

Survival - introduction

Melania Pintilie
Ontario Cancer Institute/PMH/UHN
University of Toronto



Princess Margaret Hospital
University Health Network



Outline

- **Practical considerations**
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- Hazard function
- Cox proportional hazards model

Outline

- Parametric modelling
- Sample size
- Competing risks: justification and definition
- Non-parametric estimation of the probability of event in the presence of competing risks
- Reporting

Practical considerations

- Examples: SAS and R
- **Input/Output: R <->SAS**
- **Data manipulation is easier in SAS**
- **Graphing is easier in R**

Strong recommendation:

- **CHECK the data**

Example

age	sex	hgb	clinstg	blktxcat	ch	rt	mcens
56	F	140	2	1	N	Y	0
36	F	130	2	1	N	Y	0

Input/Output in SAS

```
filename f 'c:/your_directory/follic.csv';

data follic; infile f dsd truncover delimiter=',' firstobs=2
  obs=542;
input age sex $ hgb clinstg blktxcat
  ch $ rt $ mcens;
*****.;
filename outf 'c:/your_directory/follic_out.csv';

proc export data=follic outfile=outf dbms=csv replace;
```

Details for R software

<http://www.r-project.org> (CRAN)

- basic: general functions like **mean**, **plot**, **sum**, **distributions**, **tests**, **linear regression**
- libraries: **survival** and **cmprsk**
- The libraries need to be installed once
- The libraries need to be loaded every time you start a new session
 - Menu: Packages-> Load
 - **library(survival)** or **library(cmprsk)**

Details for R

- Change the directory to your working directory
 - Menu: File -> Change dir...
 - `getwd("C:/your_directory")`
- Save the workspace
 - Menu: File -> Save Workspace
 - `save.image("C:/your_directory/.RData")`
 - Will save all objects (datasets, variables) but need to load the libraries every time
- Useful commands: **help, help.search, class**
- Use `source('name of file')` to read in a function



.RData

Input/Output in R

```
follic=read.table('c:/your_directory/follic.csv',  
  sep=',',na.strings='.',header=T)  
names(follic)=casefold(names(follic))  
#####  
write.table(follic,'follic_out.csv',sep=',')  
## created in your home directory
```

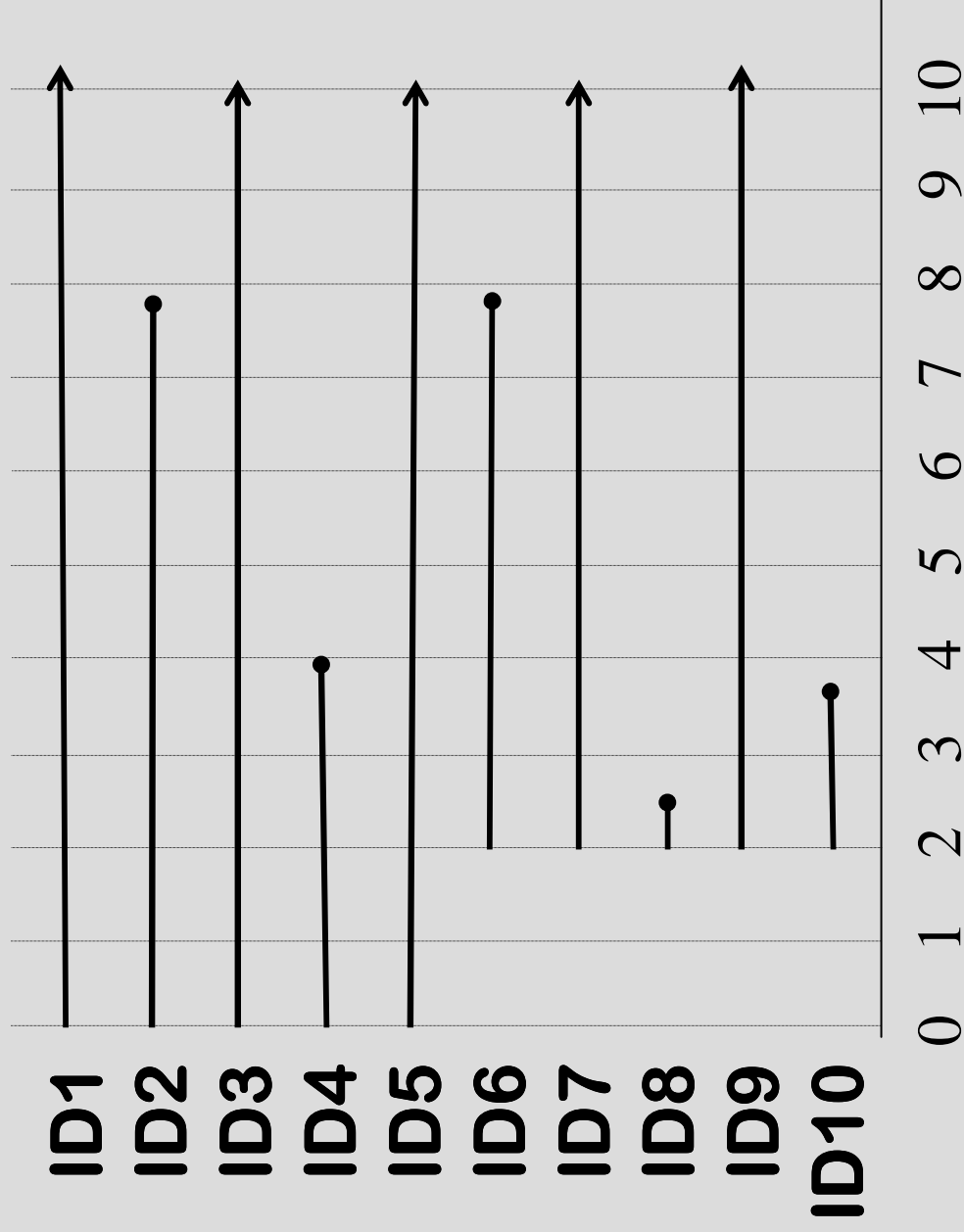
Outline

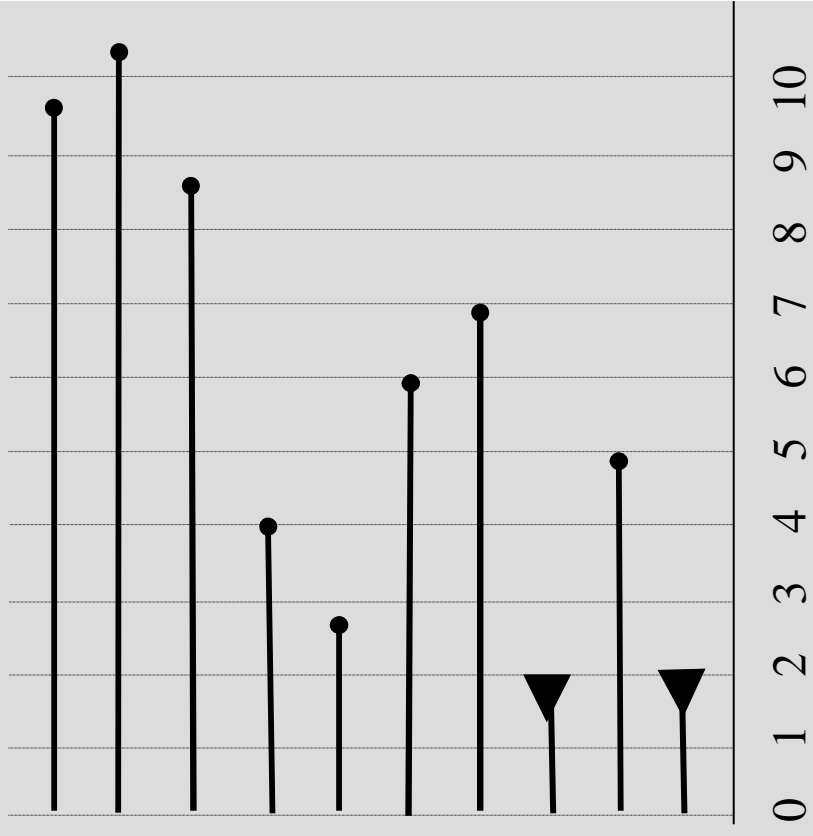
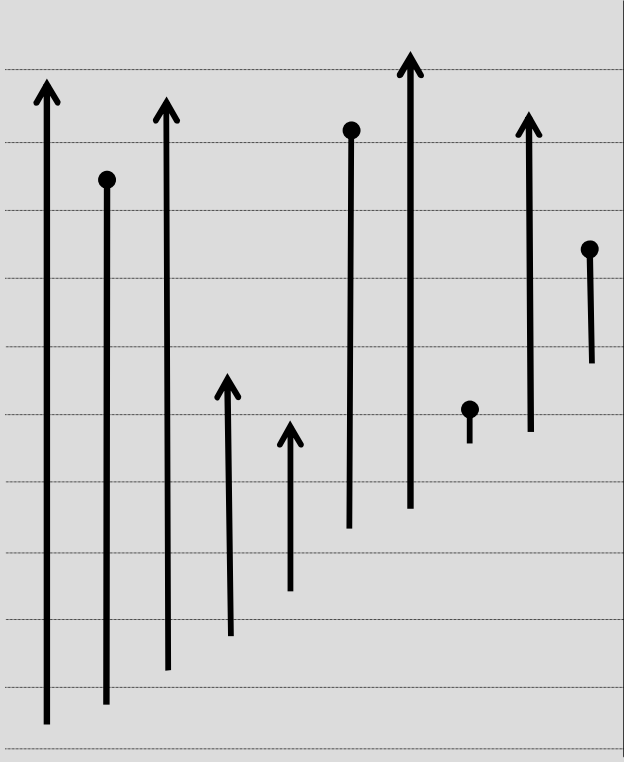
- Practical considerations
- **Characteristics of survival type data**
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- Hazard function
- Cox proportional hazards model

Survival type data

Incomplete data

- Time to death of a group of individuals working in construction (hence *survival*)
- Time that takes for specific light bulbs to burn out
- Time to death of patients diagnosed with heart disease
- The amount of barium in water
- Time of hospitalization





Survival type data

- 2 components:
 - A positive value representing the quantity which we measure
 - Circle/arrow which signifies whether the event occurred
- We will refer to these:
 - Time variable
 - Censor variable (indicator variable 1 = for event observed, 0 = event not observed).

Censoring

- Right censoring at t . The value of the observation is not known, but only that is greater than t .
 - *Example: Time to death*
- Left censoring at t . The value of the observation is not known, but only that is less than t .
 - *Example: The level of a certain contaminant in the water (say barium).*

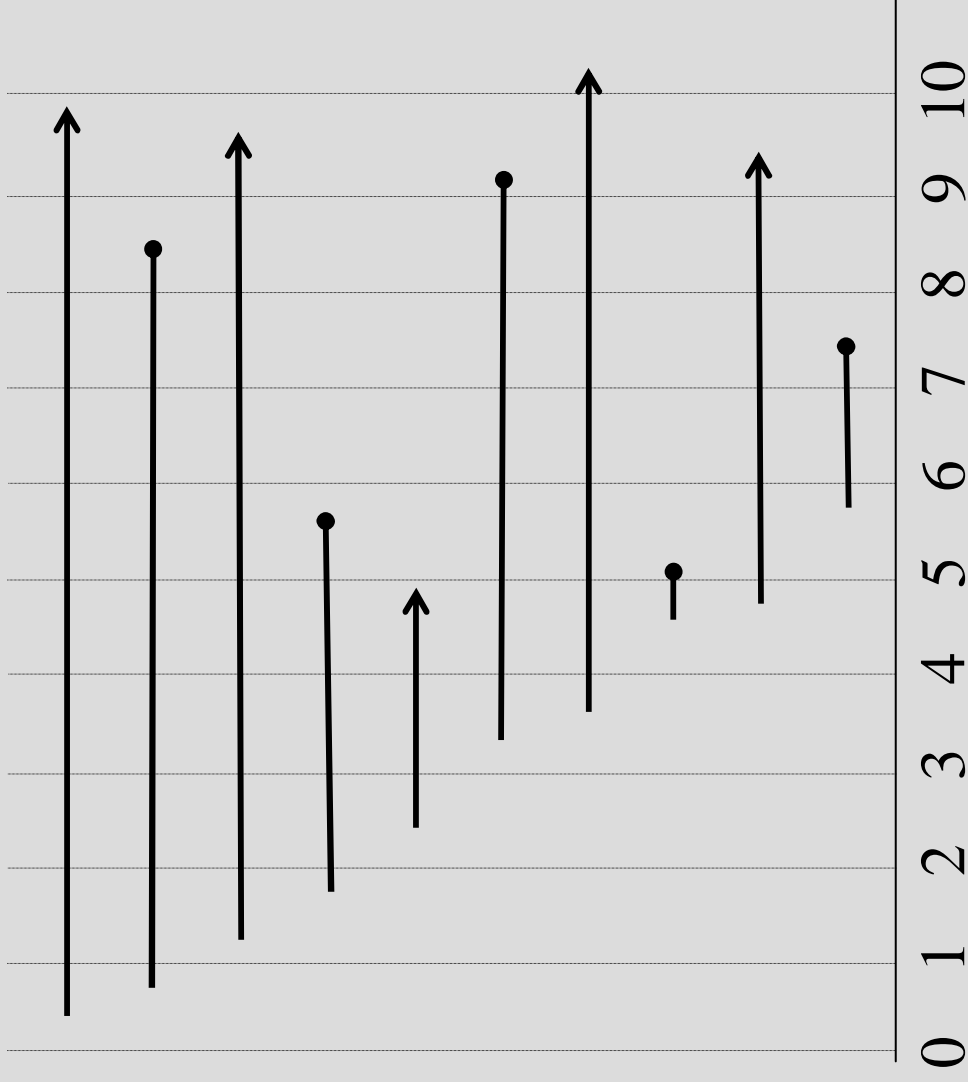
- Examples from cancer research
- Time to death: diagnosis of disease to death, years
- Right censoring

Survival type data

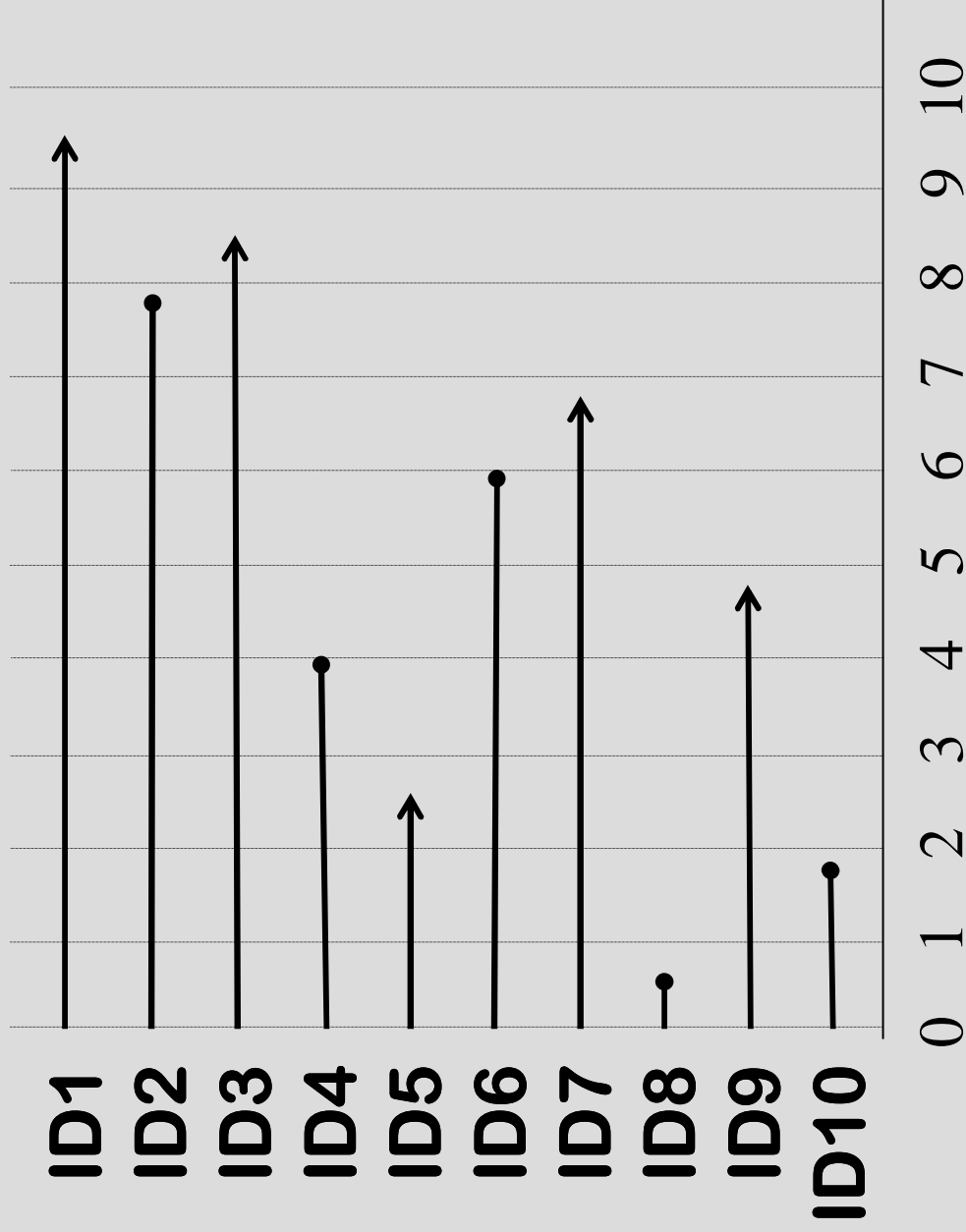
- If all deaths (events) are observed => continuous data
- If all subjects are followed for 3 years => categorical data
- Otherwise: incomplete observations.
- Need to account for the time under observation even if the event was not ultimately observed

