),xlab="time (years)",
                  ylab="cumulative hazard", main="Basic survfit")
> surv2=survfit(coxph(Surv(lifetime,status==1)~1,method="breslow",
                                                    data=mel))
> plot(surv2,fun="cumhaz",ylim=c(0,0.6),xlab="time (years)",
                      ylab="cumulative hazard",main="Using coxph")
```

The Nelson-Aalen plots

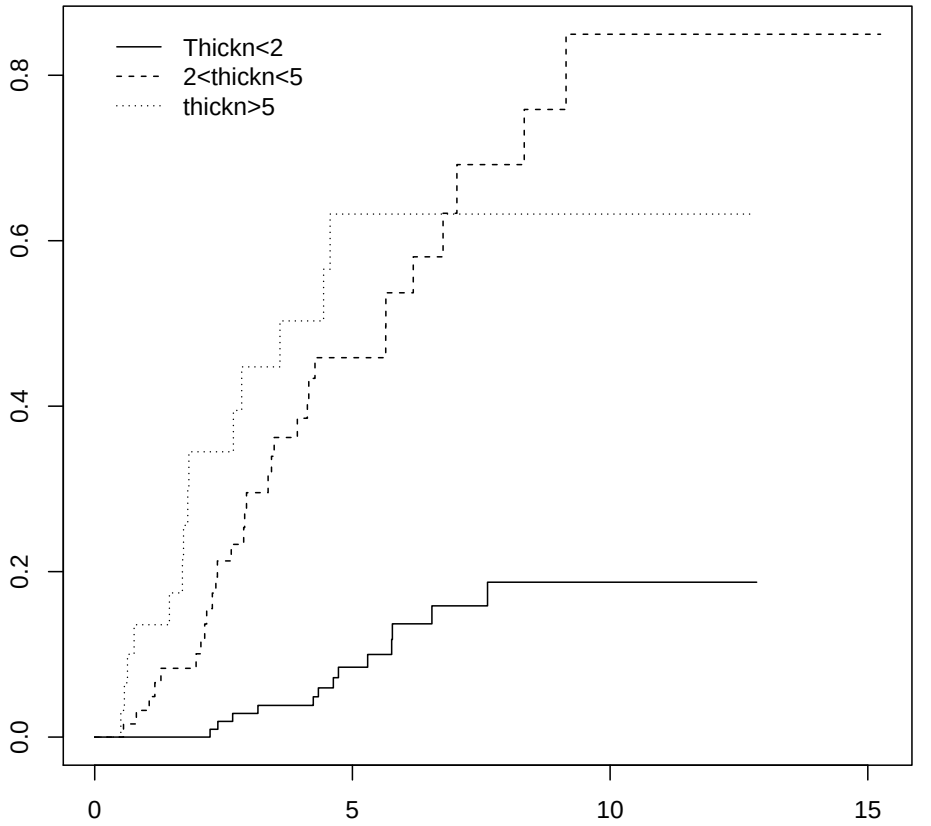
The commands will produce the identical plots:



Comparing groups graphically

We can compare groups by plottting survival or cumulative hazards by the groups

```
plot(survfit(Surv(lifetime,status==1)~grthick,data=mel,type="fh"),
                                                fun="cumhaz",lty=1:3)
legend(0,0.87,lty=1:3,c("Thickn<2","2<thickn<5","thickn>5"),bty="n")
```



Comparing groups formally

by hypothesis testing (e.g. log-rank test):

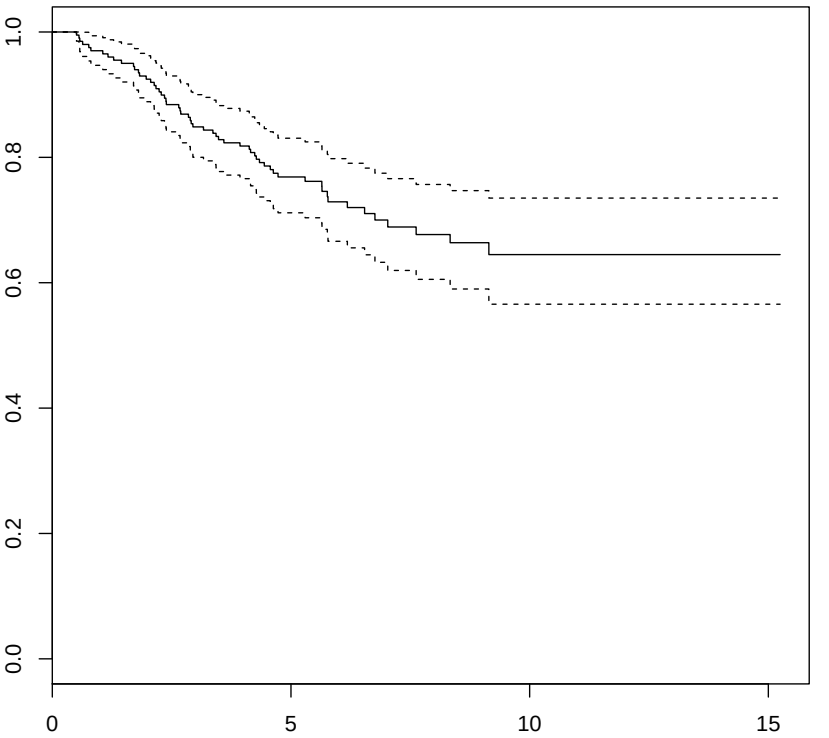
```
> survdiff(Surv(lifetime,status==1)~grthick,data=mel)
Call:
survdiff(formula = Surv(lifetime, status == 1) ~ grthick, data = mel)

             N Observed Expected (O-E)^2/E (O-E)^2/V
grthick=1 109       13    33.75     12.75     31.36
grthick=2  64       30    16.39     11.30     15.88
grthick=3  32       14     6.86      7.42      8.45

 Chisq= 31.6  on 2 degrees of freedom, p= 1.39e-07
```

Estimation of survival: Kaplan-Meier

