INTRODUCTION TO SURVIVAL ANALYSIS

Wendy Yi-Ying Wu

wendy.wu@regionvasterbotten.se

Questions

- Have you ever had the experience of buying a brand new phone, and just a week after the warranty expires, it suddenly stop working?

- Everyone will die but who can survive longer?

"Death is certain, the time is not"

Outline

- What is the survival analysis and when can we use the survival analysis?
- Survival data structure
- Basic concepts
- Visualizing and comparing the survival curves
- Modeling the effect of covariates on survival
- When can a model not be used?

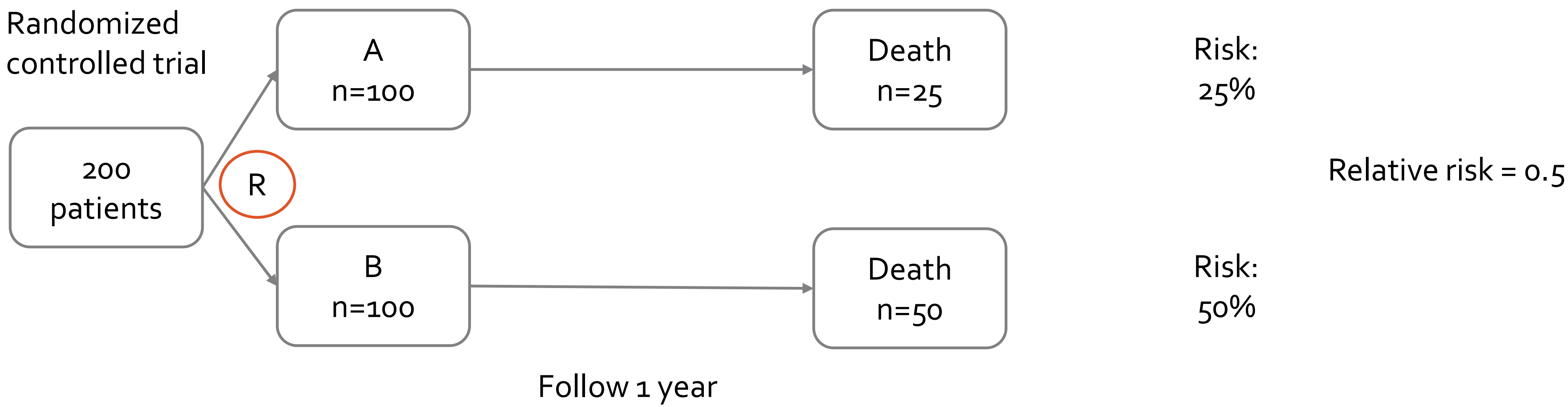
Survival analysis

- Is not restricted to death and survival

- ‘**time-to-event analysis**’ or ‘**event history analysis**’

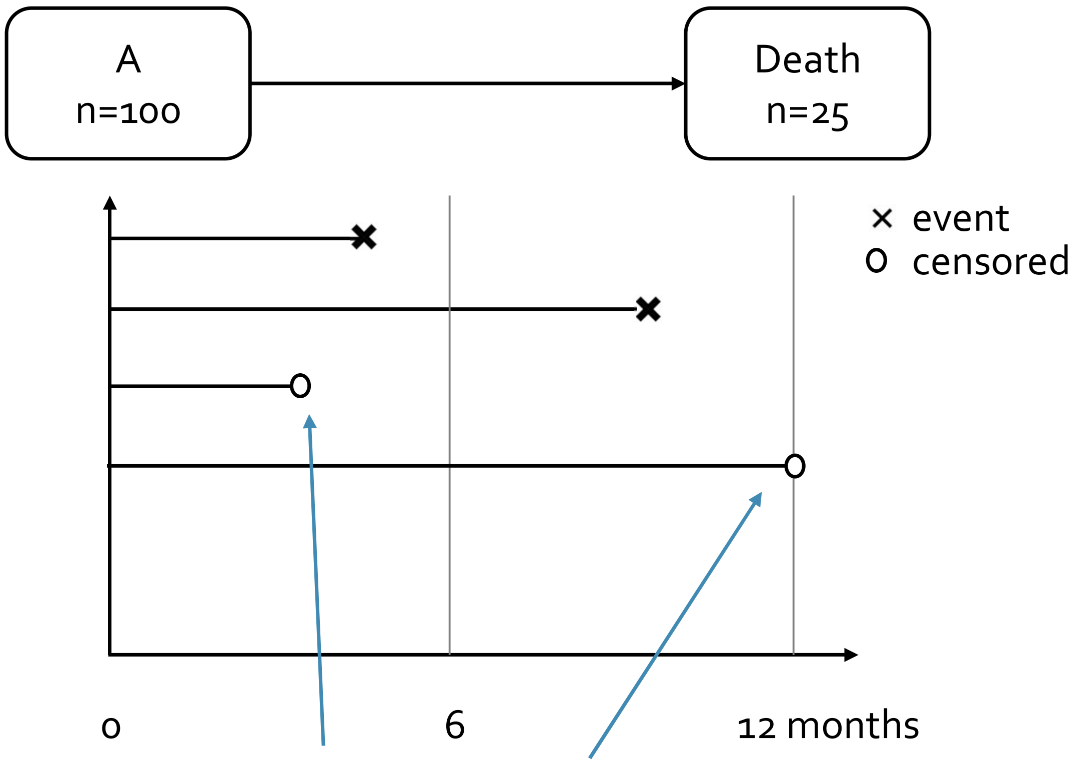
- Examples
  - Time from surgery to death
  - Time from HIV infection to development of AIDS
  - Time to machine malfunction