Outline

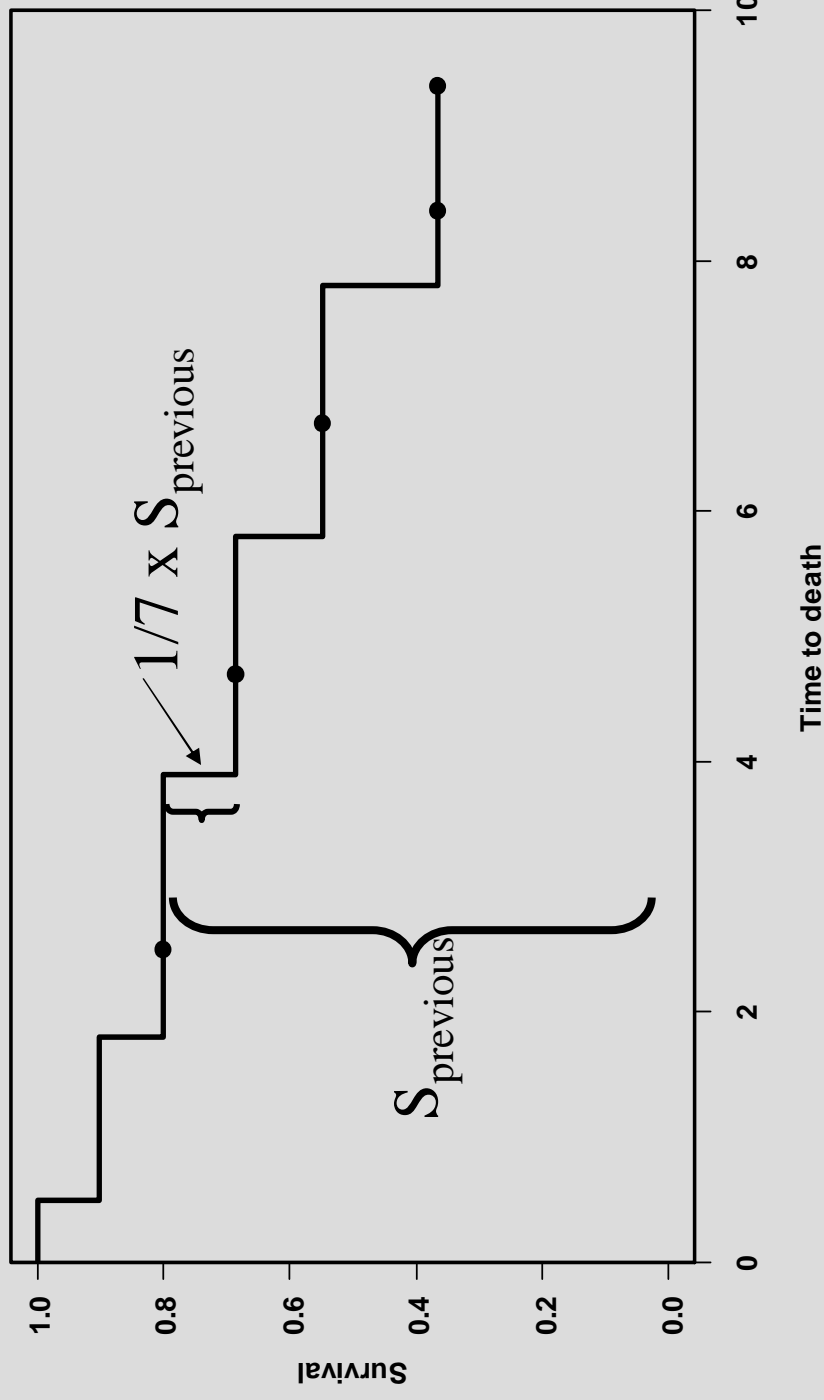
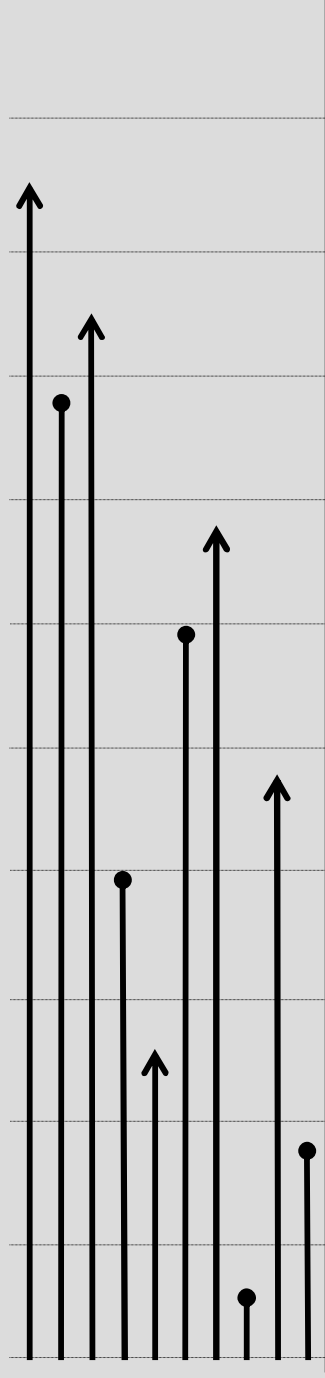
- Practical considerations
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- Hazard function
- Cox proportional hazards model



Survival at 1 year 90%
Survival at 2 years 80%
Survival at 4 years could be 60% or 70%



Non-parametric estimation for the survivor function



Non-parametric estimation for the survivor function

$$S(t_3) = S(t_2) - \frac{1}{7} S(t_2) = \frac{9}{10} \frac{8}{9} \left(1 - \frac{1}{7}\right) = \frac{9}{10} \frac{8}{9} \frac{6}{7}$$

$$\begin{aligned} S(t_3) &= S(t_2) - \frac{1}{7} S(t_2) \\ &= S(t_2) - \frac{d_3}{n_3} S(t_2) \\ &= S(t_2) \left(1 - \frac{d_3}{n_3}\right) \\ &= S(t_2) \left(\frac{n_3 - d_3}{n_3}\right) \end{aligned}$$

d_j = number of events at t_j

n_j = number at risk at t_j

$$S(t_3) = \left(\frac{n_1 - d_1}{n_1}\right) \left(\frac{n_2 - d_2}{n_2}\right) \left(\frac{n_3 - d_3}{n_3}\right)$$

- Non-parametric estimation of survival
- Kaplan-Meier estimation
- Product limit estimation
- Estimates the probability that an individual survives beyond t

$$S(t) = P(T > t)$$

$$F(t) = P(T \leq t)$$

SAS – estimation and plot

```
proc lifetest data=follic plots=(s)
  censoredsymbol='|' timelist=
  0,1,2,3,4,5;
time survtime*stat(0);
survival out=your_data conftype=loglog;
proc print data=your_data;
run;
```

SAS output

The LIFETEST Procedure

Product-Limit Survival Estimates

Timelist	survtime	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00000	0.0000	1.0000	0	0	0	541
1.00000	0.9172	0.9686	0.0314	0.00750	17	524
2.00000	1.9986	0.9186	0.0814	0.0118	44	496
3.00000	2.9897	0.8759	0.1241	0.0142	67	471
4.00000	3.9671	0.8366	0.1634	0.0159	88	441
5.00000	4.9144	0.8018	0.1982	0.0173	106	406

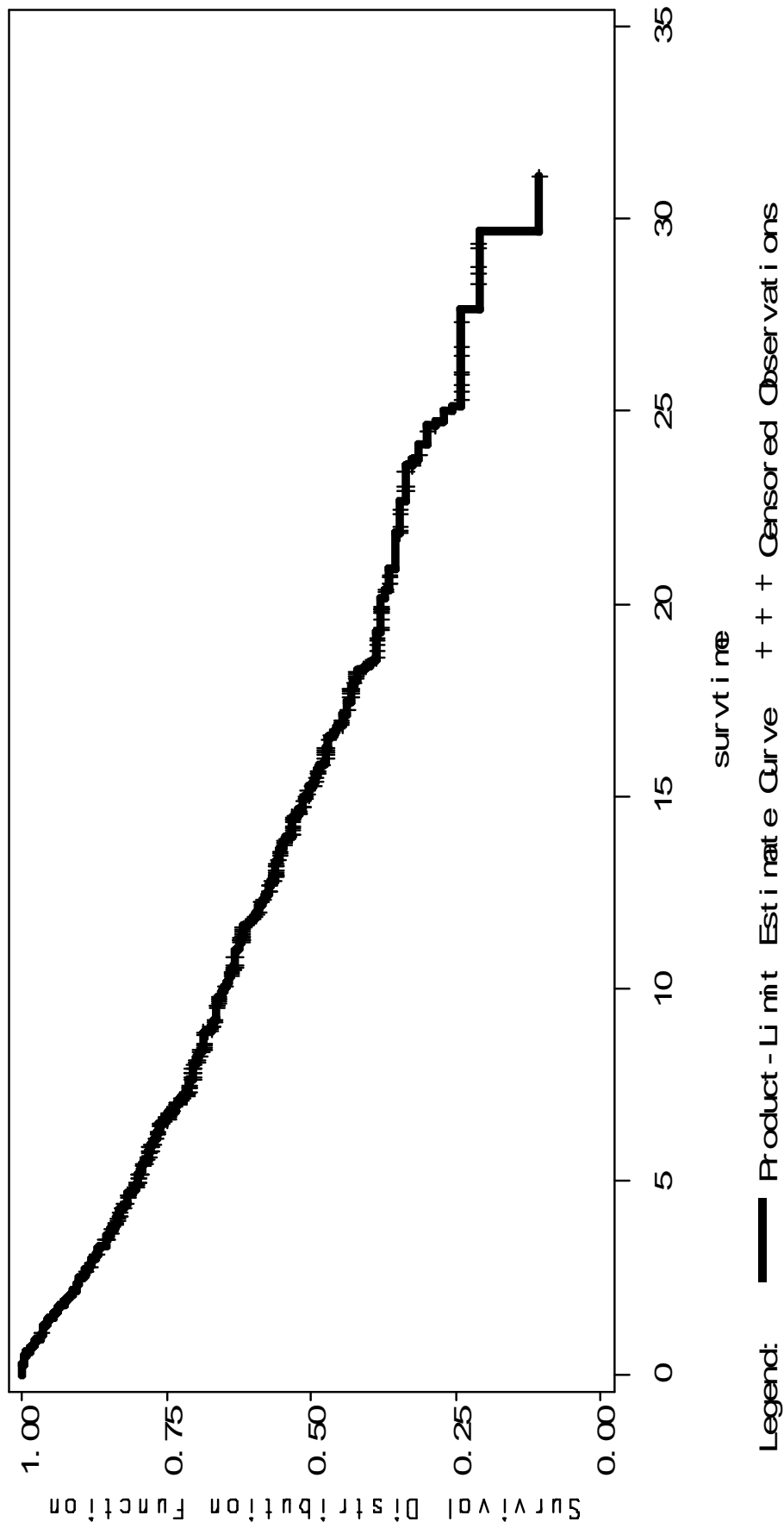
Summary Statistics for Time Variable survtime

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval	
		[Lower	Upper)
75	25.1499	23.7645	.
50	15.2799	13.6728	17.1116
25	6.6721	5.5852	7.6550

25

Follicular lymphoma



SAS output – data your_data

Obs	survtime	_CENSOR_	SURVIVAL	CONFTYPE	SDF_LCL	SDF_UCL
1	0.00000	.	1.00000		1.00000	1.00000
2	0.23546	0	0.99815	LOGLOG	0.98695	0.99974
3	0.43258	0	0.99630	LOGLOG	0.98530	0.99907
4	0.45996	0	0.99445	LOGLOG	0.98291	0.99821
5	0.55305	0	0.99261	LOGLOG	0.98042	0.99722
.....						
18	0.91718	0	0.96858	LOGLOG	0.94994	0.98035
19	1.00753	0	0.96673	LOGLOG	0.94771	0.97891
20	1.09514	1	0.96673	.	.	.
21	1.23203	0	0.96488	LOGLOG	0.94549	0.97745
22	1.23751	0	0.96302	LOGLOG	0.94327	0.97599
23	1.27584	0	0.96117	LOGLOG	0.94107	0.97451

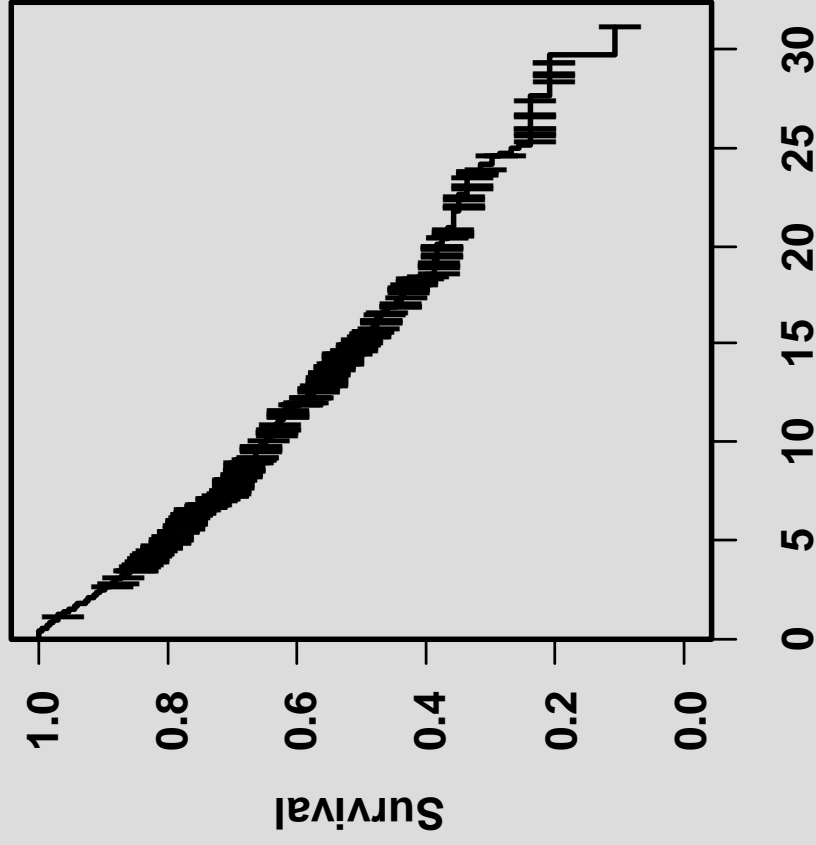
18	0.91718	0	0.96858	LOGLOG	0.94994	0.98035
19	1.00753	0	0.96673	LOGLOG	0.94771	0.97891

R – estimation and plot

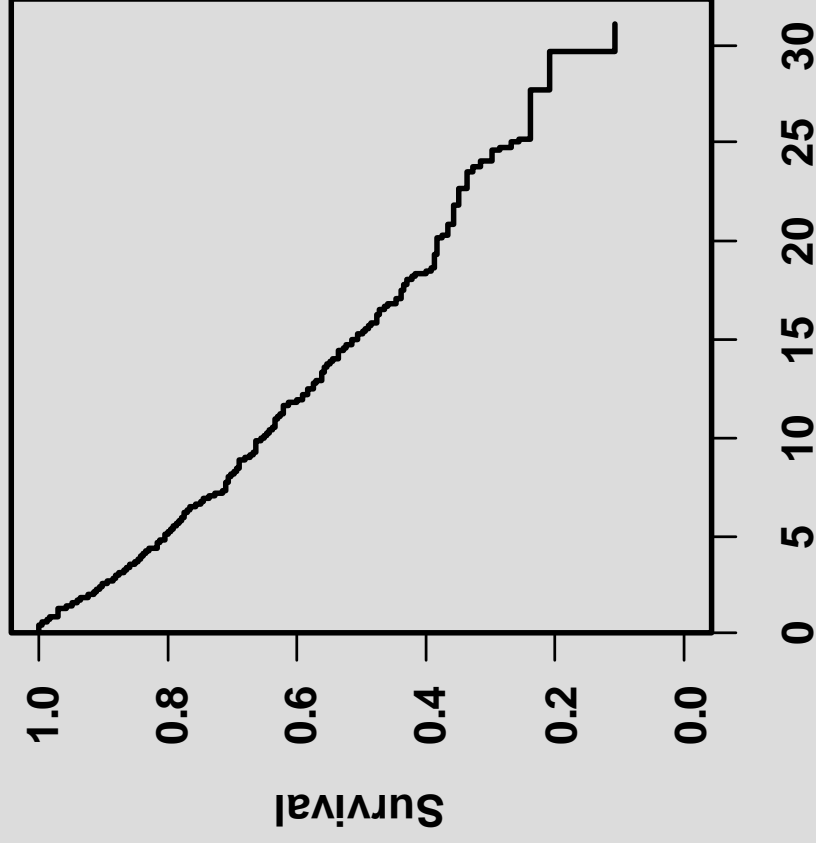
```
library(survival)
fit=survfit(Surv(survtime*stat)~1,
            data=follic,conf.type='log-log')
summary(fit,times=c(0:5))

plot(fit,lwd=2,mark='|',conf.int=F,
      xlab='Time to death (years)', ylab='Survival')

help(survfit.plot)
```



Time to death (years)



Time to death (years)

Outline

- Practical considerations
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- Hazard function
- Cox proportional hazards model

Parametric estimation

	Distribution	Survivor	Density
	$F(t)=P(T \leq t)$	$S(t)=1-F(t)$	$f = \frac{\partial F}{\partial t} = -\frac{\partial S}{\partial t}$
Exponential	$1 - e^{-\lambda t}$	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$
Weibull	$1 - e^{-(\lambda t)^\nu}$	$e^{-(\lambda t)^\nu}$	$\lambda \nu (\lambda t)^{\nu-1} e^{-(\lambda t)^\nu}$
Log-normal	$\Phi \left(\frac{\log(t) - \mu}{\sigma} \right)$	$1 - \Phi \left(\frac{\log(t) - \mu}{\sigma} \right)$	

Estimating the survivor function - exponential

$$S(t) = e^{-\lambda t} \qquad S(t) = e^{-\frac{t}{\theta}} \qquad f(t) = \frac{1}{\theta} e^{-\frac{t}{\theta}}$$

$$\theta = e^{\beta}$$

α = scale parameter (assumed 1)

β = intercept

$$L = \prod_D f(t_i) \prod_C S(t_i) \\ = \prod_D \frac{1}{\theta} \exp\left(-\frac{t_i}{\theta}\right) \prod_C \exp\left(-\frac{t_i}{\theta}\right) \\ \hat{\beta} = \log\left(\frac{\sum_{\text{all}} t_i}{\text{number of deaths}}\right)$$