```
plot(survfit(Surv(lifetime,status==1)~1,data=mel))
```

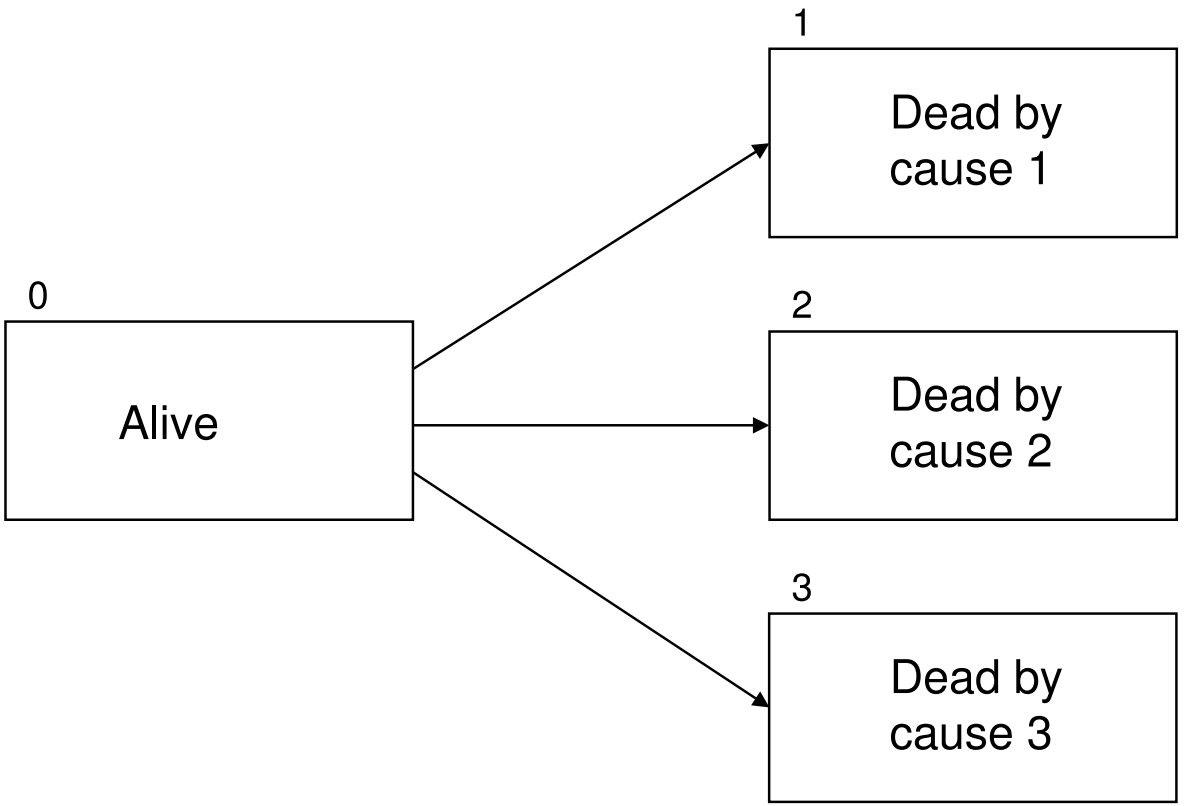


But wait a sec - the outcome here is death by melanoma,

so the (optimistic) interpretation of the curve is

"probability of alive if melanoma is the only cause of death"

Competing risks



Denotes the state "alive" by 0 and state "death from cause h" by $h(= 1, 2, \ldots, k)$

- Transition (hazard) rate $\alpha_{0h}(t)$ of cause h
- k causes of death (can only observe one cause)
- Will call $\alpha_{0h}(t)$ cause-specific hazard rates

Competing risk melanoma data

```
> table(mel$status)
  1   2   4
 57 134  14
```

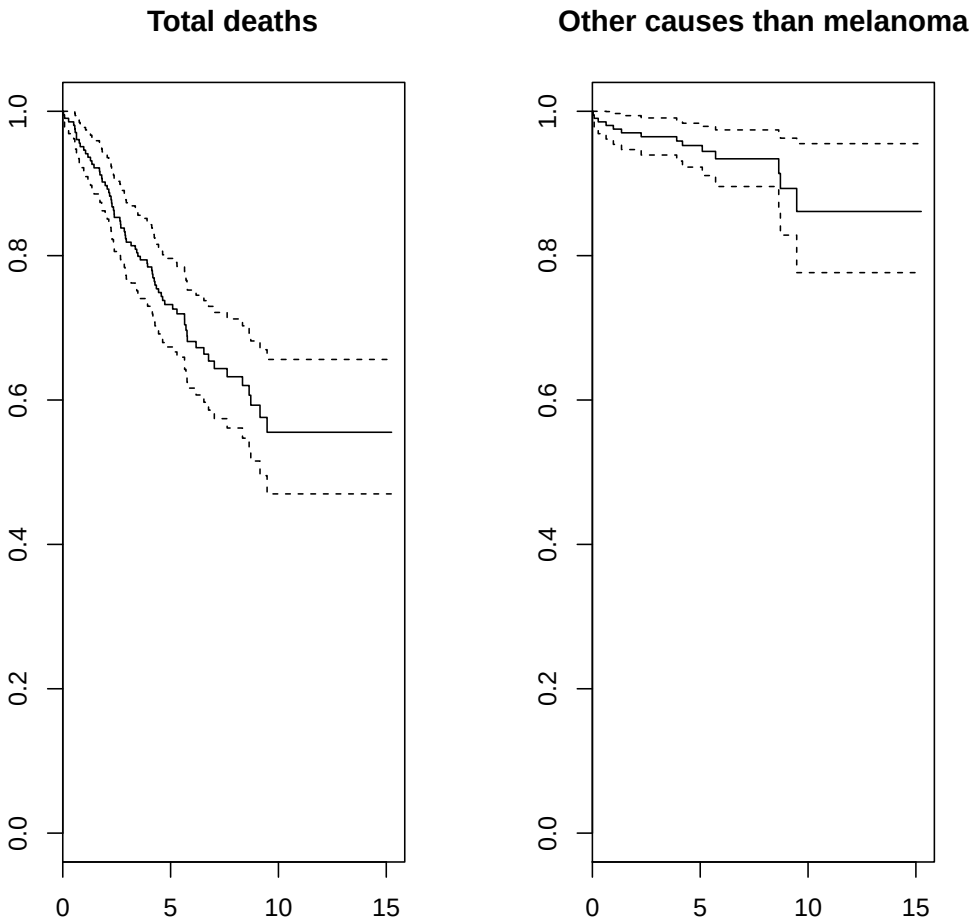
Thus there are 57 melanoma deaths and 14 deaths of other causes.

We could deal with this by

- Calculating total mortality
- Estimating mortality due to other causes.
  This second option would optimistically be interpreted as probability of alive given no melanoma mortality
- Estimating cumulative incidences $\mathrm{P}_{0h}(0,t)$ = probability of death by cause h before time t

All and other causes