- Can patients survive longer if they recieve a new treatment A ?



Complication: people might drop out of studies!

Censoring

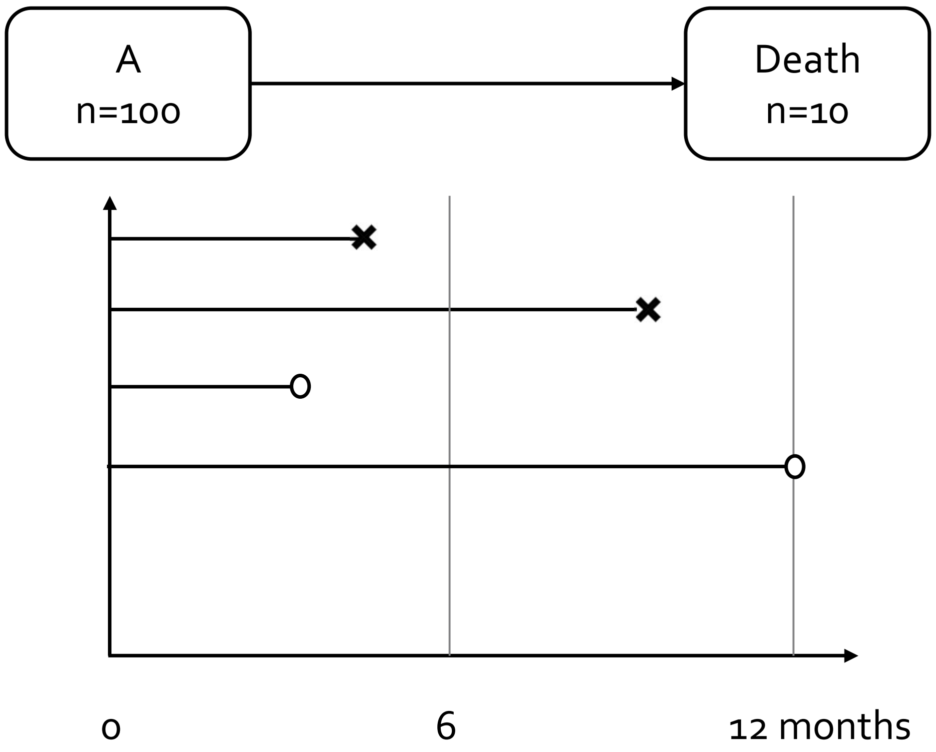


We don’t know when they will die, only that they are still alive at a certain time

Censoring may arise in the following ways

- has not (yet) experienced the event of interest within the study time period
- lost to follow-up
- a patient experiences a different event that makes further follow-up impossible.