SAS: Estimating the survivor function - exponential

```
proc lifereg data=follic;  
model survtime*stat(0)=/dist=exponential;  
run;
```

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	3.0757	0.0625	2.9532 3.1982	2421.69	<.0001
Scale	0	1.0000	0.0000	1.0000 1.0000		
Weibull Scale	1	21.6643	1.3540	19.1665 24.4875		
Weibull Shape	0	1.0000	0.0000	1.0000 1.0000		

$$S(t) = e^{-\frac{t}{\theta}}$$

$$\theta = e^{\beta}$$

R: Estimating the survivor function - exponential

```
fitw=survreg(Surv(survtime,stat)~1,  
data=follic,dist='exponential')
```

```
summary(fitw)
```

```
Call:
```

```
survreg(formula = Surv(survtime, stat) ~ 1, data = follic,  
dist = "exponential")
```

	Value	Std. Error	z	p
(Intercept)	3.08	0.0625	49.2	0

Scale fixed at 1

Exponential distribution

Loglik(model)= -1043.4 Loglik(intercept only)= -1043.4

Number of Newton-Raphson Iterations: 4

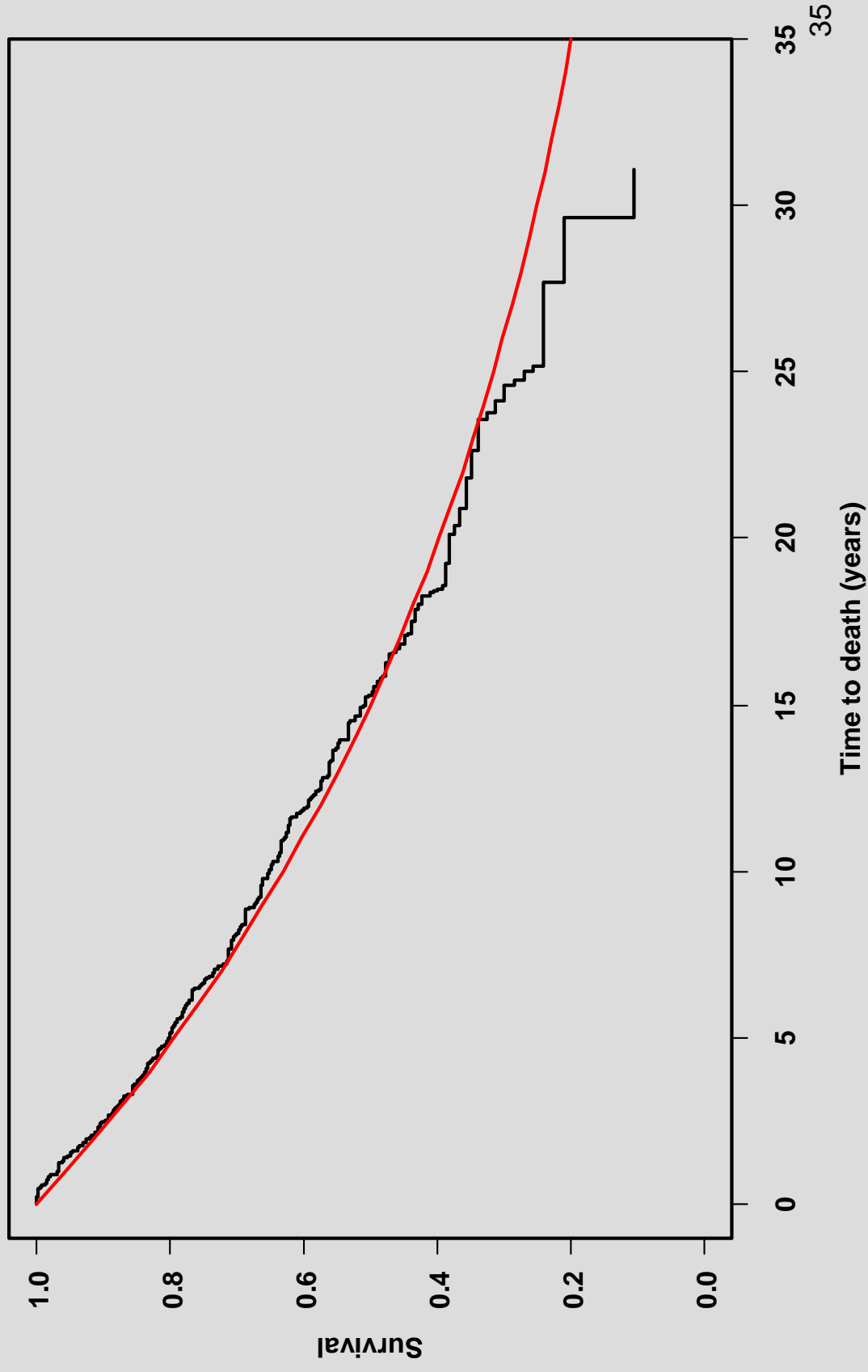
n= 541

$$S(t) = e^{-\frac{t}{\theta}}$$

$$\theta = e^{\beta}$$

Non-parametric vs. exponential distribution

Follicular lymphoma



Estimating the survivor function

Weibull distribution

$$S(t) = \exp\left(-(\lambda t)^\alpha\right)$$

$$S(t) = \exp\left(-\left(\frac{t}{\theta}\right)^{\alpha}\right), \quad \theta = \exp(\beta)$$

α = scale parameter

β = intercept

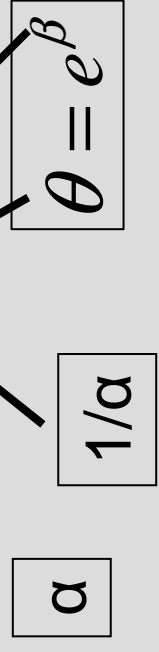
$$L = \prod_D f(t_i) \prod_C S(t_i)$$

$$= \prod_D \frac{t_i^{\alpha-1}}{\alpha[\theta]^\alpha} \exp\left(-\left(\frac{t_i}{\theta}\right)^{\alpha}\right) \prod_C \exp\left(-\left(\frac{t_i}{\theta}\right)^{\alpha}\right)$$

SAS: Estimating the survivor function: Weibull distribution

```
proc lifereg data=follic;  
model survtime*stat(0)=/dist=weibull;  
run;
```

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square Pr > ChiSq
Intercept	1	3.0157	0.0596	2.8989 3.1324	2562.52 <.0001
Scale	1	0.8821	0.0466	0.7953 0.9783	
Weibull scale	1	20.4027	1.2154	18.1542 22.9295	
Weibull shape	1	1.1337	0.0599	1.0222 1.2573	



$$S(t) = \exp\left(-\left(\frac{t}{\theta}\right)^{1/\alpha}\right)$$

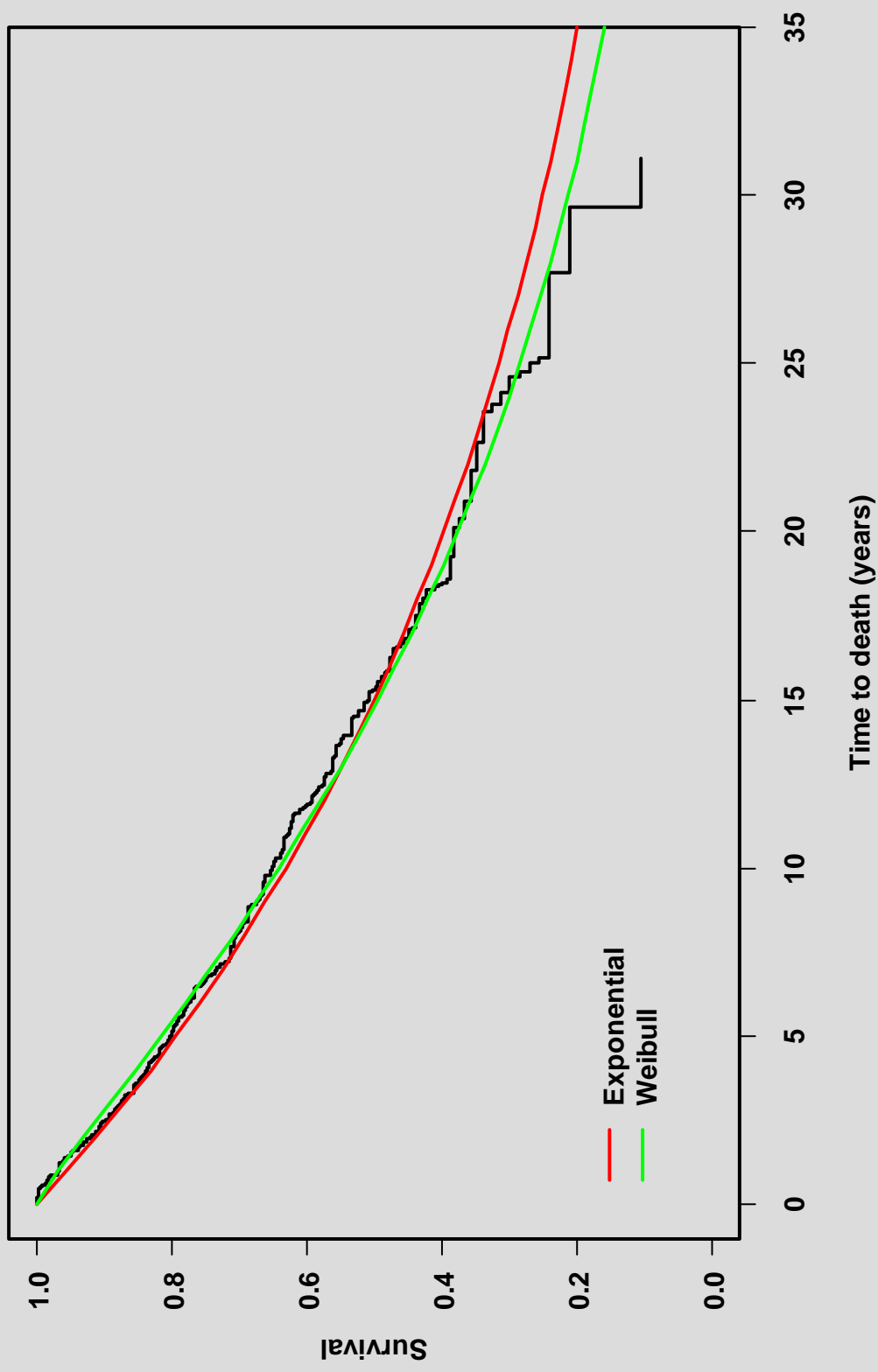
R: Estimating the survivor function: Weibull distribution

```
fitw=survreg(Surv(survtime,stat)~1,  
  data=follic,dist='weibull')  
summary(fitw)
```

```
Value Std. Error      z      p  
(Intercept)  3.016    0.0596  50.62  0.0000  
Log(scale)  -0.125   0.0528 -2.38 0.0175  
Scale= 0.882
```

There is some evidence that the scale is not 1, and thus Weibull fits better than exponential.

Follicular lymphoma



Modelling log-normal distribution

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)$$

σ = scale parameter

μ = intercept

SAS: Estimating the survivor function: lognormal distribution

```
proc lifereg data=follic;  
model survtime*stat(0)=/dist=lognormal;  
run;
```

Parameter	DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square
Intercept	1	2.6948	0.0731	2.5515 2.8381	1358.15
Scale	1	1.3352	0.0630	1.2172 1.4646	<.0001



$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)$$

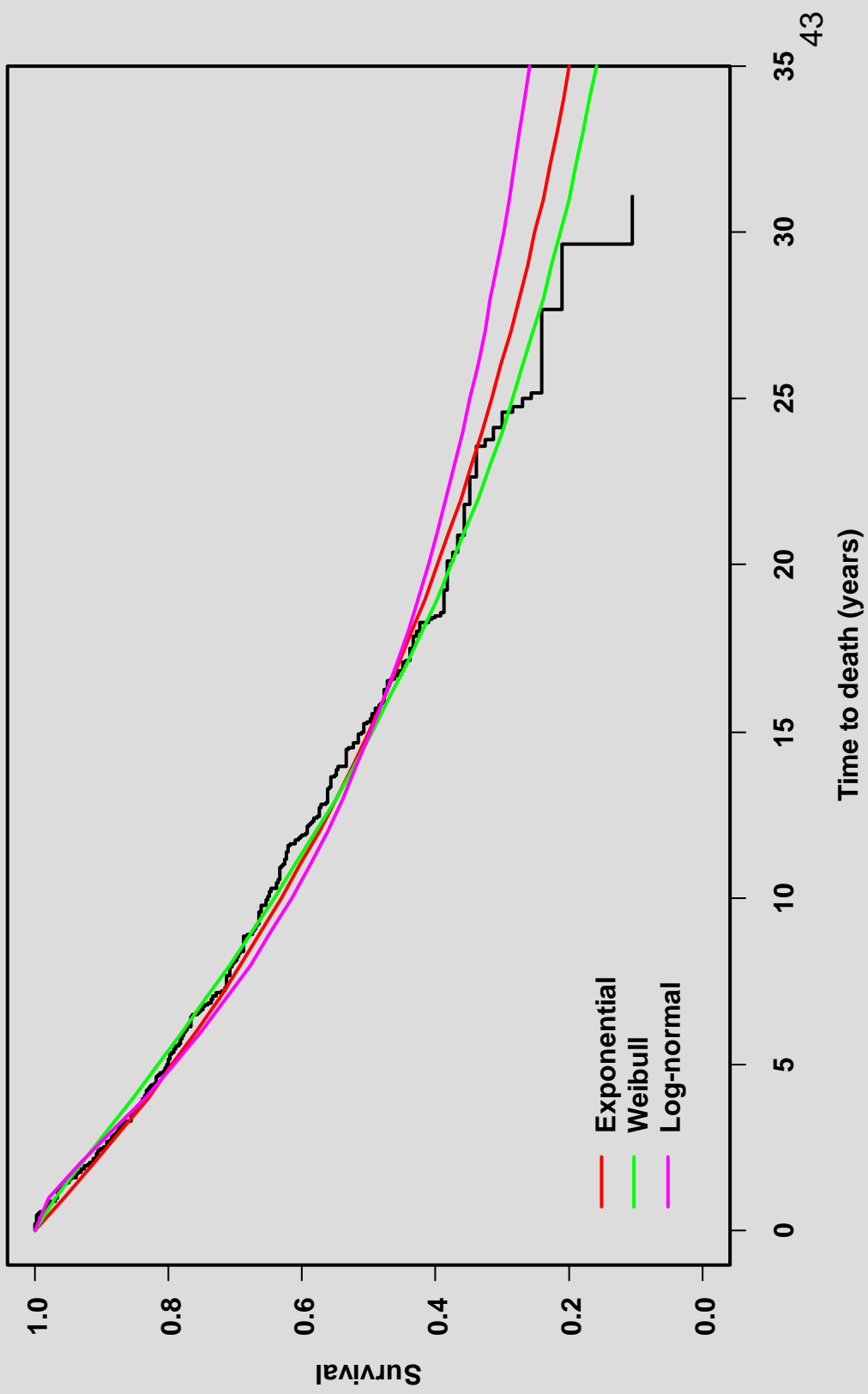
R: Estimating the survivor function: lognormal distribution

```
fitw=survreg(Surv(survtime,stat)~1,  
data=follic,dist='lognormal')  
summary(fitw)
```

	Value	Std. Error	Z	P
(Intercept)	2.695	0.0731	36.85	2.61e-297
Log(scale)	0.289	0.0472	6.13	9.02e-10

```
Scale= 1.34
```

Follicular lymphoma



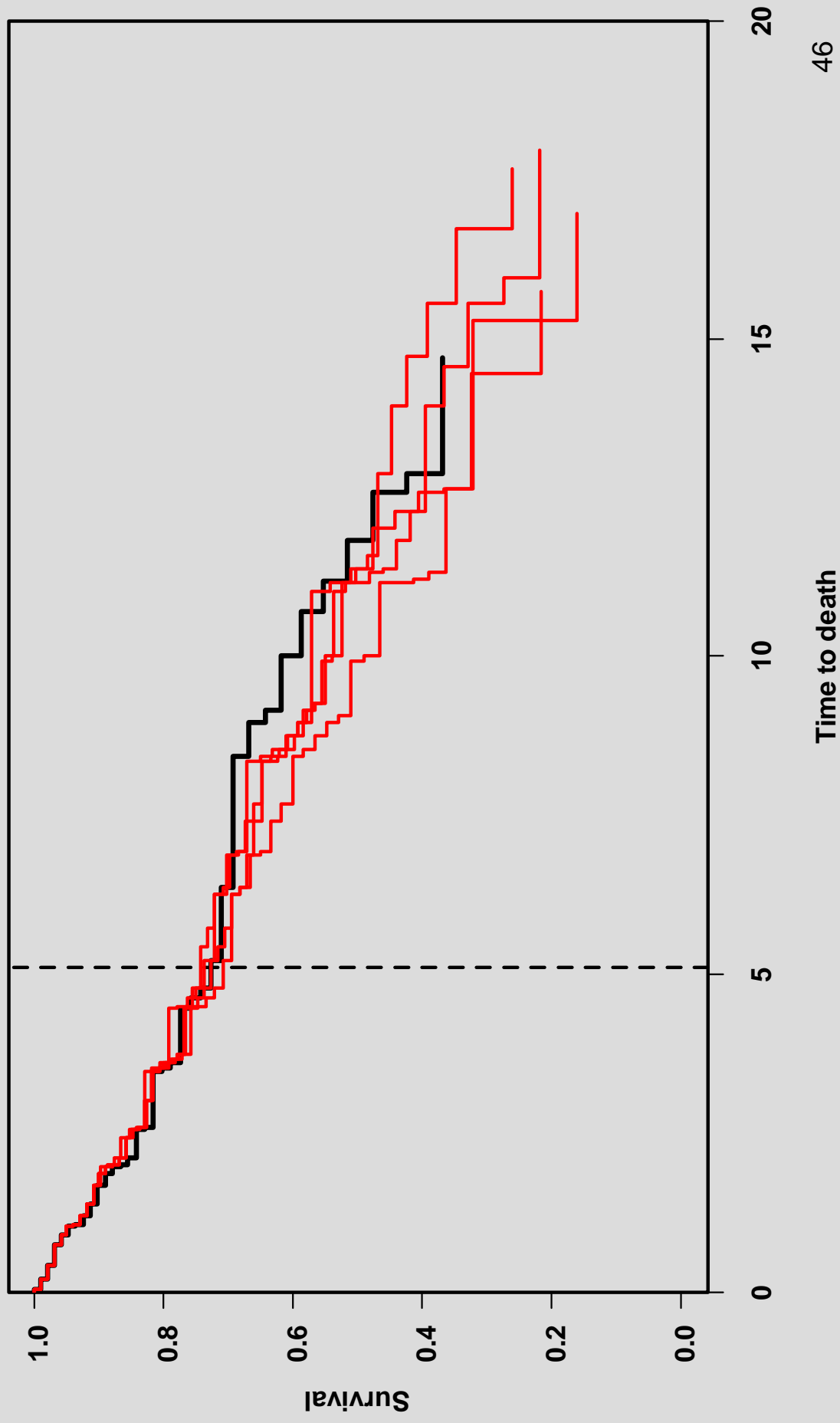
Outline

- Practical considerations
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- Hazard function
- Cox proportional hazards model