```
> mel$dead=1*(mel$status!=2)
> par(mfrow=c(1,2))
> plot(survfit(Surv(lifetime,dead==1)~1,data=mel),main="Total deaths")
> plot(survfit(Surv(lifetime,dead==4)~1,data=mel),
                    main="Other causes than melanoma")
```



Cumulative incidence functions

The transition probabilities for competing risks are given as

$$P_{0h}(s,t) = P(X(t) = h | X(s) = 0)$$

can be estimated

$$\widehat{\mathrm{P}}_{00}(s,t) = \prod_{s<u\leq t} \left[1 - \frac{dN_0(u)}{Y_0(u)}\right],$$

for $h = 0$ and

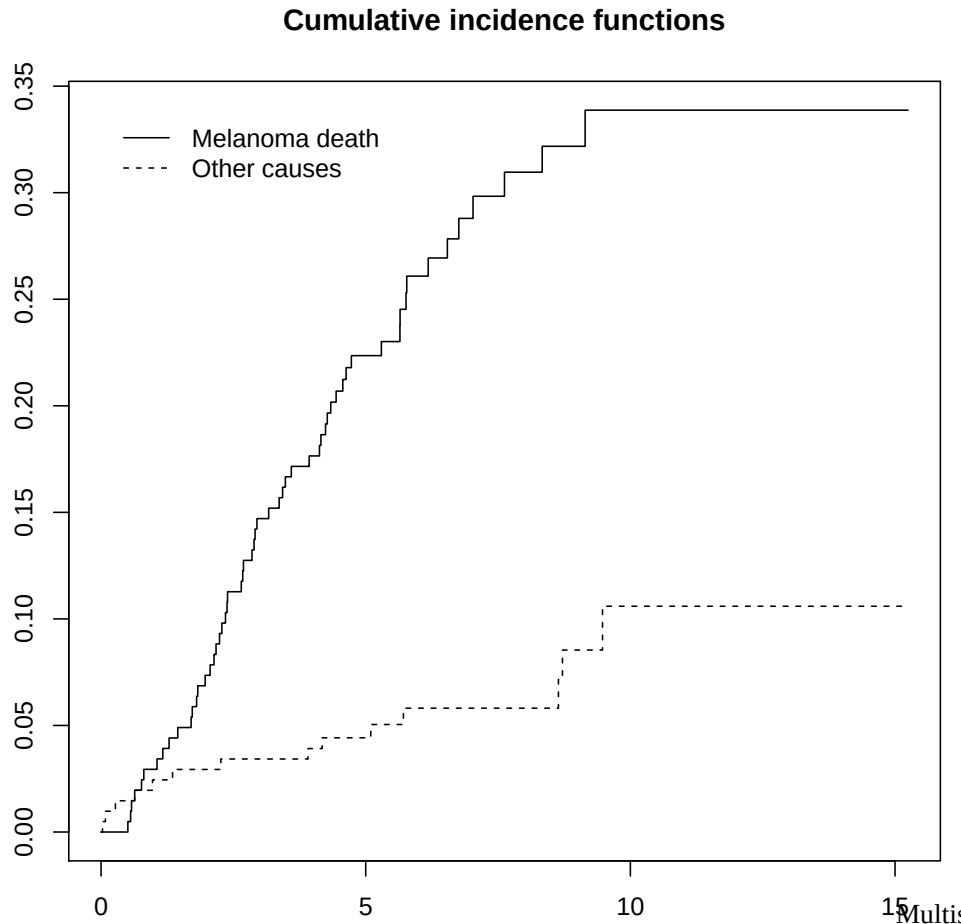
$$\hat{P}_{0h}(s,t) = \int_s^t \hat{P}_{00}(s,u-)\frac{dN_{0h}(u)}{Y_0(u)}$$

for $h = 1, \ldots, K$. The $\hat{P}_{0h}(s,t)$ for $h > 0$ are called cumulative incidence functions.

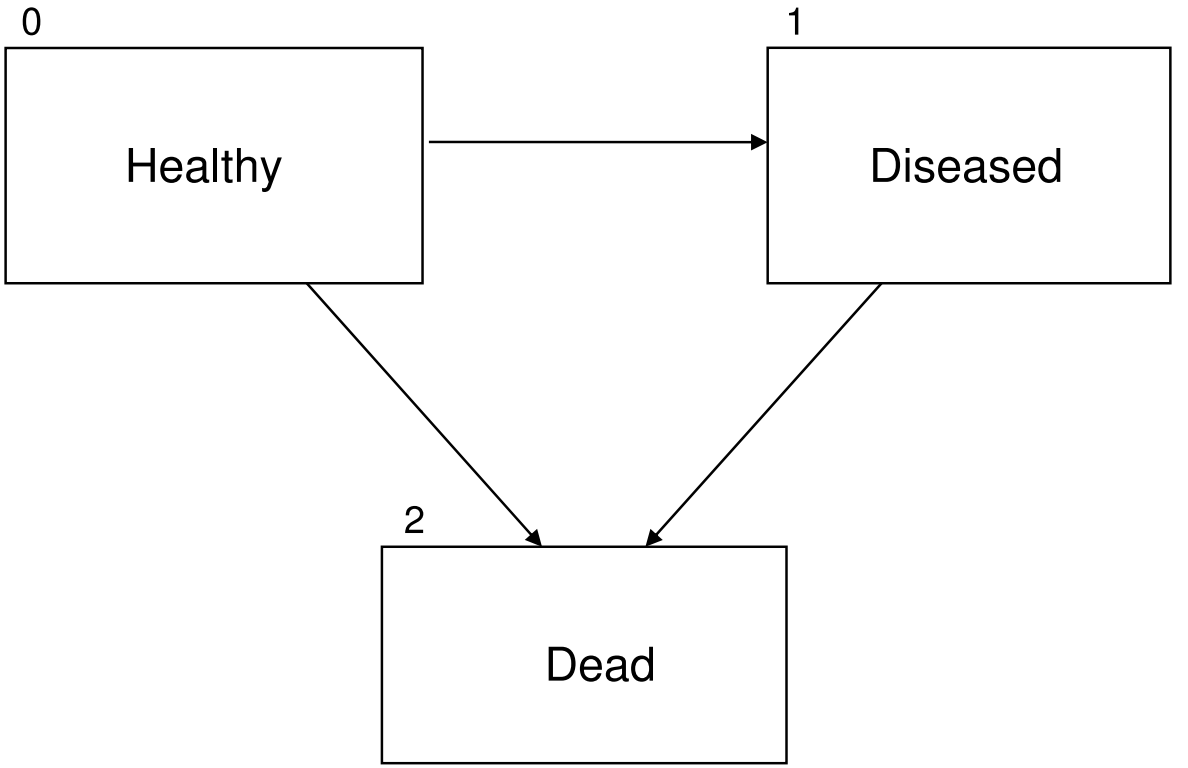
Cumulative incidence for melanoma data

can be obtained in R through the `survfit` function.