Survival data structure



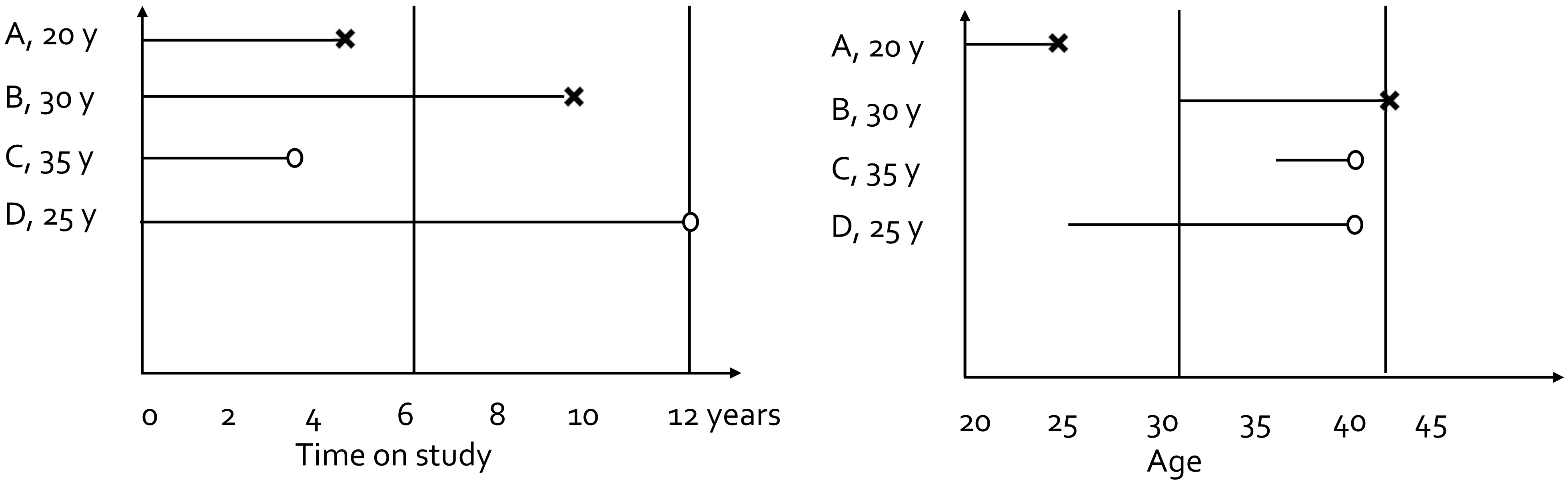
ID	Time	Event	Treatment
1	5	1	A
2	8	1	A
3	4	0	A
4	12	0	A

Survival time

- Starting time of the true survival time (Time=0)
  - Study entry
  - Beginning of treatment
  - Disease diagnosis
  - Surgery
  - Point in calendar time
  - Birth…

Survival time

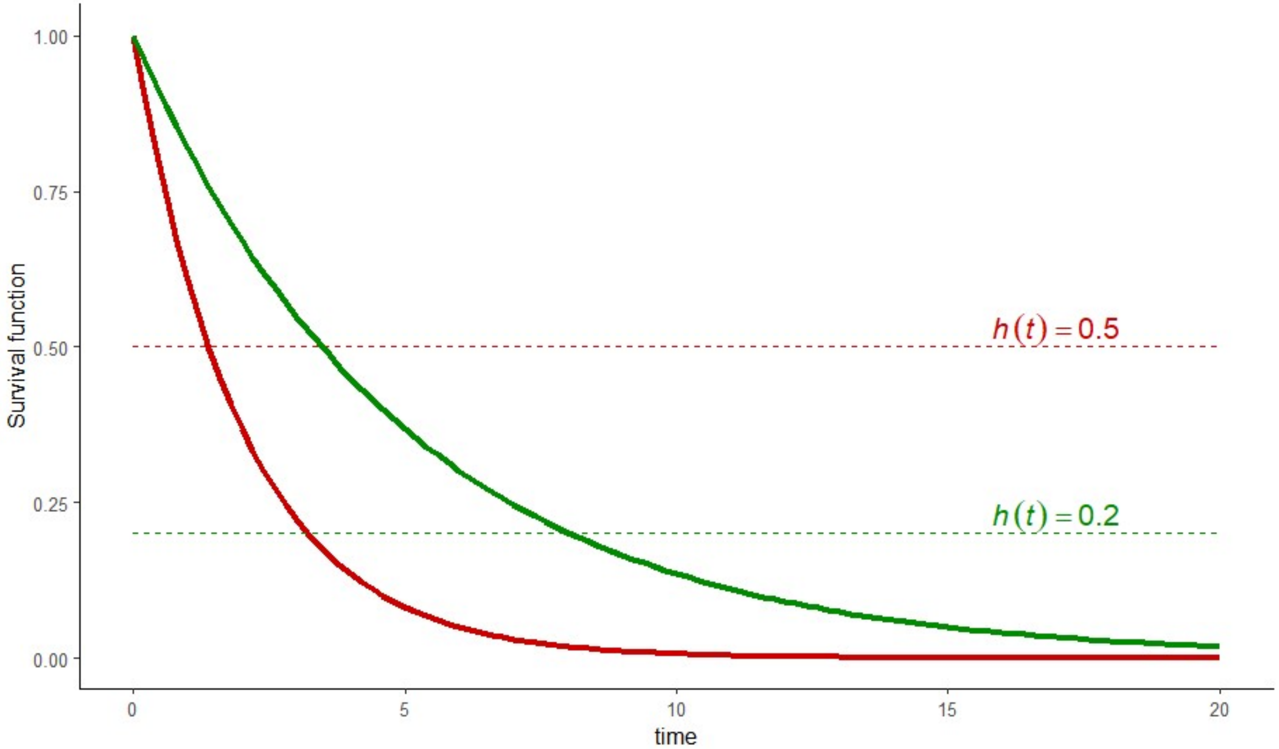
- Time until an event occurs (time scale)
  - Time on study
  - Age at follow-up