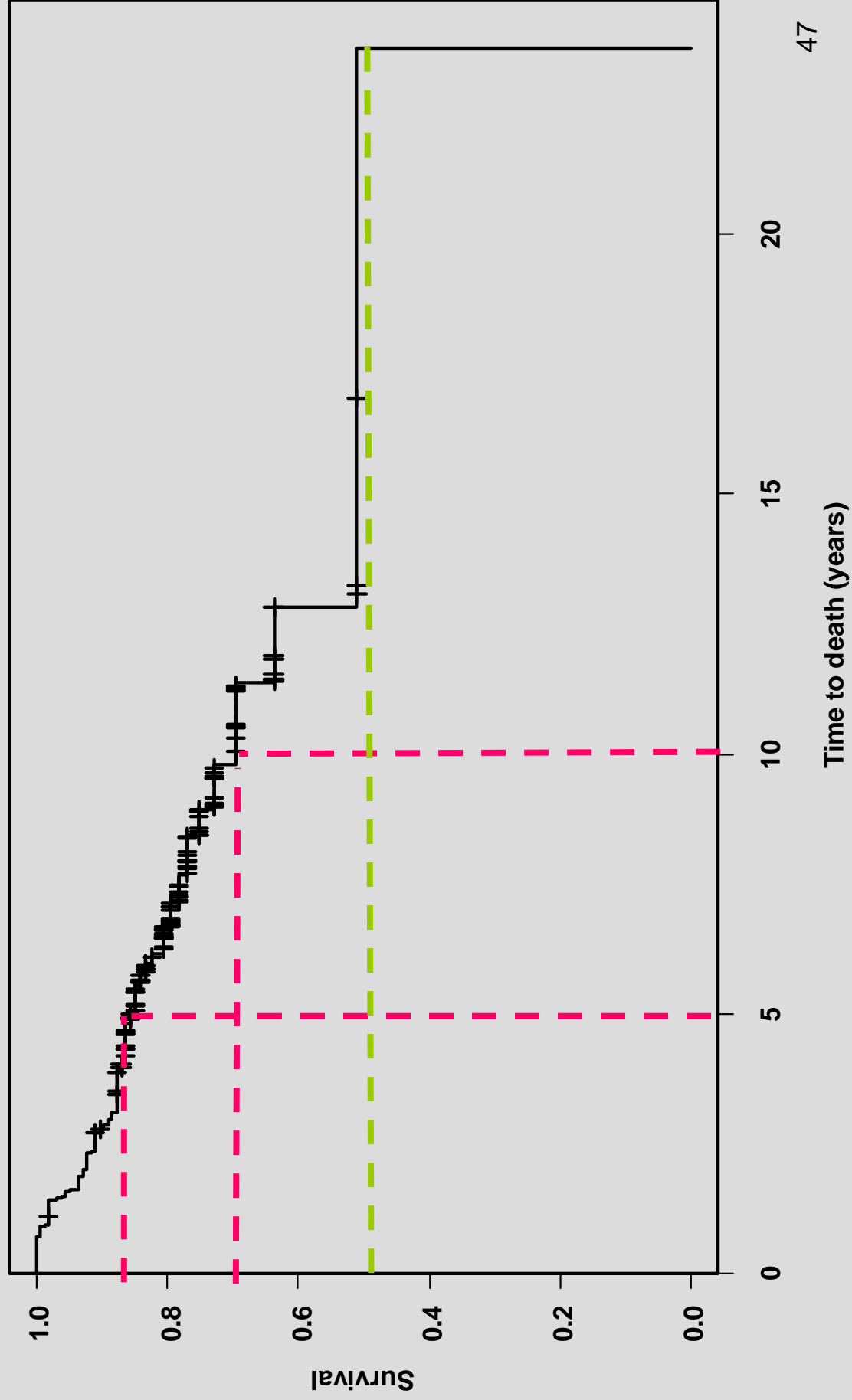
Summarizing the curve

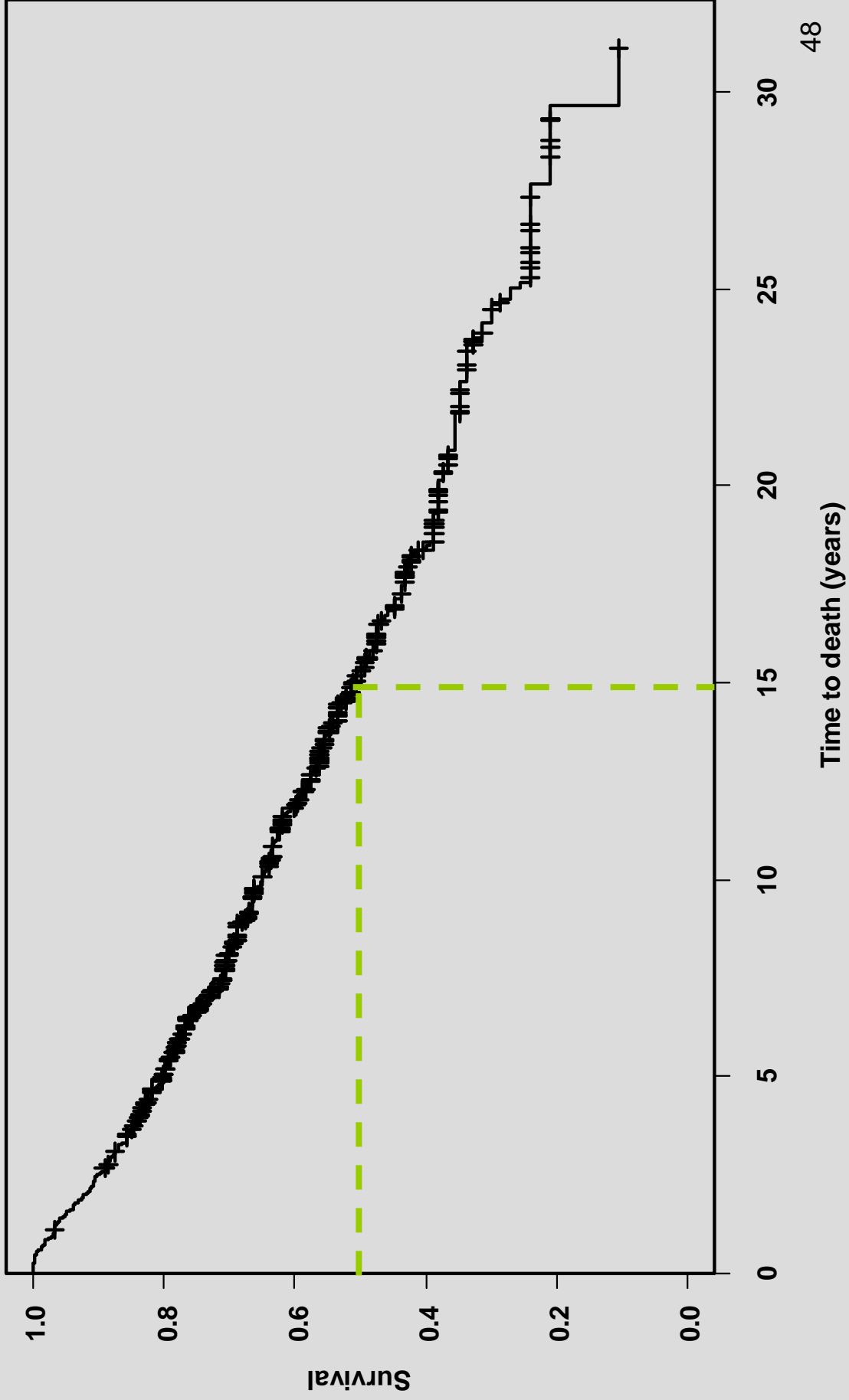
- Give the whole curve
- Survival at time at a certain time point
- Median survival = the time at which survival is 50%.
 - It may not exist
 - It may not be reliable

Accrual 15 years, Follow-up=0, 1, 2, 3, 4



Survival at 5 years = 86%
Survival at 10 years = 70%





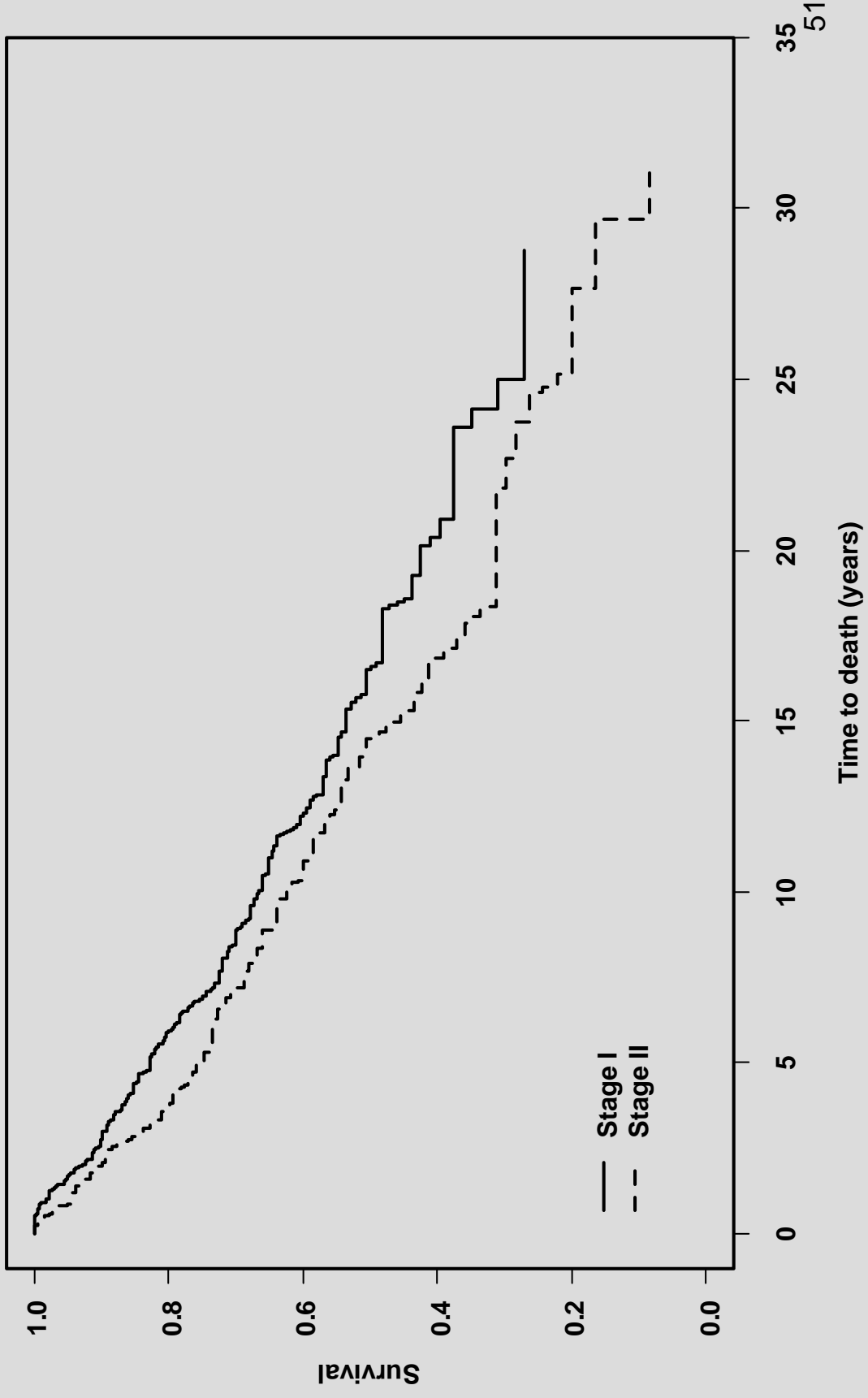
Outline

- Practical considerations
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- **Testing a covariate**
- Hazard function
- Cox proportional hazards model

Testing a covariate

- Stage = how severe is the disease at the time of diagnosis
- Are the stage II doing worse?
- Are the survivor distributions of stage I and stage II different

Follicular lymphoma



At 5 years:

```
clinstg=1
time  n.risk  n.event  survival  std.err
5.0000 281.0000 62.0000 0.8265 0.0201

clinstg=2
time  n.risk  n.event  survival  std.err
5.0000 125.0000 44.0000 0.7520 0.0325
```

At each point in time , t_j

	Alive	Dead	Total
Group 1	a_{1j}	d_{1j}	n_{1j}
Group 2	a_{2j}	d_{2j}	n_{2j}
Total	a_j	d_j	n_j

$$O - E = d_{1j} - \frac{n_{1j}}{n_j} \times d_j$$

Testing a categorical variable

Logrank test

$$\frac{U^2}{\text{Var}(U)} \sim \chi^2_{k-1} \quad k=2, \quad \chi^2_1$$

$$\begin{aligned} U &= \sum_{\text{all } t_j} \left(d_{1j} - \frac{d_j}{n_j} n_{1j} \right) \\ &= \sum_{\text{all } t_j} \left\{ n_{1j} \left(\frac{d_{1j}}{n_{1j}} - \frac{d_j}{n_j} \right) \right\} \end{aligned}$$

SAS – logrank test

```
proc lifetest data=follic plots=(s) censoredsymbol='|'  
  timelist= 0,1,2,3,4,5;  
time survtime*stat(0);strata clinstg;  
run;
```

Test	Chi-Square	DF	Chi-Square
Log-Rank	3.4200	1	0.0644
Wilcoxon	3.1678	1	0.0751
-2Log(LR)	4.2029	1	0.0404

R: Logrank test

```
fit=survdiff(Surv(survtime,stat)~clinstg,data=follic)
fit
Call:
survdiff(formula = Surv(survtime, stat) ~ clinstg,
          data = follic)

          N Observed Expected (O-E) ^2/E (O-E) ^2/V
clinstg=1 362      153      167      1.16      3.42
clinstg=2 179      103      89      2.18      3.42
```

Chisq= 3.4 on 1 degrees of freedom, p= 0.0644

Other tests:

Wilcoxon (SAS)

Wilcoxon-Peto (SAS, R)

$$\frac{U_w^2}{\text{Var}(U_w)}$$

$$U_w = \sum_{\text{all } t_j} \left\{ n_j \left\{ d_{1j} - \frac{d_j}{n_j} n_{1j} \right\} \right\}$$
$$= \sum_{\text{all } t_j} \left\{ n_j n_{1j} \left(\frac{d_{1j}}{n_{1j}} - \frac{d_j}{n_j} \right) \right\}$$

$$\frac{U_w^2}{\text{Var}(U_w)}$$

$$U_w = \sum_{\text{all } t_j} \left\{ \hat{S}(t_j) \left(d_{1j} - \frac{d_j}{n_j} n_{1j} \right) \right\}$$
$$= \sum_{\text{all } t_j} \left\{ \hat{S}(t_j) n_{1j} \left(\frac{d_{1j}}{n_{1j}} - \frac{d_j}{n_j} \right) \right\}$$

SAS – Wilcoxon and Wilcoxon-Peto

test

```
proc lifetest data=follic plots=(s) censoredsymbol='|'  
timelist= 0,1,2,3,4,5;  
time survtime*stat(0);strata clinstg/test=(wilcoxon peto);  
run;
```

Test	Chi-Square	DF	Chi-Square
Log-Rank	3.4200	1	0.0644
Wilcoxon	3.1678	1	0.0751
Peto	3.4729	1	0.0624

R: Wilcoxon-Peto test

```
fit=survdiff(Surv(survtime,stat)~clinstg,
```

```
data=follic,rho=1)
```