```
> mel$status2=1*(mel$status==1)+2*(mel$status==4)
> plot(survfit(Surv(lifetime,status2,type="mstate")~1,data=mel),lty=1:2
> title("Cumulative incidence functions")
> legend(0,0.34,lty=1:2,c("Melanoma death","Other causes"),bty="n")
```



(Healthy-)Illness-Death (ID)



Hazards (intensities) $\alpha_{gh}(t)$ for transition from state g to state h at time t with $\alpha_{01}(t) > 0$, $\alpha_{02}(t) > 0$ and $\alpha_{12}(t) > 0$ (at least for some t) and all other $\alpha_{gh}(t) = 0$.

Transition probabilities in ID-process

Let $X(t)$ be the state of the process at time t, $X(t) \in \{0, 1, 2\}$ and define transition probabilities

$$P_{gh}(s,t) = P(X(t) = h | X(s) = g)$$

We then get that

$$P_{00}(s,t) = \exp(-\int_s^t (\alpha_{01}(u) + \alpha_{02}(u))du),$$

$$P_{11}(s,t) = \exp(-\int_s^t \alpha_{12}(u)du)$$

and

$$P_{01}(s,t) = \int_s^t P_{00}(s,u)\alpha_{01}(u)P_{11}(u,t)du$$

where the integrand is the "density" of staying in 0 until u, then moving to 1 at u and staying in 1 up to t.

Transition probabilities in ID-process, II

We may also calculate

$$P_{02}(s,t) = \int_s^t P_{00}(s,u)\alpha_{02}(u)du + \int_s^t P_{01}(s,u)\alpha_{12}(u)du$$

where the first terms come from the direct move from 0 to 2 (death without previous disease) and the second from death after disease.

Finally

$$\begin{aligned} P_{12}(s,t) &= \int_s^t \exp(-\int_s^u \alpha_{12}(v)dv)\alpha_{12}(u)du \\ &= 1 - \exp(-\int_s^t \alpha_{12}(u)du) = 1 - P_{11}(s,t) \end{aligned}$$

and all other $P_{gh}(s,t) = 0$ (except for $P_{22}(s,t) = 1$).

Estimation transition probabilities, ID

Let

- $Y_h(t) =$ no. in state h at time $t-$
- $N_{gh}(t) =$ no. of direct transitions from g to h in $[0, t]$
- $N_{g\bullet}(t) = \sum_h N_{gh}(t)$ no. transitions out of g in $[0, t]$

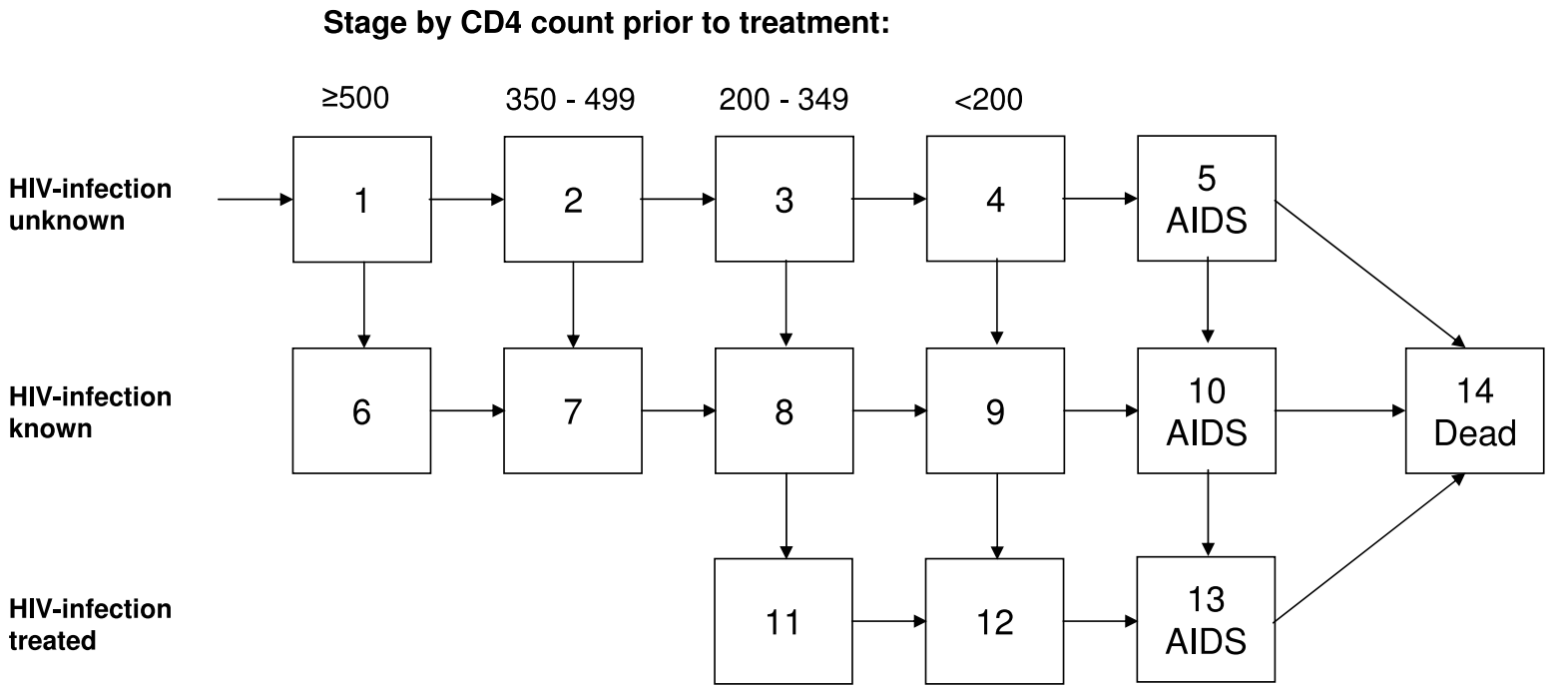
We get estimates

$$\hat{P}_{00}(s,t) = \prod_{s<u\leq t}[1 - \frac{dN_{0\bullet}(u)}{Y_0(u)}]$$
$$\hat{P}_{11}(s,t) = \prod_{s<u\leq t}[1 - \frac{dN_{12}(u)}{Y_1(u)}]$$
$$\hat{P}_{01}(s,t) = \int_s^t \hat{P}_{00}(s,u)\frac{dN_{01}(u)}{Y_0(u)}\hat{P}_{11}(u,t)$$

and

$$\hat{P}_{02}(s,t) = \int_s^t \hat{P}_{00}(s,u)\frac{dN_{02}(u)}{Y_0(u)} + \int_s^t \hat{P}_{01}(s,u)\frac{dN_{12}(u)}{Y_1(u)}$$