- Survivor function $S(t) = \Pr(T > t)$
  - The probability that an individual survives longer than t.

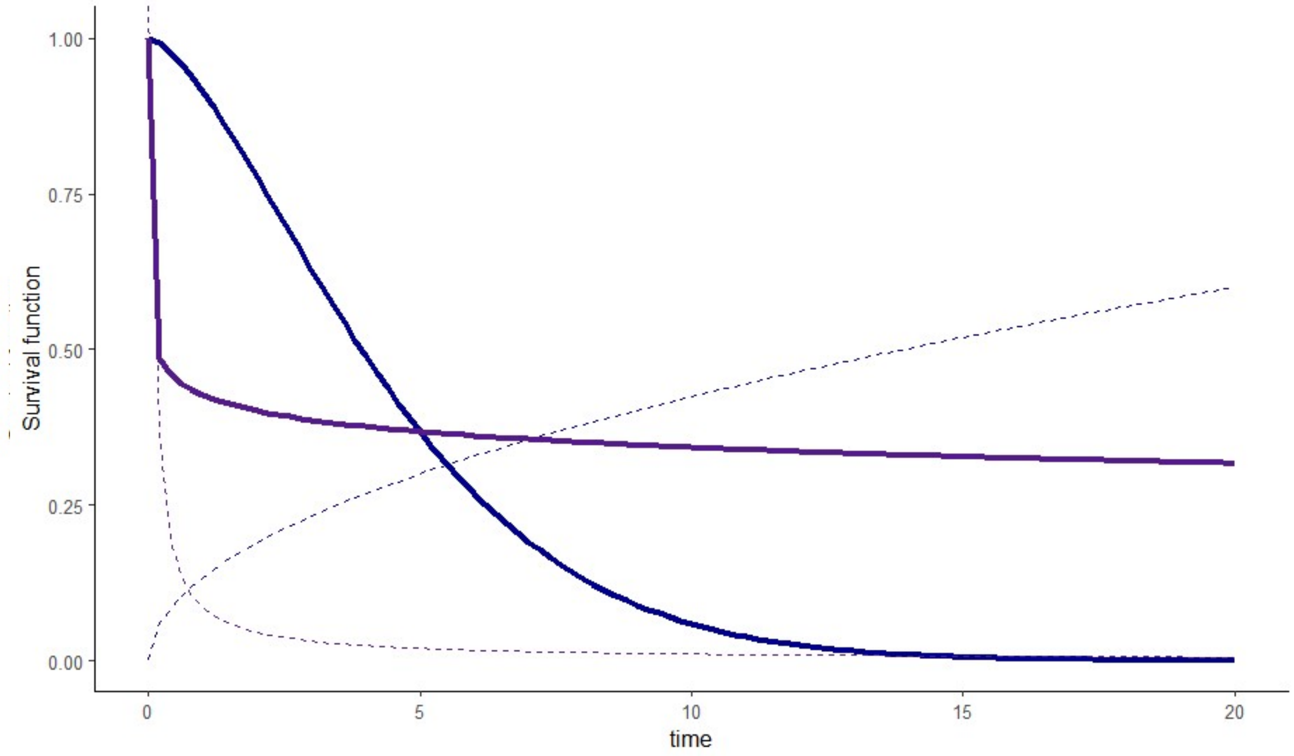
- Hazard function $h(t) = \lim_{\Delta t \to 0} \left(\frac{t \leq T < t + \Delta t | T \geq t}{\Delta t} \right) = \frac{f(t)}{S(t)}$
  - Instantaneous risk of the event
  - The probability that an individual who is under observation (still alive) at a time t has an event at that time t