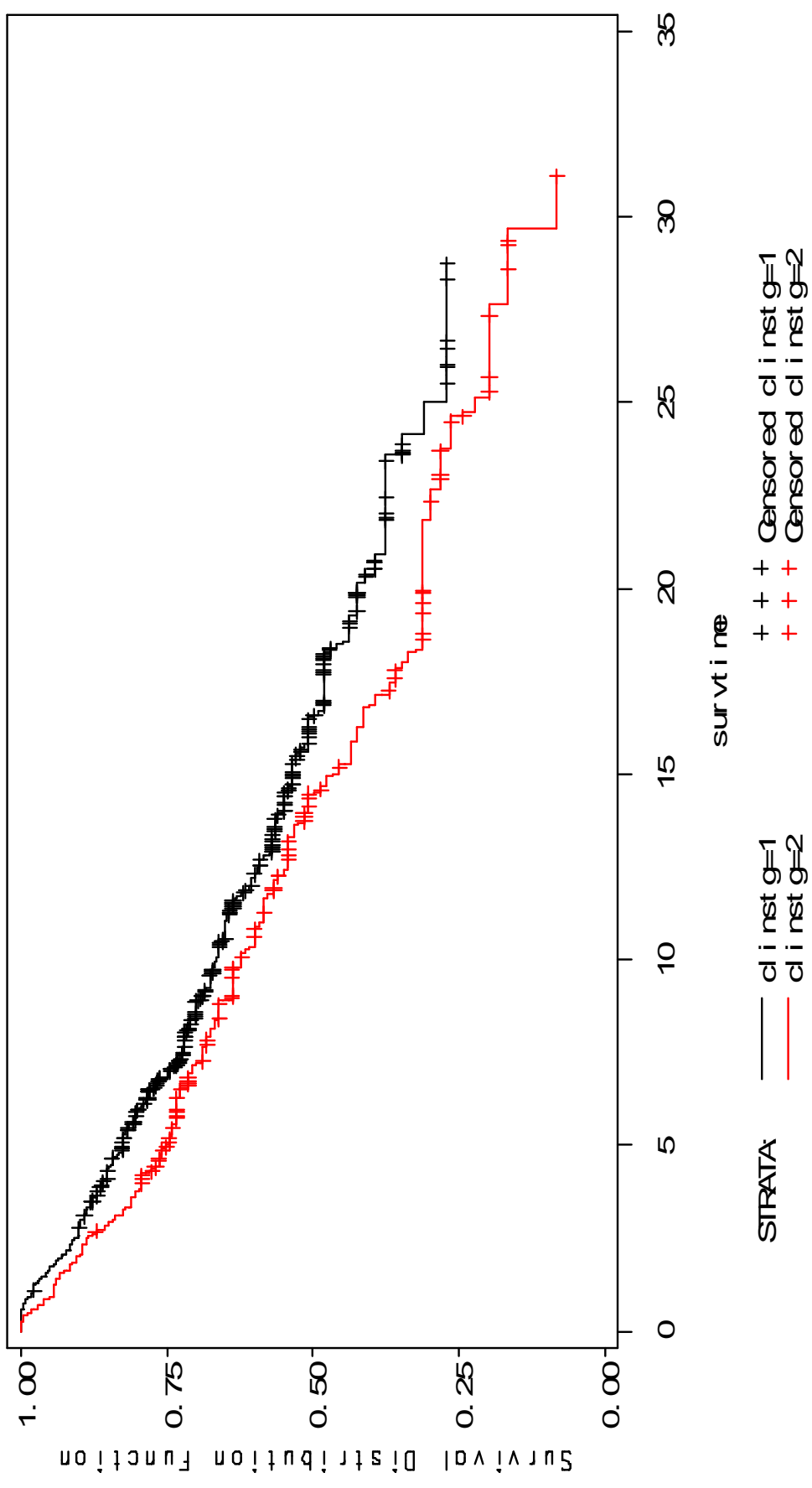
```
fit
```

```
      N Observed Expected (O-E) ^2/E (O-E) ^2/V
clin1 362    113.0    123.5    0.892    3.48
clin2 179    72.9     62.4    1.768    3.48
```

```
Chisq= 3.5 on 1 degrees of freedom, p= 0.0621
```

SAS: Graph 2 or more curves

```
proc lifetest data=follic plots=(s)
  censoredsymbol='|' timelist= 0,1,2,3,4,5;
time survtime*stat(0);strata clinstg;
run;
```



R: Graph of 2 or more curves

```
fit=survfit(Surv(survtime,stat)~clinstg,data=follic)

plot(fit,lty=c(1,2),xlim=c(-1,35), mark.time=F, conf.int=F,lwd=2,
      xlab='Time to death (years)',
      ylab='Survival',
      main='Follicular lymphoma')

legend(0,0.2,lty=c(1,2),bty='n',
       legend=c('Stage I','Stage II'))

mtext = help add elements outside the graph
text = help add elements inside the graph

text(0,0.3,adj=0,paste('logrank p-value=',0.064))
```

R: Working with *survfit* object

```
fit=survfit(Surv(survtime,stat)~clinstg,  
            data=follic)  
sfit=summary(fit)
```

Event times for stage 1: `sfit$time[sfit$strata=='clinstg=1']`

Number at risk for stage 2:

```
sfit$n.risk[sfit$strata=='clinstg=2']
```

Survival estimates for stage 2:

```
sfit$surv[sfit$strata=='clinstg=2']
```

Note: `==` as logical operator

R: Working with *survfit* object

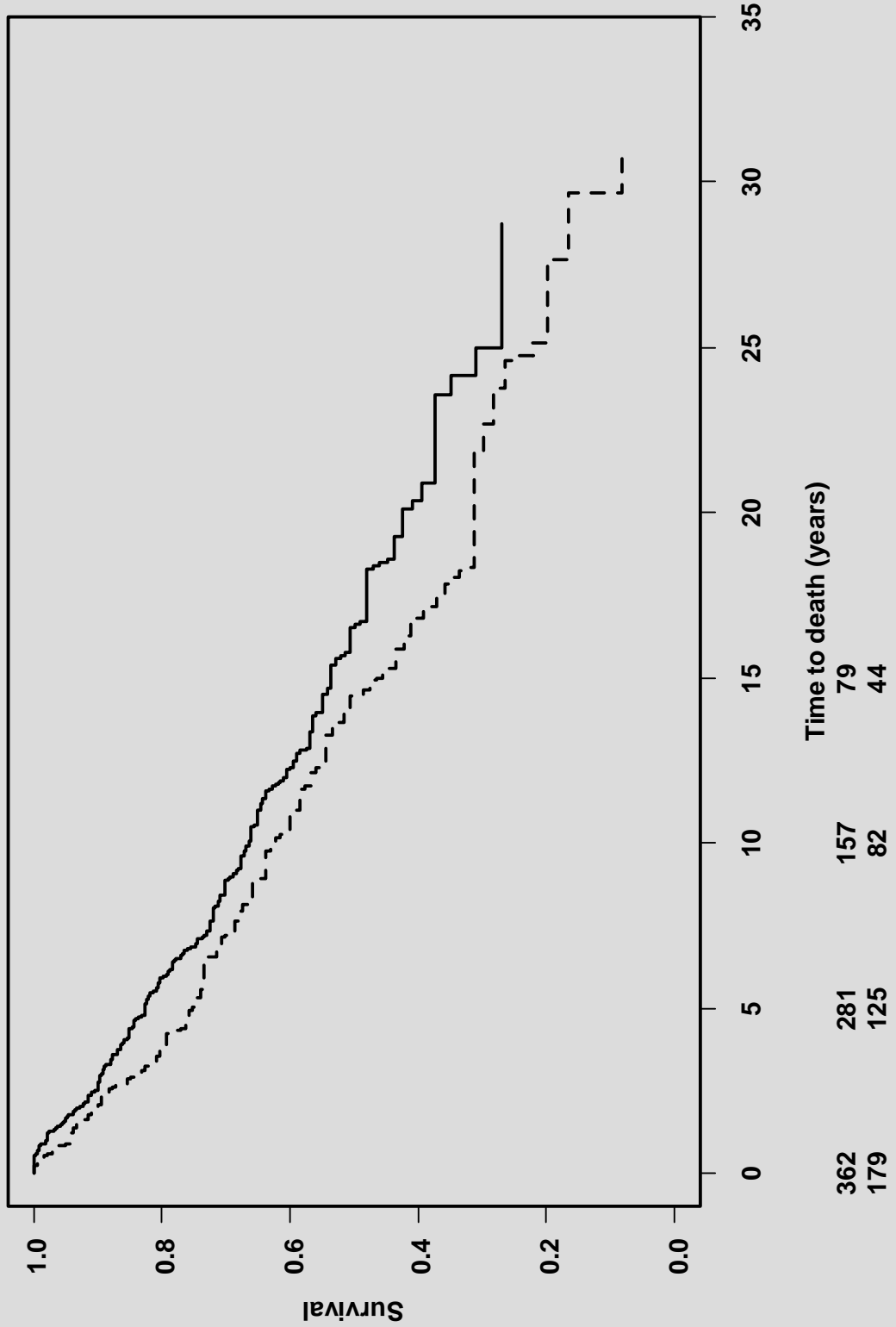
Add number of risks to the graph.

```
fit=survfit(Surv(survtime,stat)~clinstg,data=follic)
sfit=summary(fit,times=c(0,5,10,15))
nrisk1=sfit$n.risk[sfit$strata=='clinstg=1']
nrisk2=sfit$n.risk[sfit$strata=='clinstg=2']

par(las=1,mfrow=c(1,1),font=2,font.axis=2,font.lab=2,lwd=2,
    mar=c(8,5,2,2))

plot(fit,lty=c(1,2),xlim=c(-1,35),mark.time=F,conf.int=F,lwd=2,
     xlab='Time to death (years)',
     ylab='Survival',
     main='Follicular lymphoma')
mtext(nrisk1,at=c(0,5,10,15),side=1,line=4)
mtext(nrisk2,at=c(0,5,10,15),side=1,line=5)
```

Follicular lymphoma



Outline

- Practical considerations
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- **Hazard function**
- Cox proportional hazards model

Hazard

- The instantaneous death rate
- Force of mortality
- Formal definition

$$h(t) = \lim_{\delta t \rightarrow 0} \frac{P(t \leq T, t + \delta t | T \geq t)}{\delta t} = \frac{f(t)}{S(t)}$$

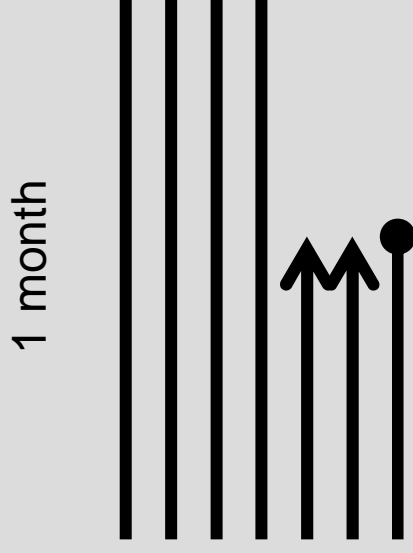
- Cumulative hazard

$$H(t) = \int_0^t h(s) ds$$

$$S(t) = e^{-H(t)}$$

Hazard non-parametric estimation

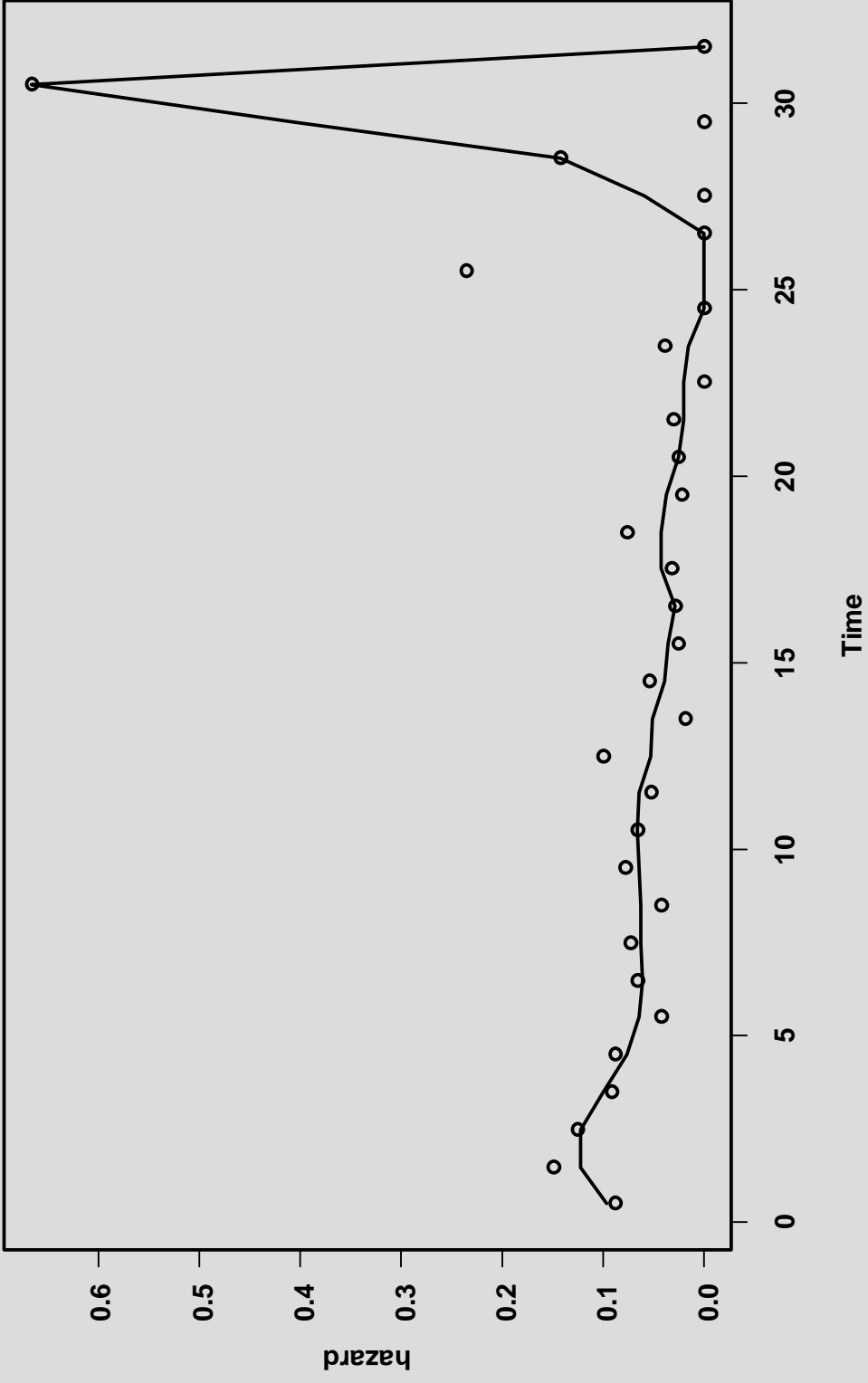
- Time axes is divided in intervals (not necessary equal)



- The j -th interval:

- δ_j = length
- n_j = number at risk at the beginning
- c_j = number censored
- d_j = number dead

$$\delta_j = 1 \quad n_j = 7 \quad c_j = 2 \quad d_j = 1$$
$$\hat{h}_j = \frac{d_j}{\delta_j \left(n_j - \frac{c_j}{2} - \frac{d_j}{2} \right)}$$



Outline

- Practical considerations
- Characteristics of survival type data
- Non-parametric estimation of the survivor function
- Parametric estimation
- Summarizing the survival curve
- Testing a covariate
- Hazard function
- Cox proportional hazards model

Cox proportional hazards modelling

$$h(t | x) = h_0(t) \exp(\beta x)$$

$$PL(\beta) = \prod_{j=1}^r \left(\frac{\exp(\beta x_j)}{\sum_{i \in R_j} \exp(\beta x_i)} \right)$$

$R_j =$ risk set at time t_j

Cox proportional hazards modelling

Semi-parametric

$$h(t | x) = h_0(t) \exp(\beta x)$$

stage = 1 or 2

$$h(t | \text{stage} = 1) = h_0(t) \exp(\beta)$$

$$h(t | \text{stage} = 2) = h_0(t) \exp(2\beta)$$

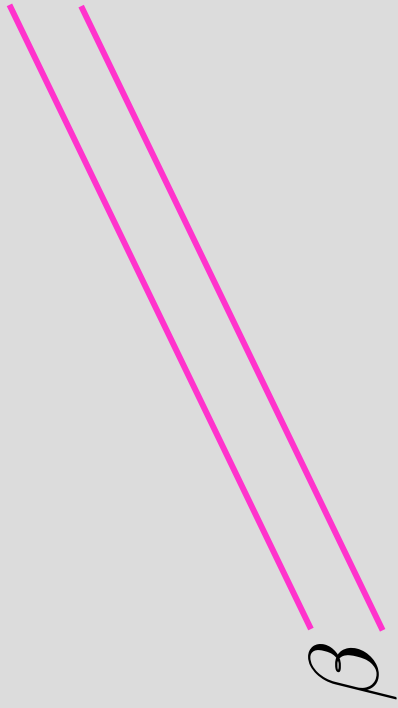
$$\exp(\beta) = \frac{h(t / \text{stage} = 2)}{h(t / \text{stage} = 1)}$$