General (complicated) event scheme



General event schemes

Let $X(t)$ be the state of the process at t

$\alpha_{gh}(t) =$ hazard/intensity of moving from g to h at t

$\alpha_{gg}(t) = -\sum_{h \neq g} \alpha_{gh}(t)$

$P_{gh}(s,t) = P(X(t) = h | X(s) = g) =$ transition probabilities

and the matrices of transition probabilities and hazards

$$\mathbf{P}(s,t) = [P_{gh}(s,t)]_{g,h=0}^{k} \text{ and } \alpha(u) = [\alpha_{gh}(u)]_{g,h=0}^{k}$$

which may be written as a continuous product

$$\mathbf{P}(s,t) = \prod_{s<u\leq t} \mathbf{P}(u, u+du) = \prod_{s<u\leq t} [\mathbf{I} + \alpha(u)du]$$

where $\mathbf{I}$ is the identity matrix.

General event schemes, estimation

In place of $\alpha_{gh}(u)du$ use

$$d\hat{A}_{gh}(u) = \frac{dN_{gh}(u)}{Y_g(u)} \text{ for } g \neq h$$
$$d\hat{A}_{gg}(u) = -\frac{dN_{g\bullet}(u)}{Y_g(u)}$$

Here

- $Y_h(t) =$ no. in state h at time $t-$
- $N_{gh}(t) =$ no. of direct transitions from g to h in $[0, t]$
- $N_{g\bullet}(t) = \sum_h N_{gh}(t)$ no. transitions out of g in $[0, t]$

The estimator of the transition probabilities

$$\hat{\mathbf{P}}(s,t) = \prod_{s<u\leq t} [\mathbf{I} + d\hat{\mathbf{A}}(u)]$$

is called the **Aalen-Johansen** estimator

Aalen-Johansen estimator, simulation

The Aalen-Johansen estimator may be applied to the Healthy-Illness-Death process.

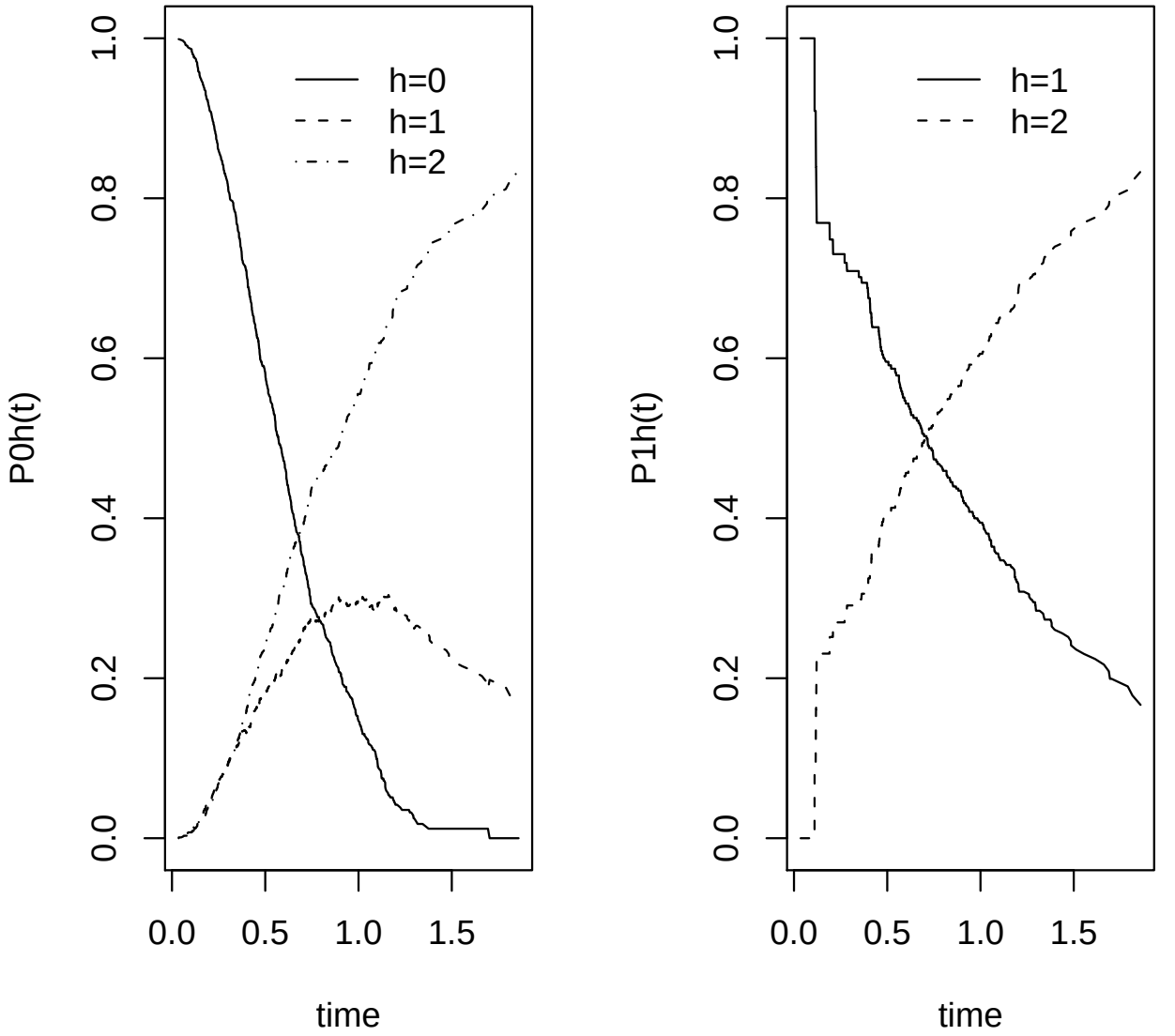
Let

- $\alpha_{01}(t) = t$ (Weibull, $k = 2$)
- $\alpha_{02}(t) = t$ (Weibull, $k = 2$)
- $\alpha_{12}(t) = 1$ (exponential)
- Censoring uniform on $[0, 1]$
- $n = 1000$ individuals

Last week we simulated such process and programmed estimation of the transition probabilities.

Today we will show how the R-library `mstate` can be used to obtain the estimates.

Estimated trans.prob. from simulation



Framework for `mstate`

We are going to

- Simulate the ID-process again and set up a wide format data file with one record per individual
- Define the possible transitions for the ID process
- Transfer the wide format data to long format data with one record per (possibly censored) transition
- Calculate increments to the Aalen-Johansen estimator
- and finally calculate and plot the Aalen-Johansen

Simulation of ID and "wide-format" data-setup