Relations between hazard rate and survival curves



T ~ Exponential distribution

Relations between hazard rate and survival curves



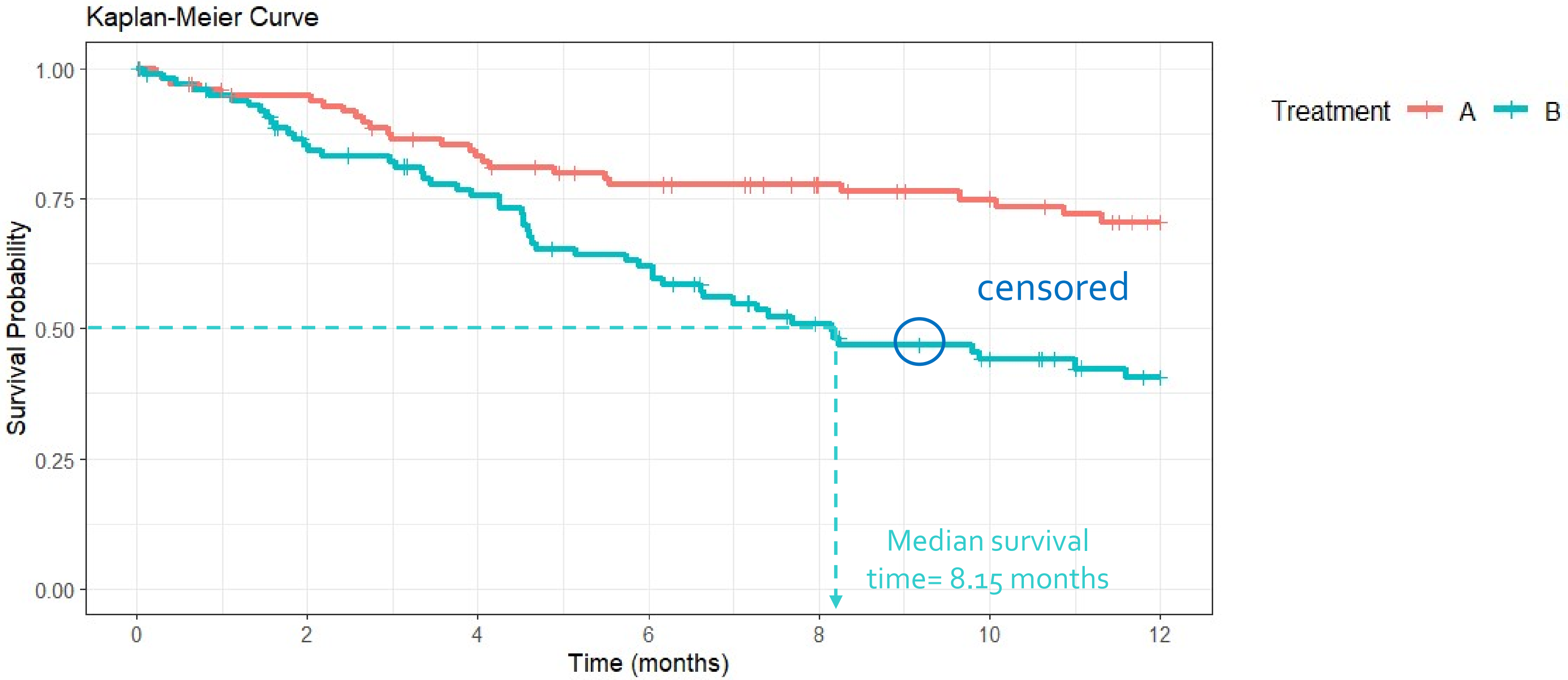
T ~ Weibull distribution

Goals of survival analysis

- Estimate and interpret survival and/or hazard functions
- Compare survival and/or hazard functions
- Modeling the effect of covariates on survival