Hazard ratio

$$\beta = \log(h(t / \text{stage} = 2)) - \log(h(t / \text{stage} = 1))$$

β (HR) does not depend on time = proportional hazards

Hazard rate, log scale



β

Time, log scale

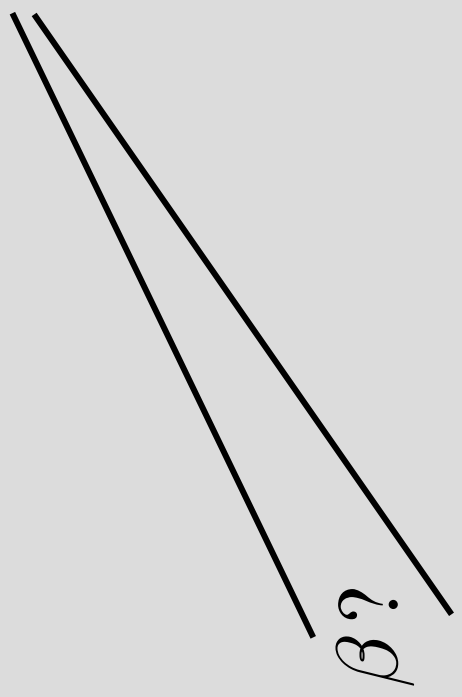
Hazard rate, log scale



β

Time, log scale

Hazard rate, log scale



$\beta?$

Time, log scale

Hazard rate, log scale



β

Time, log scale

Cox proportional hazards modelling

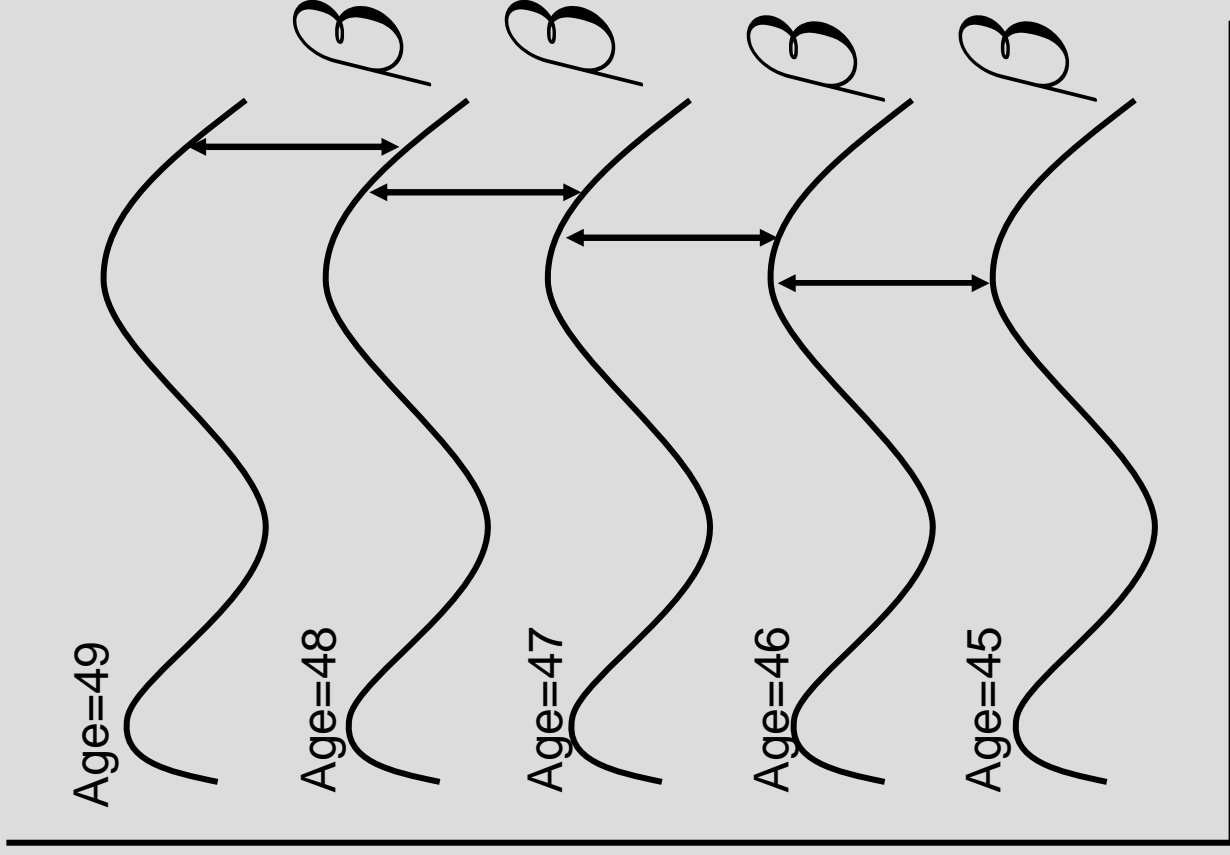
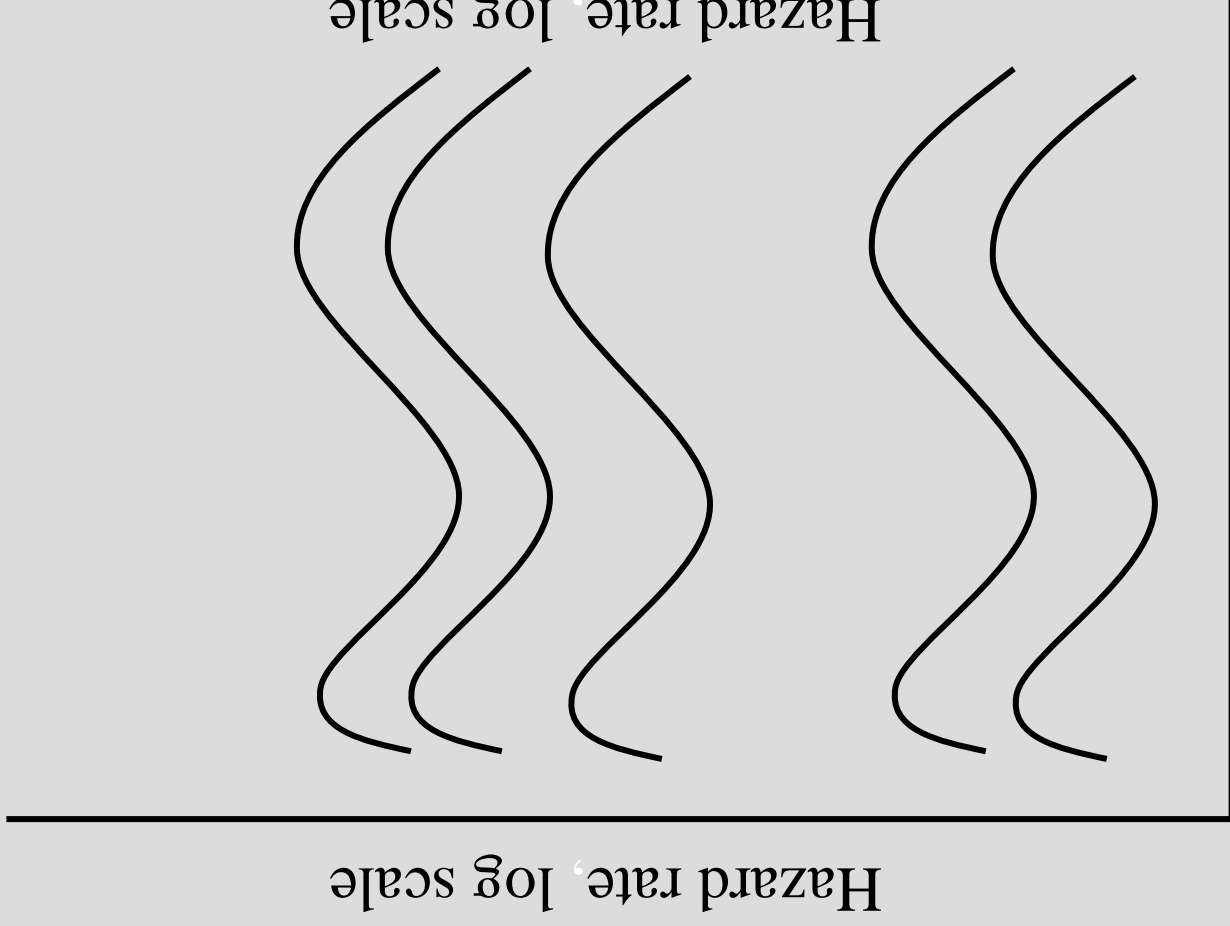
$x = \textit{continuous (age)}$

$$h(t | x = 47) = h_0(t) \exp(\beta \times 47)$$

$$h(t | x = 46) = h_0(t) \exp(\beta \times 46)$$

$$\exp(\beta) = \frac{h(t / x = 47)}{h(t / x = 46)} = \frac{h(t / x = a + 1)}{h(t / x = a)}$$

HR = fold increase of the hazard for 1 unit increase of x =>log hazard and x are linear =>linearity



Time (log scale)

Time (log scale)

SAS: Cox proportional hazards model

```
proc phreg data=follic;  
model survtime*stat(0)= clinstg chn/ties=efron ri;  
run;
```

Where age<30;

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
clinstg	1	0.25227	0.13163	3.6729	0.0553
chn	1	-0.09136	0.17570	0.2704	0.6031

β

Analysis of Maximum Likelihood Estimates

Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits
----------	--------------	------------------------------------

clinstg	1.287	0.994 1.666
chn	0.913	0.647 1.288

$HR = e^{\beta}$

R: Cox proportional hazards model

```
coxph(Surv(survtime,stat)~clinstg+ch,data=follic)
```

Call:

```
coxph(formula = Surv(survtime, stat) ~ clinstg + ch, data = follic)
```

	coef	exp(coef)	se(coef)	z	p
clinstg	0.2523	1.287	0.132	1.92	0.055
<u>ch</u>	-0.0914	0.913	0.176	-0.52	<u>0.600</u>

```
Likelihood ratio test=3.62 on 2 df, p=0.164 n= 541
```

ch is a factor variable (Y/N). The coefficient is $h(Y)/h(N)$

p-value is given with 2 significant digits.

The first is 0.55 the second is 0.60.

R: Working with *coxph* objects

```
fit=coxph(Surv(survtime,stat)~clinstg+ch,data=follic)
```

```
fit=coxph(Surv(survtime,stat)~clinstg+ch,data=follic,
```

```
subset=(age<30))
```

```
summary(fit)
```

```
summary(fit)$coef
```

```
summary(fit)$conf.int
```

```
fit$coef
```

```
fit$var
```

Unit of measure

```
> summary(hgb1)
  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
  4.00  13.00  14.00  13.86  15.00  18.90
> fit=coxph(Surv(survtime,stat)~hgb1,data=follic)
> summary(fit)$coef
      coef      exp(coef)    se(coef)      z      p
hgb1 -0.02266176  0.977593  0.04280909 -0.5293678  0.6

> summary(follic$hgb)
  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
  40.0  130.0  140.0  138.6  150.0  189.0
> fit=coxph(Surv(survtime,stat)~hgb,data=follic)
> summary(fit)$coef
      coef      exp(coef)    se(coef)      z      p
hgb -0.002266176  0.9977364  0.004280909 -0.5293678  0.6
```

Categorical variables

```
table(follic$blktxcat)
0    1    2    3
199 104 137 101
```

0=no bulk
1=small bulk
2=medium bulk
3=large bulk

```
fit=coxph(Surv(survtime,stat)~clinstg+ch+blktxcat, data=follic)
      coef exp(coef) se(coef)      z      p
clinstg  0.0465    1.048    0.1397  0.333  7.4e-01
chY      -0.2712    0.762    0.1806 -1.502  1.3e-01
blktxcat  0.2510    1.285    0.0604  4.153  3.3e-05
```

SAS: analyzing bulk

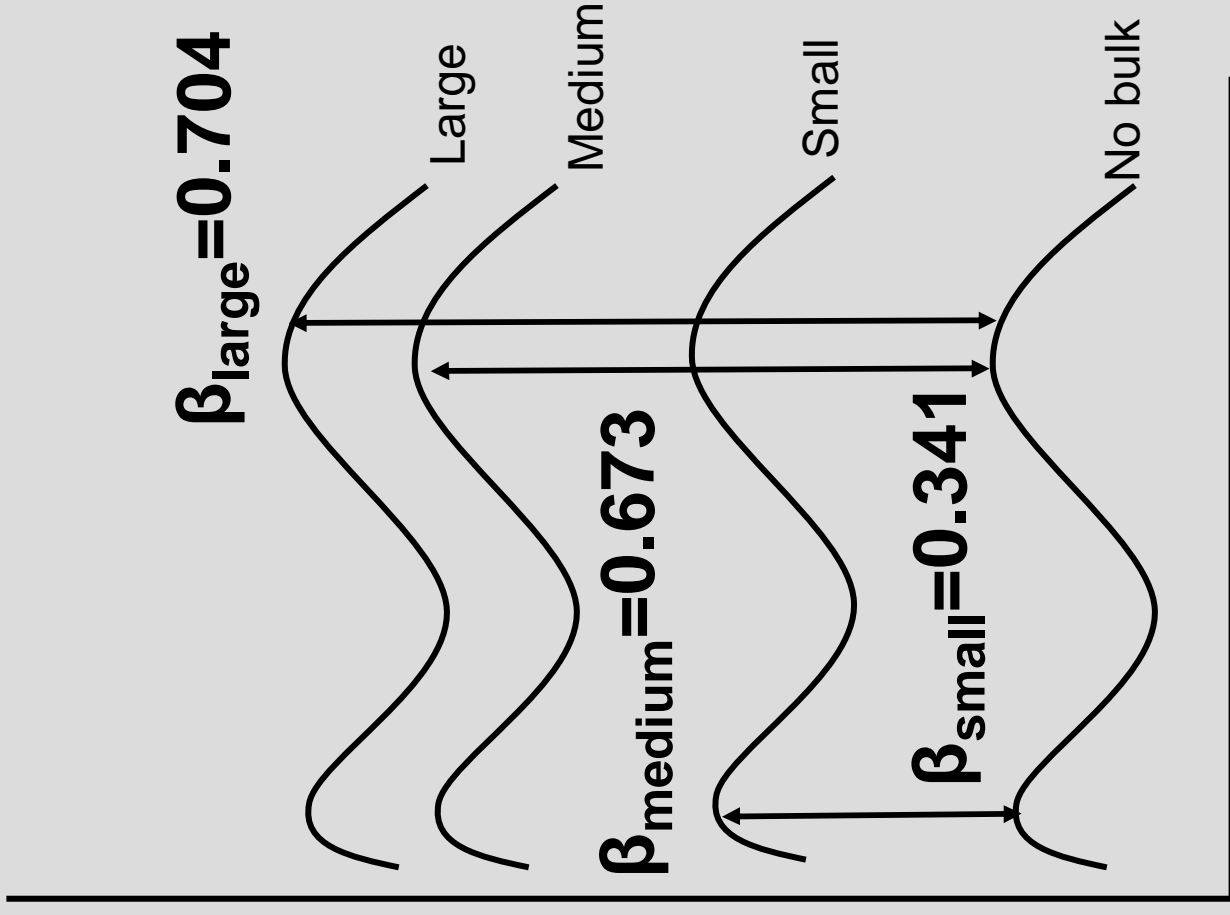
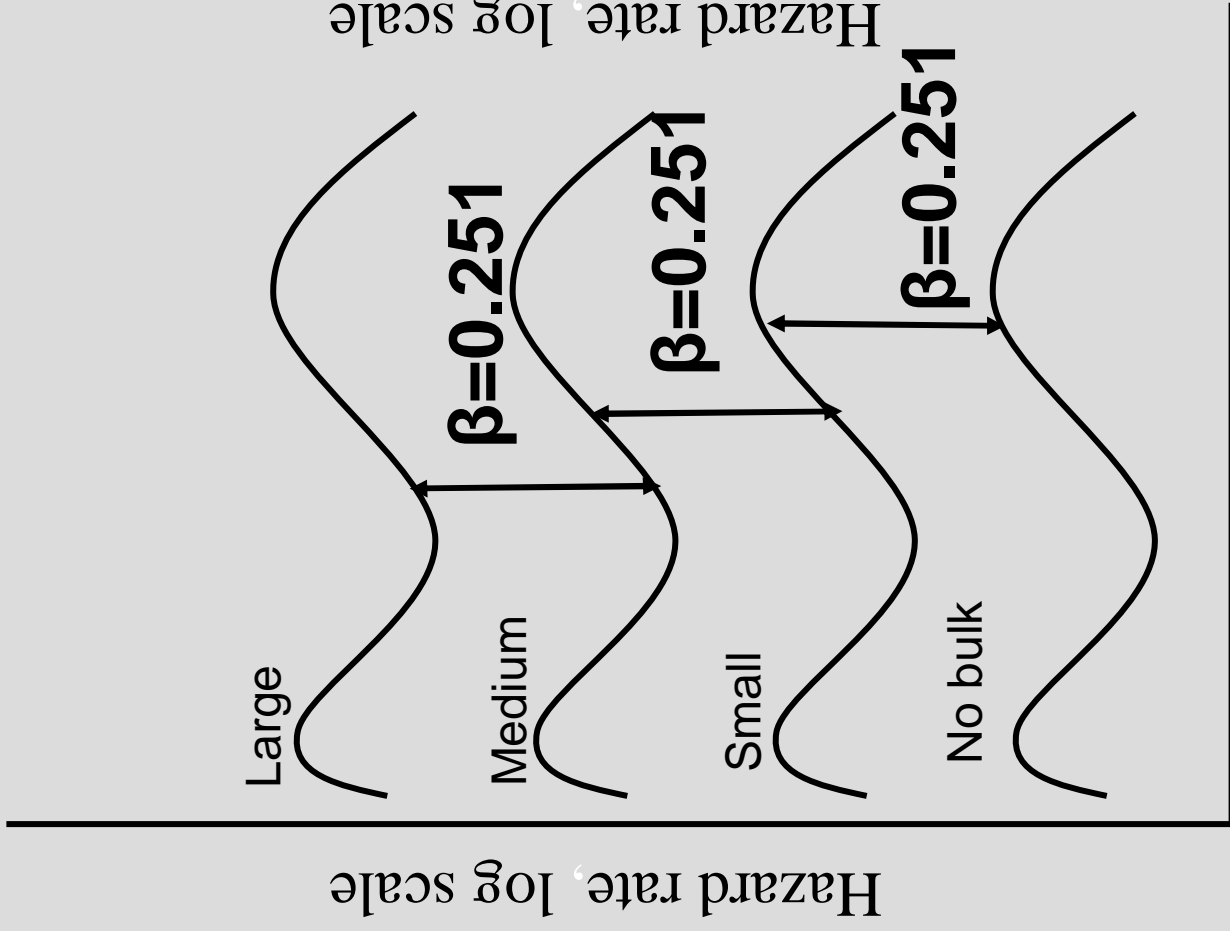
```
if blktxcat=0 then do;small=0;medium=0;large=0;end;  
if blktxcat=1 then do;small=1;medium=0;large=0;end;  
if blktxcat=2 then do;small=0;medium=1;large=0;end;  
if blktxcat=3 then do;small=0;medium=0;large=1;end;
```

```
bulk: test small, medium, large;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr >	ChiSq	Hazard Ratio
clinstg	1	0.02116	0.14124	0.0224	0.8809	1.021	
chn	1	-0.23416	0.18289	1.6393	0.2004	0.791	
small	1	0.34057	0.18649	3.3351	0.0678	1.406	
medium	1	0.67329	0.17448	14.8907	0.0001	1.961	
large	1	0.70379	0.19318	13.2731	0.0003	2.021	

Linear Hypotheses Testing Results

Label	Wald		
	Chi-Square	DF	Pr > ChiSq
bulk	18.4575	3	0.0004



R: analyzing bulk

```
fit=coxph(Surv(survtime,stat)~clinstg+ch+as.factor(blktxcat),data=follic)
```

```
wald.test(fit$var,fit$coef, Terms=c(3,4,5))
```

```
wald.test in library aod
```

```
Load library beforehand: library(aod)
```

Outline

- Parametric modelling
- Sample size
- Competing risks: justification and definition
- Non-parametric estimation of the probability of event in the presence of competing risks
- Reporting

Parametric modelling: exponential

$$S(t) = e^{-\frac{t}{\theta}} \quad \theta = e^{\beta} \quad \beta = \text{intercept}$$

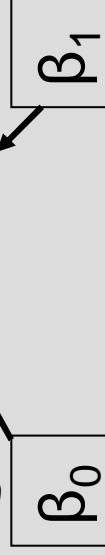
$$S(t | x) = \exp\left(-\frac{t}{\theta(x)}\right), \quad \theta(x) = \exp(\beta_0 + \beta_1 x)$$

β_0 and β_1 = coefficients

SAS/R: exponential distribution