```
# Simulates n=100 replicates of the ID process and censoring time
n<-100
timesick<-rweibull(n,2)
timedeath<-rweibull(n,2)
timesickdeath<-rexp(n)
censtime<-runif(n)*2
# Defines indicators of three events
D01<-1*(timesick<pmin(timedeath,censtime))
D02<-1*(timedeath<pmin(timesick,censtime))
D12<-1*(D01==1)*((timesick+timesickdeath)<censtime)
# Defines censored transition times to sickness and death states
obstimesick=timesick*D01+pmin(timedeath,censtime)*(1-D01)
obstimedeath=timedeath*D02+censtime*D01*(1-D12)+
             (timesick+timesickdeath)*D12+censtime*(1-D01-D02)
# Defines individual process and indicator of death
id=1:n
D2=D02+D01*D12
# Data are put in "wideformat" file with one line per individual
simdat=data.frame(cbind(id,obstimesick,D01,obstimedeath,D2))
head(simdat)
```

Wide format data file - without covariates

```
> head(simdat)
  id obstimesick D01 obstimedeath D2
1  1   0.6635938   1    1.6769120  0
2  2   0.5448141   1    1.1839140  1
3  3   0.4957327   0    0.4957327  1
4  4   0.4916671   1    0.6039680  1
5  5   0.3334876   1    1.2061065  0
6  6   0.2723386   0    0.2723386  0
```

The wide format data (above) need

- One line per individual
- An id-number
- The possibly censored time until illness
- An indicator for whether time until illness was observed
- The possibly censored time until death
- An indicator for whether time until death was observed

Defining the possible transitions

The transitions for ID can be set up in two ways

- Using an ID-specific function `trans.illdeath`
- Using a general function `transMat` for transitions

```
> library(survival)
> library(mstate)
> tmat2=trans.illdeath(names=c("Healthy","Sick","Dead"))
> tmat2
         to
from      Healthy Sick Dead
  Healthy      NA    1    2
  Sick         NA   NA    3
  Dead         NA   NA   NA
> tmat=transMat(x=list(c(2,3),c(3),c()),names=c("Healthy","Sick","Dead"
> tmat
         to
from      Healthy Sick Dead
  Healthy      NA    1    2
  Sick         NA   NA    3
  Dead         NA   NA   NA
```

Getting the long format data - one line per transition

```
> mssimdat=msprep(data=simdat,trans=tmat,
  time=c(NA,"obstimesick","obstimedeath"),status=c(NA,"D01","D2"))
> mssimdat[c(1:8,15:16),]
An object of class 'msdata'

Data:
   id from to trans    Tstart     Tstop      time status
1   1    1  2     1 0.0000000 0.6635938 0.6635938      1
2   1    1  3     2 0.0000000 0.6635938 0.6635938      0
3   1    2  3     3 0.6635938 1.6769120 1.0133181      0
4   2    1  2     1 0.0000000 0.5448141 0.5448141      1
5   2    1  3     2 0.0000000 0.5448141 0.5448141      0
6   2    2  3     3 0.5448141 1.1839140 0.6390999      1
7   3    1  2     1 0.0000000 0.4957327 0.4957327      0
8   3    1  3     2 0.0000000 0.4957327 0.4957327      1
15  6    1  2     1 0.0000000 0.2723386 0.2723386      0
16  6    1  3     2 0.0000000 0.2723386 0.2723386      0
```

Overview over observed transitions

using the `mstate` function `events` on the long format data.