1. Estimate survival function

Kaplan-Meier estimate of survival function



A

At Risk	100	90	77	68	58	53	43
Events	0	5	16	21	21	23	26

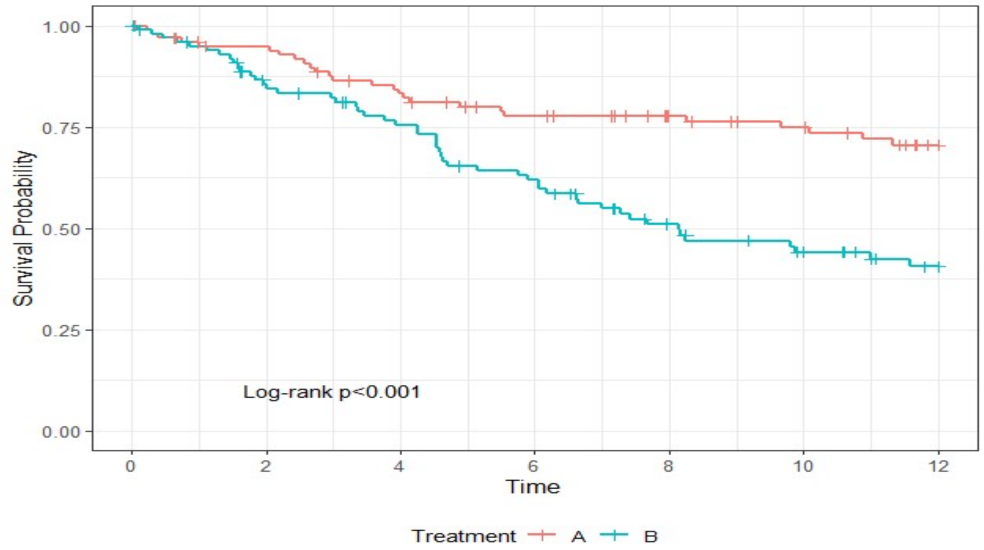
B

At Risk	100	79	67	54	38	29	21
Events	0	14	23	35	44	49	51

2. Compare survival function

Are KM curves statistically equivalent?

- Log-rank test
- Alternatives
  - Wilcoxon (Breslow)
  - Peto
  - Flemington-Harrignton




```
Call:
survdiff(formula = Surv(time = obst, event = event) ~ tx, data = dkm1)

        N Observed Expected (O-E)^2/E (O-E)^2/V
tx=A 100       26     42.3      6.29      14.1
tx=B 100       51     34.7      7.68      14.1

 Chisq= 14.1  on 1 degrees of freedom, p= 2e-04
```

Cox proportional hazards regression

The hazard rate at time t baseline Treatment

$$\underbrace{h(t)}_{\text{The hazard rate at time } t} = \underbrace{h_0(t)}_{\text{baseline}} \times \exp(\beta \underbrace{X}_{\text{Treatment}})$$

hazard rate at time t in the treatment A (X=1) group: $h_0(t) \times \exp(\beta)$
hazard rate at time t in the treatment B (X=0) group: $h_0(t)$

Hazard ratio: $\frac{h_0(t)\times\exp(\beta)}{h_0(t)} = \exp(\beta)$

How to interpret the results?

```
Call:
coxph(formula = Surv(time = obst, event = event) ~ tx, data = dkm1)

  n= 200, number of events= 77

       coef exp(coef) se(coef)      z Pr(>|z|)
txA -0.8803    0.4147   0.2420 -3.638 0.000275 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
txA    0.4147      2.412    0.2581    0.6663

Concordance= 0.598  (se = 0.029 )
Likelihood ratio test= 14.13  on 1 df,   p=2e-04
Wald test            = 13.24  on 1 df,   p=3e-04
Score (logrank) test = 14.09  on 1 df,   p=2e-04
```

Hazard ratio: 0.42 (95%CI= 0.26-0.67)

The hazard rate in the treatment A group is 0.42 times the hazard rate in the control group.

A patient in the treatment group A has 0.42 times probability (hazard) of death than a patient in the treatment group B.