```
proc lifereg data=follic;  
model survtime*stat(0)=clinstg/dist=exponential;  
run;  
  
> fit=survreg(Surv(survtime,stat)~clinstg,data=follic,dist='exponential')  
> summary(fit)
```

```
Call:  
survreg(formula = Surv(survtime, stat) ~ clinstg, data =  
follic,  
dist = "exponential")  
Value Std. Error z p  
(Intercept) 3.438 0.189 18.15 1.18e-73  
clinstg -0.264 0.127 -2.07 3.84e-02
```



Outline

- Parametric modelling
- **Sample size**
- Competing risks: justification and definition
- Non-parametric estimation of the probability of event in the presence of competing risks
- Reporting

Sample size calculation

HR = hazard ratio
to be detected
 σ = standard deviation
of the covariate
 $Z_{1-\alpha/2}$ = quantile of the
standard normal

$$\sqrt{n_{ev}} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})}{\sigma \ln(HR)}$$

$$P_{ev} = 1 - \frac{e^{-\lambda f} - e^{-\lambda(f+a)}}{\lambda a}$$

$$N = \frac{n_{ev}}{P_{ev}}$$

a = accrual time
f = follow-up time
 λ = the hazard rate
for the overall

Assumption: Exponential distribution

Example 1

- $N=150$
- $n_{ev} = 70$
- Covariate: p53, measured IHC
 - 40% negative ($p=0.4$)
 - 60% positive
- Assume: $\alpha=0.05$, $\beta=0.2$ (power=0.8)
- Calculate effect size which could be detected under these conditions

Example 1

The screenshot shows the 'PASS: Regression: Cox' software interface. The 'Data' tab is active, displaying various input parameters and calculated values. The 'Find (Solve For):' dropdown is set to 'B'. The 'Hypothesis Test:' is 'Two-Sided'. The 'N (Sample Size):' is 150. The 'Alpha (Significance Level):' is 0.05. The 'Beta (1-Power):' is 0.20. The 'S (Std Deviation of X1):' is 0.49. The 'R-Squared Other X's:' is 0.0. The 'P (Overall Event Rate):' is 0.467. The 'B (Log Hazard Ratio):' is 0.2. The 'SD' dropdown is set to 'SD'. The 'Template Id:' field is empty. The 'Template' tab is also visible, showing a message: 'Run this procedure with the current settings and view the output report and plots.' Below the message are 'Reset', 'Next', and 'Last' buttons.

Find (Solve For): B

Hypothesis Test: Two-Sided

N (Sample Size): 150

Alpha (Significance Level): 0.05

Beta (1-Power): 0.20

S (Std Deviation of X1): 0.49

R-Squared Other X's: 0.0

P (Overall Event Rate): 0.467

B (Log Hazard Ratio): 0.2

SD

Template Id: _____

Run this procedure with the current settings and view the output report and plots.

Reset Next Last

$$\frac{n_{ev}}{N} = \frac{70}{150} = 0.467$$

$$p(1-p) = \sqrt{0.4 \times 0.6} = 0.49$$

Example 1

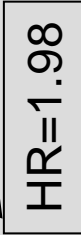
	Sample Size	Reg. Coef. (B)	S.D. of X1 (SD)	Event Rate (P)	R-Squared X1 vs Other X' (R2)	Two-Sided Alpha	Beta
Power	0.80000	0.6831	0.4900	0.4670	0.0000	0.05000	0.20000

Refereces

Report definitions

Summary statements

Plot



HR=1.98

Example 2

- $a=7$
- Accrual rate 60/year
- $f=5$
- Half are randomized to a new treatment
- S at 3 years for new treatment=0.65
- S at 3 years for standard treatment =0.55
- Assume: $\alpha=0.05$
- Calculate power

Example 2

Find (Solve For): Beta and Power
Proportion Surviving Past T0:
 S1 0.65
 S2 0.55
R (Accrual Time): 7
% Time Until 50% Accrual: 50
Follow-Up Time, T-R: 5
Proportion Lost to Follow-Up:
 1 0 2 0

Alternative Hypothesis:
 Ha: S1 <> S2
T0 (Fixed Time Point): 3
N (Total Sample Size): 420
Proportion in Group 1: 0.5
Alpha (Significance Level): .05
Beta (1-Power): 0.20

Template Id: _____

Power=0.82

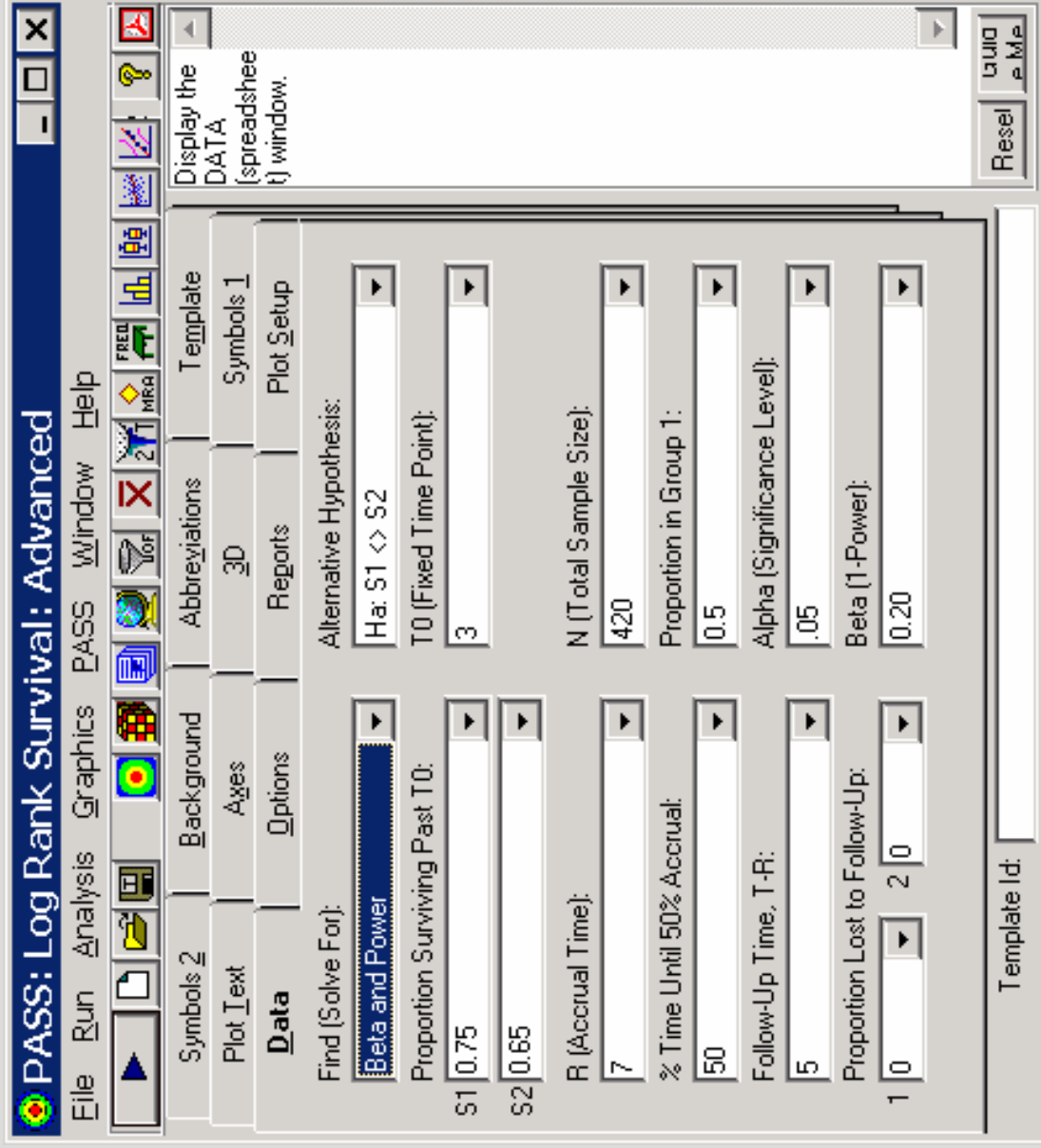
$$HR = \frac{\log(S2)}{\log(S1)} = 1.39$$

Example 2

S at 3 years in standard treatment = 0.65

Power = 0.90

$$HR = \frac{\log(S2)}{\log(S1)} = 1.5$$

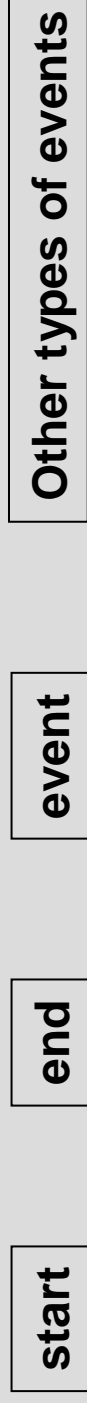


Outline

- Parametric modelling
- Sample size
- **Competing risks: justification and definition**
- Non-parametric estimation of the probability of event in the presence of competing risks
- Reporting

Other endpoints

- Survival: time to death
- Disease free survival: time to first failure (relapse or death)
- Progression free survival
- Progression free rate
- Local relapse rate
- Second malignancy rate



Example follicular lymphoma

- Follicular lymphoma= type of cancer
- Early stage (disease is not spread), high cure rates, long term side effects of the treatment
- Event of interest is second malignancy
- Other events: death without second malignancy (most likely from lymphoma)

Example follicular lymphoma

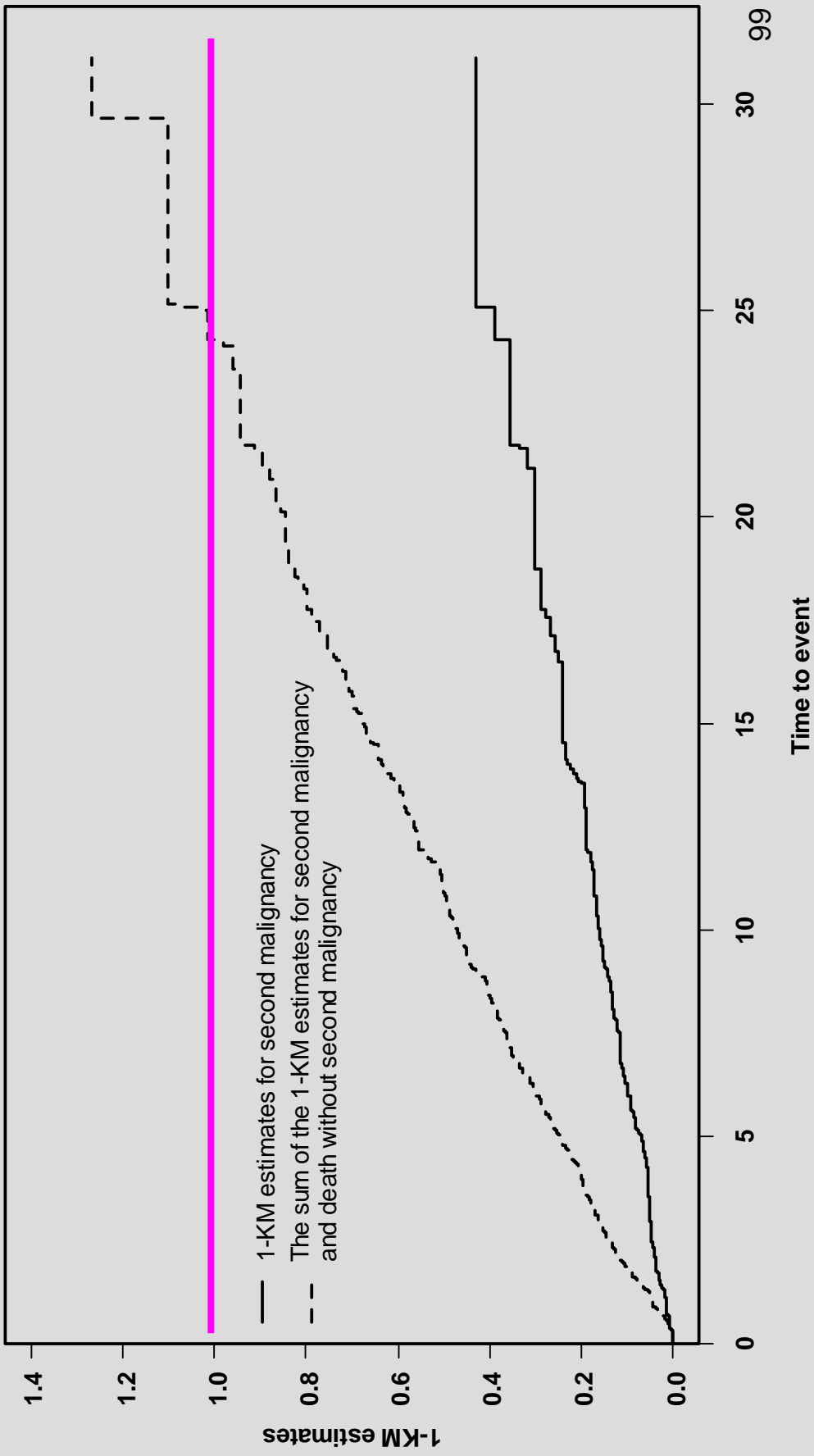
- Time to second malignancy or death or last follow-up
- Censor variable: 1 for second malignancy and 0 for death without second malignancy or alive

Problem 1

- Those who die without second malignancy will never experience the event of interest

KM (t) estimated probability of surviving beyond t
1-KM estimated the probability of death before t

Problem 2
The 1-KM for second malignancy is not a probability



Definition

Competing risks type of event=
the event whose occurrence either precludes the occurrence of another event under investigation or fundamentally alters the probability of occurrence of this other event.

Gooley, TA; Leisenring, W; Crowley, J; Storer, BE,

"Estimation of failure probabilities in the presence of competing risks: new representations of old estimators" *Statistics in*

Medicine 1999 pp. 695-706

Definition, Examples

the event whose occurrence either precludes the occurrence of another event under investigation

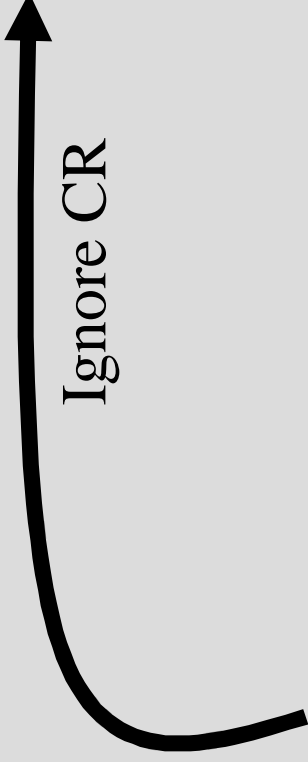
- **Death: due to disease (MI) and due to other causes**
- **First relapse: local relapse and distant relapse**

fundamentally alters the probability of occurrence of this other event

Example

- Event of interest: Death to MI
- CR event: Death of other causes
- Alive

Censor CR event
Apply usual
survival techniques



Code CR event
Apply specific
techniques



SAS macros

- http://www.uhnres.utoronto.ca/labs/hill/People_Pintilie.htm
- cuminc – estimates probability of the event of interest
- compcif
 - estimates for the probability of event of interest
 - estimates the probability for the competing risk event
 - compares the probabilities of the event of interest between 2 groups (Pepe and Mori, 1993)
- compcp - compares the conditional probabilities of the event of interest between 2 groups
 - estimates for the probability of event of interest and the competing risk event
 - estimates for the conditional probability of event of interest
 - compares the conditional probabilities of the event of interest between 2 groups (Pepe and Mori, 1993)

Software: *cmprsk* in R

<http://www.r-project.org> (CRAN)

- `cuminc` = estimates and compares CIFs
- `plot.cuminc`
- `timepoints` = gives estimates at certain points
- `crr` = modeling
- `predict.crr` = based on the hazard of subdistribution
- `plot.predict.crr`

R functions

http://www.uhnres.utoronto.ca/labs/hill/People_Pintilie.htm

- compCIF
- cifDM
- compCP
- CPvar
- btvarCP2
- plot.cp
- kly
- power

Outline

- Parametric modelling
- Sample size
- Competing risks: justification and definition
- Non-parametric estimation of the probability of event in the presence of competing risks
- Reporting

Estimation of the probability of the event of interest

$t_1 < t_2 < \dots < t_r$ Ordered time points of **all types** of events

$d_{ev j}$ Number of events of interest at time t_j

n_j Number at risk just before t_j

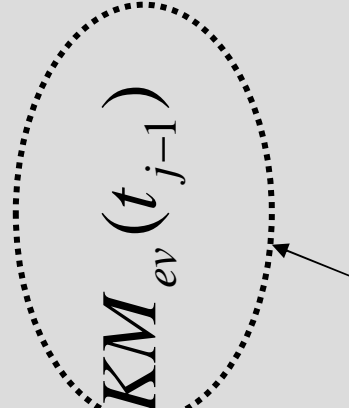
$S(t_j)$ Probability free of any event at time t_j

$\hat{F}_{ev}(t) = \sum_{all\ j, t_j \leq t} \frac{d_{ev j}}{n_j} \hat{S}(t_{j-1})$ $\hat{S}(t_j)$ = KM estimate for all types of events

Non-parametric estimation Kaplan-Meier vs. CIF

$$\hat{F}_{ev}(t) = \sum_{all\ j, t_j \leq t} \frac{d_{ev\ j}}{n_j} \hat{S}(t_{j-1})$$

← Based on all types of events

$$1 - KM_{ev}(t) = \sum_{all\ j, t_j \leq t} \frac{d_{ev\ j}}{n_j} KM_{ev}(t_{j-1})$$


Based on events of interest only

Hodgkin's disease cohort

Hodgkin's disease is a type of cancer which affects the young population.

Survival (of all causes, early stage disease) at 15 years = 71%

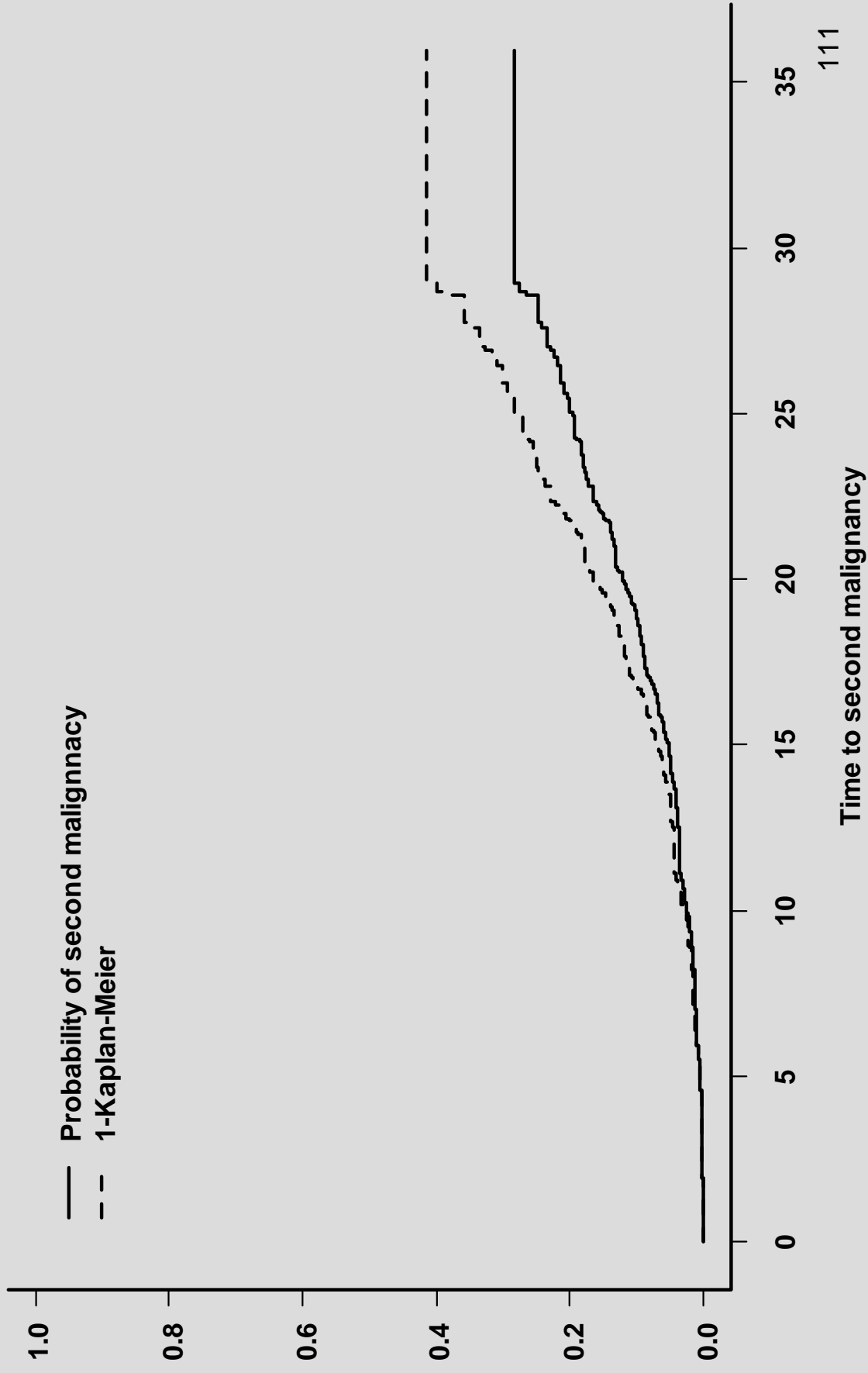
General question

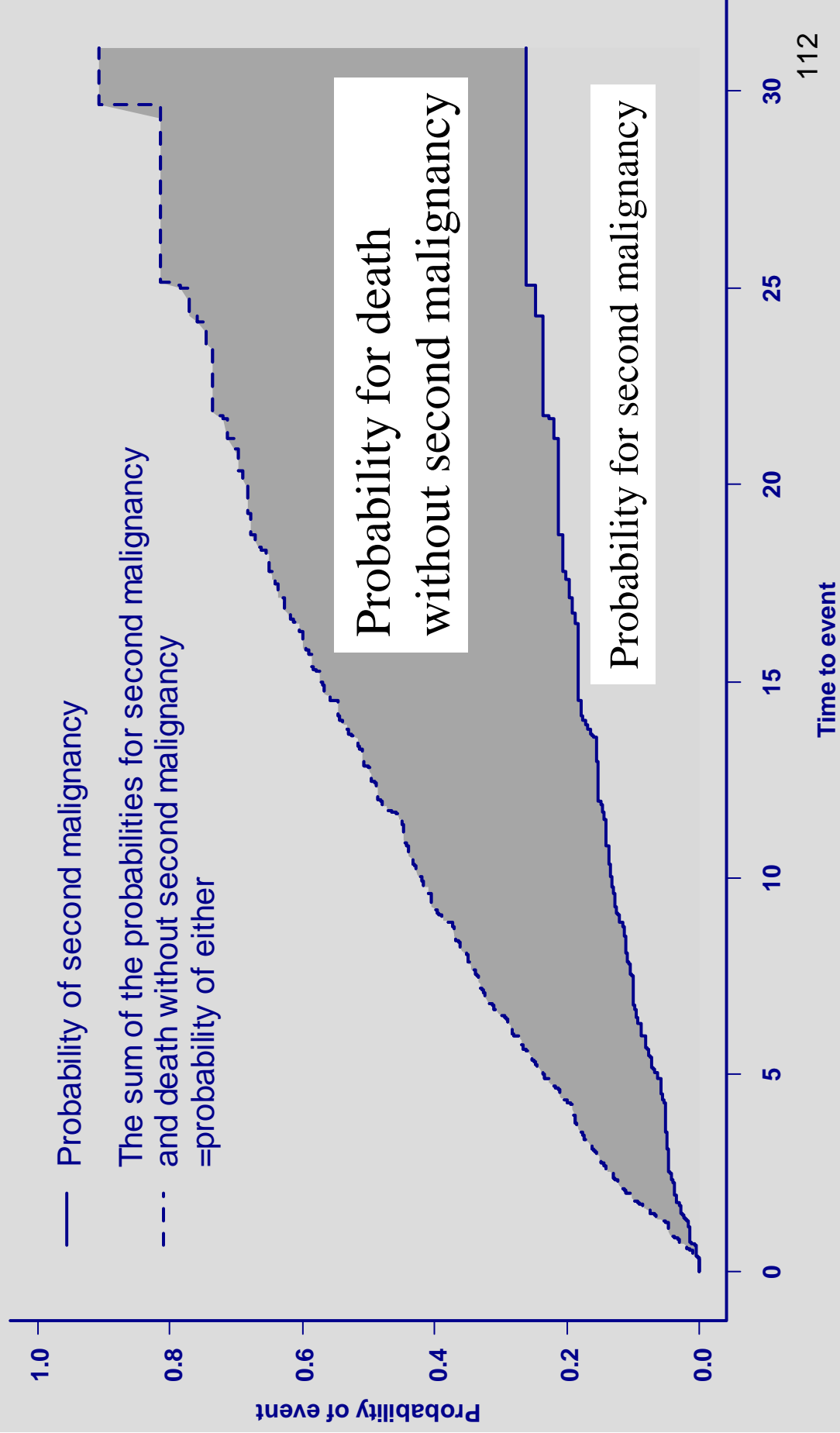
- The long term side effects of treatment, specifically the risk for other malignancies

Hodgkin's disease cohort

- Time= diagnosis to second malignancy or last follow-up
- Censor variable:
 - 1=second malignancy observed
 - 2=death without second malignancy
 - 0= alive and well

1-Kaplan-Meier > Probability of second malignancy





SAS: estimating the probability for second malignancy