```
> events(mssimdat)
$Frequencies
         to
from      Healthy Sick Dead no event total entering
  Healthy       0   37   26       37             100
  Sick          0    0   21       16              37
  Dead          0    0    0       47              47

$Proportions
         to
from        Healthy      Sick      Dead  no event
  Healthy 0.0000000 0.3700000 0.2600000 0.3700000
  Sick    0.0000000 0.0000000 0.5675676 0.4324324
  Dead    0.0000000 0.0000000 0.0000000 1.0000000
```

For instance 37 of the healthy became sick, 26 eventually died and 37 were censored.

Calculating the cumulative hazards

```
c1=coxph(Surv(Tstart,Tstop,status)~strata(trans),
                  data=mssimdat,method="breslow")
simdatfit=msfit(c1,trans=tmat)
> names(simdatfit)
[1] "Haz"    "varHaz" "trans"
> head(simdatfit$Haz)
       time        Haz trans
1 0.1681827 0.01111111     1
2 0.1886775 0.01111111     1
3 0.2001121 0.02273902     1
4 0.2316646 0.03478721     1
5 0.2354188 0.04698233     1
6 0.2385675 0.05948233     1
> table(simdatfit$Haz$trans)
 1  2  3
85 85 85
```

We do Cox-regression without covariates with "stratification" on the different possible transitions to get the cumulative hazards. Their increments will be used in the Aalen-Johansen estimator.

Finally - The Aalen-Johansen Estimator $\hat{P}_{jk}(0,t)$!

```
> simdatprob=probtrans(simdatfit,predt=0,method="aalen")
> names(simdatprob)
[1] ""          ""          ""          "trans"     "method"    "predt"
> names(simdatprob[[1]])
[1] "time"    "pstate1" "pstate2" "pstate3" "se1"     "se2"     "se3"
> names(simdatprob[[2]])
[1] "time"    "pstate1" "pstate2" "pstate3" "se1"     "se2"     "se3"
> names(simdatprob[[3]])
[1] "time"    "pstate1" "pstate2" "pstate3" "se1"     "se2"     "se3"
```

Here `simdatprob[[1]]$pstate1` give $\hat{P}_{0h}(0,t)$ for times in `simdatprob[[1]]$time` etc.

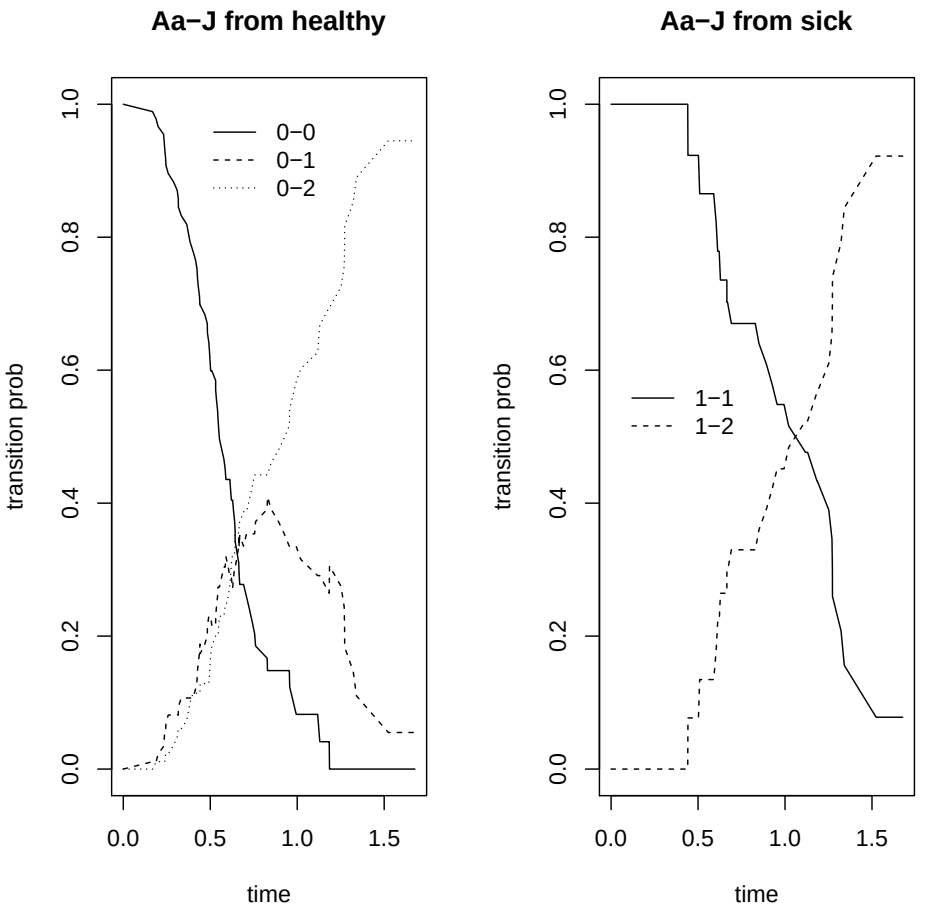
```
> cbind(simdatprob[[1]]$time,simdatprob[[1]]$pstate1,
        simdatprob[[1]]$pstate2,simdatprob[[1]]$pstate3)
      [,1]  [,2]  [,3]  [,4]
 [1,] 0.000 1.000 0.000 0.000
 [2,] 0.168 0.989 0.011 0.000
 [3,] 0.189 0.978 0.011 0.011
 [4,] 0.200 0.966 0.022 0.011
 [5,] 0.232 0.955 0.034 0.011
```

Plotting the Aalen-Johansen Estimator $\hat{P}_{jk}(0,t)$!