The hazard of death occurring in the treatment A group is 58% lower compared to the baseline group (treatment B).

Cox proportional hazards regression

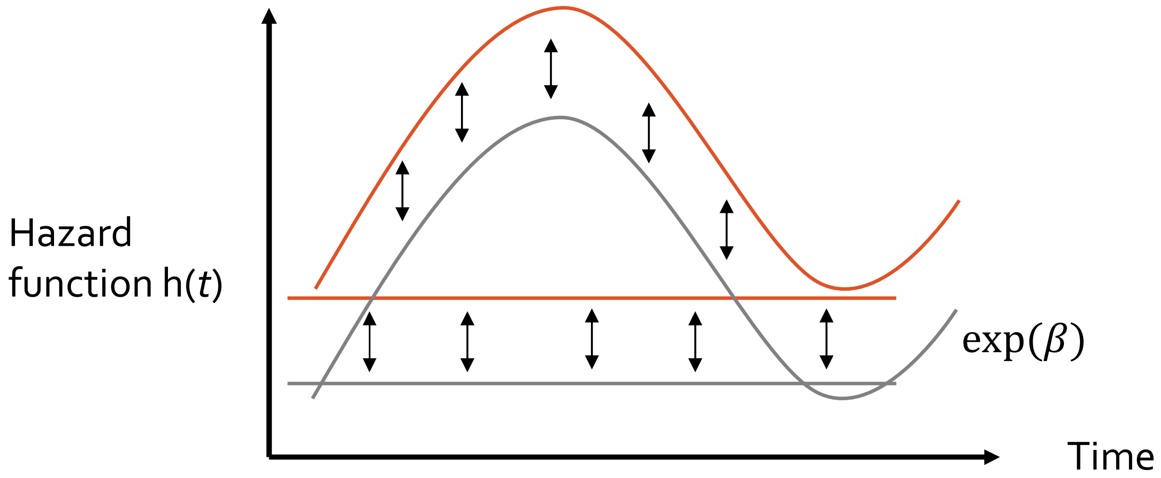
Independent variables
(exposures)

$$h(t) = h_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots)$$

baseline hazard

Regression coefficient

When can I use Cox PH regression?

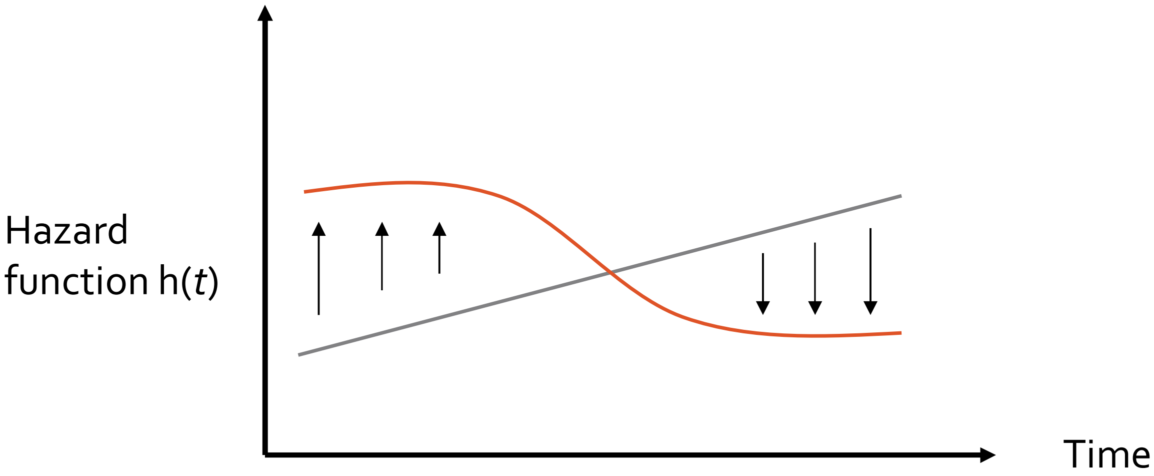


$$h(t) = h_0(t) \times \exp(\beta X)$$

Proportional hazards assumption

- The hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time
- Hazard ratio is constant over time