```
%include 'c:/your_directory/cuminc.txt';  
data hd;set hd;  
cens=(mcens=1)+ 2*(mcens=0 and stat=1);  
group=0;  
%cuminc(ds=hd,time=time,cenvble=cens,in  
terest=1,group=group);  
run;
```

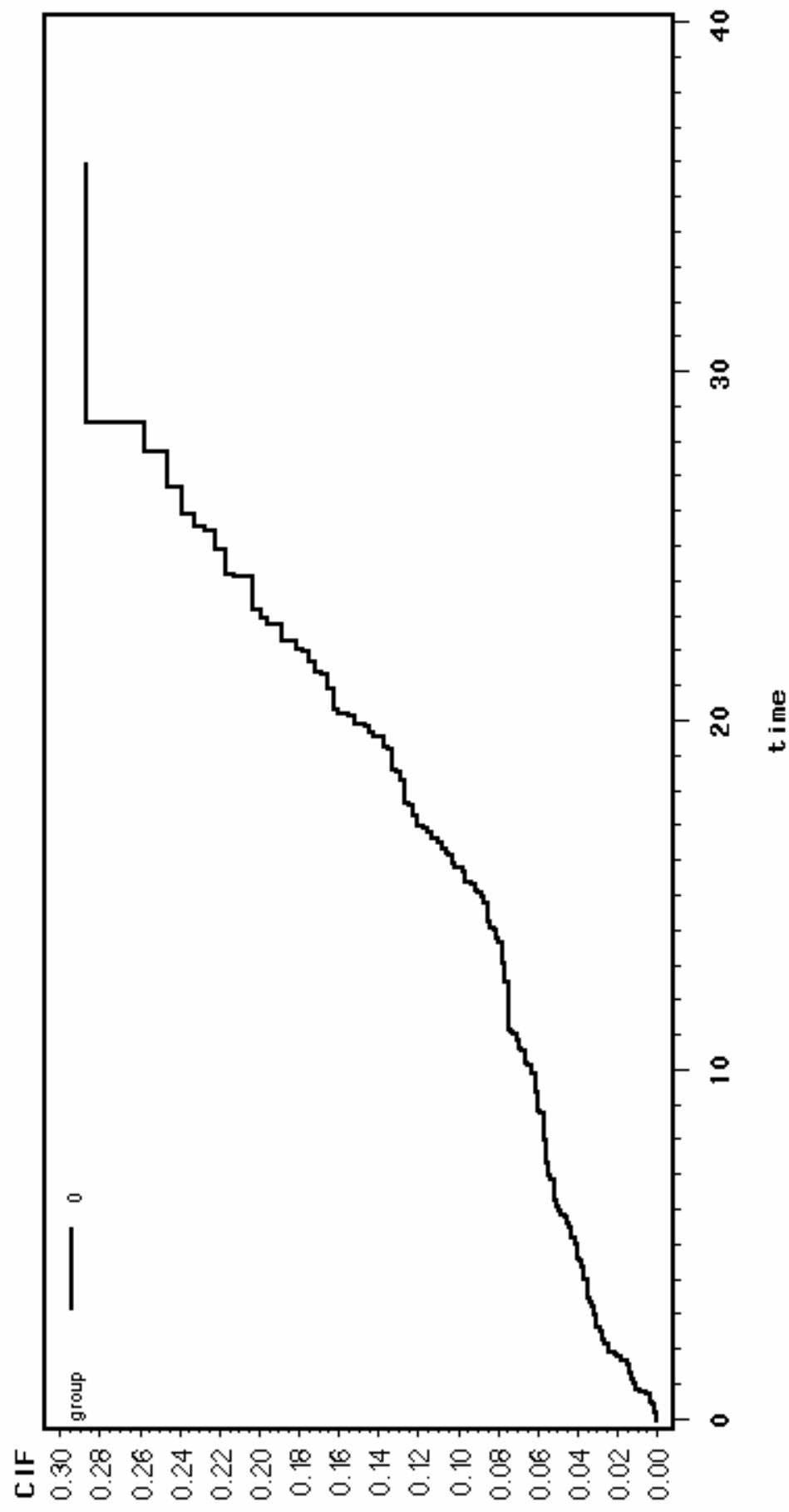
SAS: estimating the probability for second malignancy

group=0

time	Number left	Number of type 1 events	Total number of events	CIF for type 1 events	Variance for CIF type 1 events	CP for type 1 events	Variance for CP type 1 events
0.003	837	0	26	0.000000	0.000000	0.000000	0.000000
0.06	836	0	1	0.000000	0.000000	0.000000	0.000000
0.208	835	0	1	0.000000	0.000000	0.000000	0.000000
0.222	834	1	1	0.001159	0.000001	0.001198	0.000001
0.252	831	0	1	0.001159	0.000001	0.001199	0.000001
0.331	830	0	1	0.001159	0.000001	0.001200	0.000001

2.962	646	0	1	0.030010	0.000035	0.035797	0.000050
2.982	645	0	1	0.030010	0.000035	0.035850	0.000050
3.001	642	1	1	0.031265	0.000036	0.037350	0.000052
3.025	641	0	1	0.031265	0.000036	0.037406	0.000052
3.102	640	0	1	0.031265	0.000036	0.037462	0.000052
3.233	637	0	1	0.031265	0.000036	0.037519	0.000052

CIF for event 1 by group



Example in head and neck cancer

- 129 patients diagnosed with head and neck cancer
- Treated with radiation
- Side effects: inability to swallow. To prevent loss of weight a G-tube is inserted. This is a temporary measure. The G-tube can be taken out when the condition improves.

Goal:

To estimate the probability of patients still using the G-tube after say 1 year.

Example in head and neck cancer

- Assume that the G-tube is inserted at the time of starting radiation
- Some patients deteriorate rapidly and die before having the G-tube taken out.
- Death is competing risk.

SAS: estimation of probability of G-tube out

- Time= G-tube insertion to G-tube out or last follow-up
- Censor variable:
 - 1= G-tube is out
 - 2= dead before G-tube out
 - 0=alive, G-tube still in

R: Probability of G-tube out

```
> cens=(hn$gstat==1)+2*(hn$gstat==0 & hn$stat==1)
> fit=cuminc(hn$time,cens)
> fit
```

Estimates and Variances:

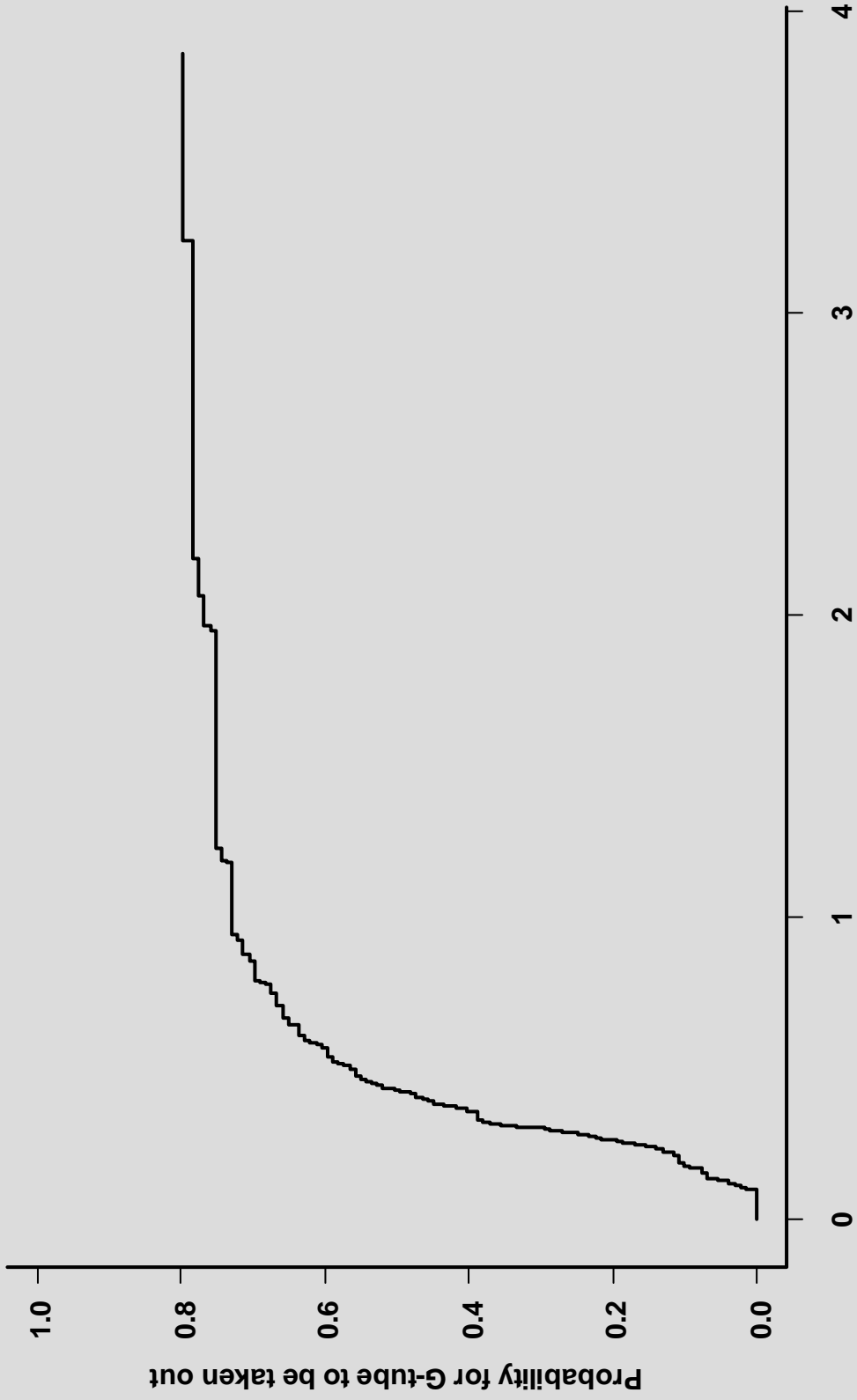
```
$est
      1      2      3
1 1 0.7286822 0.7674419 0.7829457
1 2 0.1240310 0.1705426 0.1705426
$var
      1      2      3
1 1 0.0015666011 0.001427967 0.001368097
1 2 0.0008610367 0.001133125 0.001133125
```

```
> timepoints(fit, times=c(1.5,3))
```

```
$est
      1.5      3
1 1 0.7519380 0.7829457
1 2 0.1627907 0.1705426
$var...
```

R: graphing the probability of event

```
forplot=list(list(fit$'1 1'$time,fit$'1 1'$est))  
  
plot.cuminc(forplot,lwd=2,  
xlab='Time to G-tube out (years)',  
ylab='Probability for G-tube to be taken  
out',wh=c(2,2))
```



Time to G-tube out (years)

Reporting

- Definition of endpoints
- Sample size
- Other endpoints
- Follow-up
- Number of events of interest
- Number of other events, any competing risks
- Type of analysis