```
> par(mfrow=c(1,2))
> plot(simdatprob[[1]]$time,simdatprob[[1]]$pstate1,type="l",
              ylim=c(0,1),xlab="time",ylab="transition prob")
> lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate2,lty=2)
> lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate3,lty=3)
> legend(0.4,1,lty=1:3,c("0-0","0-1","0-2"),bty="n")
> title("Aa-J from healthy")
>
> plot(simdatprob[[2]]$time,simdatprob[[2]]$pstate2,type="l",
              ylim=c(0,1),xlab="time",ylab="transition prob")
> lines(simdatprob[[2]]$time,simdatprob[[2]]$pstate3,lty=2)
> legend(0,0.6,lty=1:2,c("1-1","1-2"),bty="n")
> title("Aa-J from sick")
```

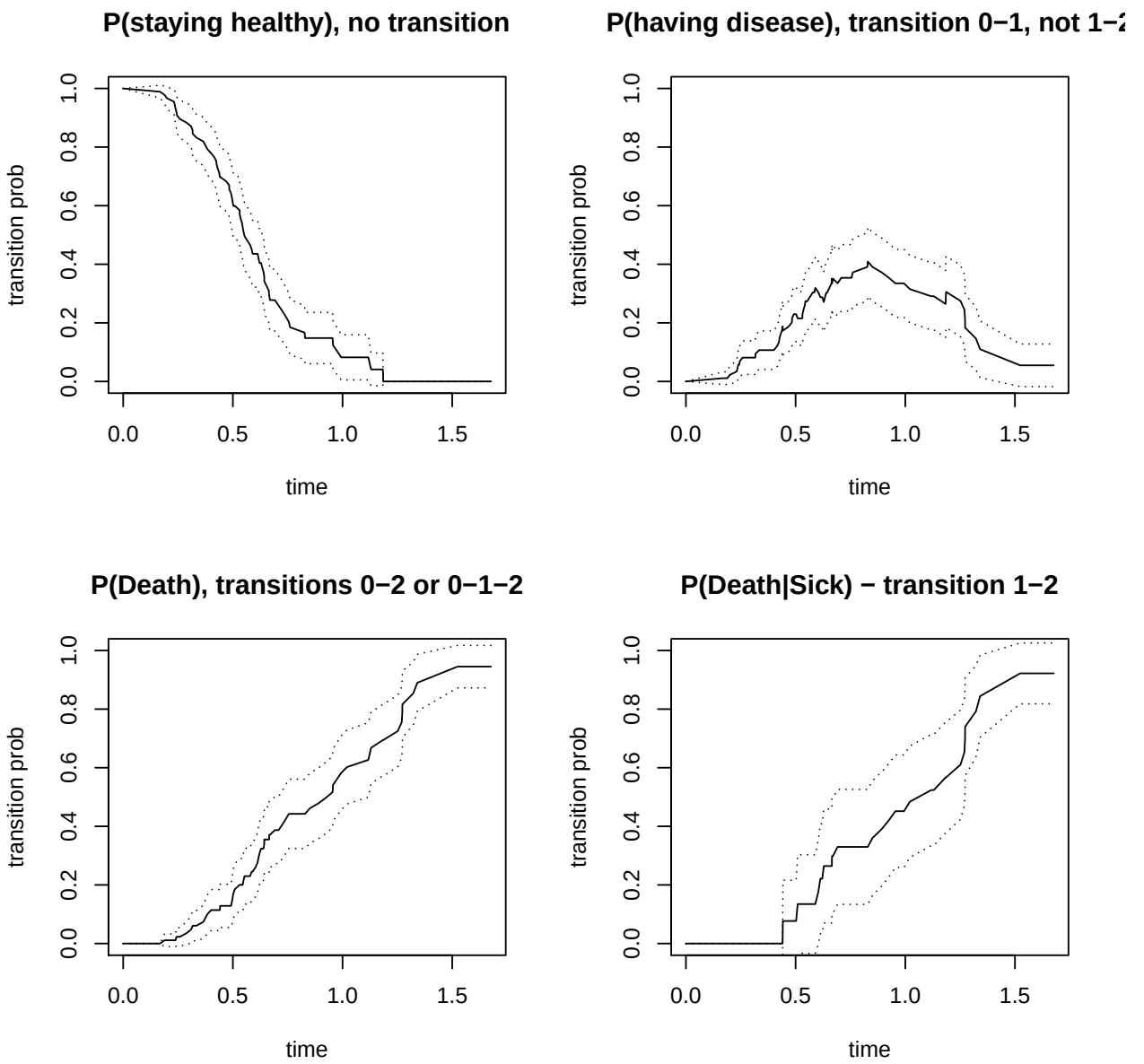
Actual plot on next slide.

Estimated trans.prob. from simulation



Note that we essentially get the same plot as programmed in last weeks slides. Differences are due to different simulation.

Aalen-Johansen estimator with confidence intervals



Plotting Aalen-Johansen with CI - commands

```
par(mfrow=c(2,2))
plot(simdatprob[[1]]$time,simdatprob[[1]]$pstate1,type="l",ylim=c(0,1),
                        xlab="time",ylab="transition prob")
lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate1+1.96*simdatprob[[1]]
lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate1-1.96*simdatprob[[1]]
title("P(staying healthy), no transition")

plot(simdatprob[[1]]$time,simdatprob[[1]]$pstate2,type="l",ylim=c(0,1),
lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate2+1.96*simdatprob[[1]]
lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate2-1.96*simdatprob[[1]]
title("P(having disease), transition 0-1, not 1-2")

plot(simdatprob[[1]]$time,simdatprob[[1]]$pstate3,type="l",ylim=c(0,1),
lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate3+1.96*simdatprob[[1]]
lines(simdatprob[[1]]$time,simdatprob[[1]]$pstate3-1.96*simdatprob[[1]]
title("P(Death), transitions 0-2 or 0-1-2")

plot(simdatprob[[2]]$time,simdatprob[[2]]$pstate3,type="l",ylim=c(0,1),
lines(simdatprob[[2]]$time,simdatprob[[2]]$pstate3+1.96*simdatprob[[2]]
lines(simdatprob[[2]]$time,simdatprob[[2]]$pstate3-1.96*simdatprob[[2]]
title("P(Death|Sick) - transition 1-2")
```