Violate PH assumption

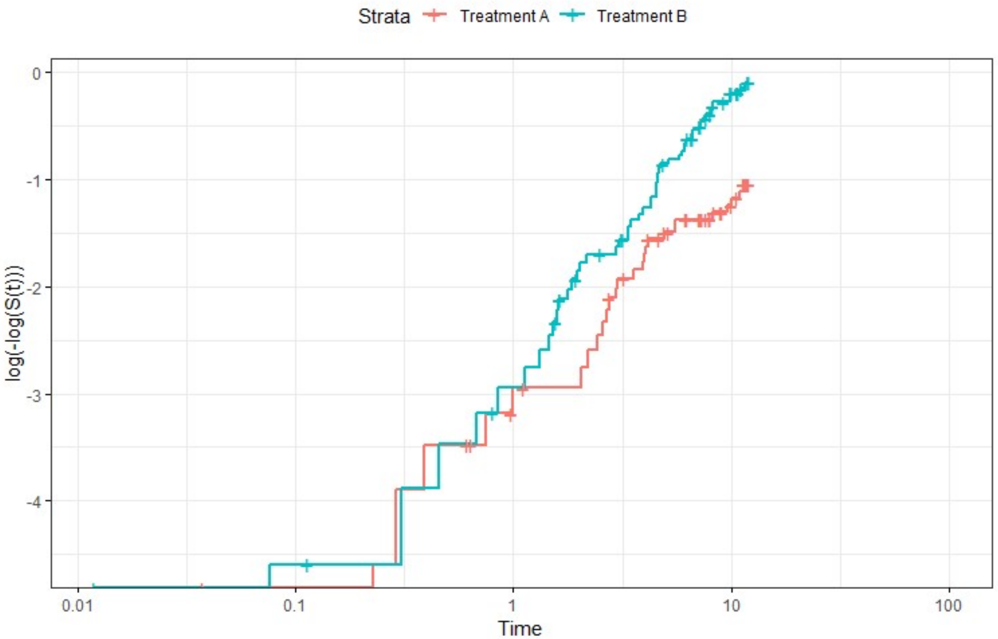


Hazard ratio is not constant over time

How do we know the PH assumption is satisfied?

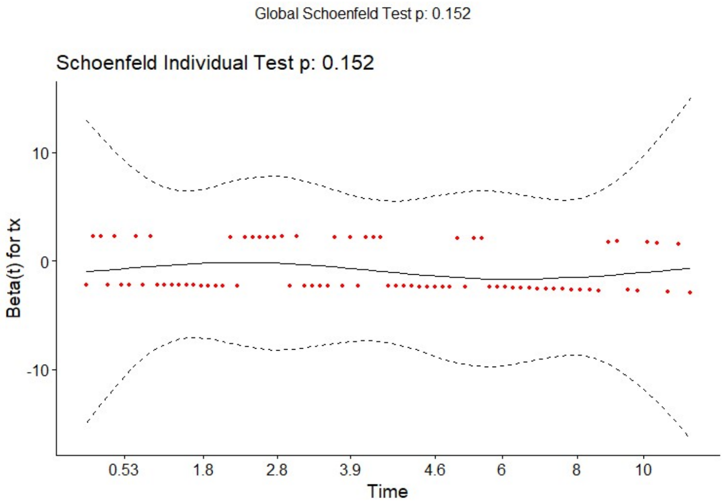
Evaluating PH assumption (3 methods)

- Graphical approach: log(-log S(t))



Parallel curves indicate the PH assumption is satisfied

- Schonefeld residuals test




```
> cox.zph(m1)
        chisq df    p
tx       2.05  1 0.15
GLOBAL   2.05  1 0.15
```

Evaluating PH assumption (3 methods)

- Using time-dependent covariates

$$h(t) = h_0(t) \times \exp(\beta_1 X + \boxed{\beta_2} X \times g(t))$$

Test $\beta_2 = 0$

if $\beta_2 \neq 0$ → PH assumption is violated

Some choices for $g(t)$:

$$g(t) = t$$

$$g(t) = \log(t)$$

$$g(t) = \begin{cases} 1\ if\ t \geq t_0 \\ 0\ if\ t < t_0 \end{cases}$$ (Heaviside function)

How can we do if PH assumption is violated?

- Landmark analysis
  - Start the analysis at time=X and analyse only those subjects who have survival until the X
    - Note: landmark time should be chosen carefully
- Fit several Cox models separately
  - Divide into shorter time periods for which the proportional hazard assumption is nearly correct

How can we do if PH assumption is violated?

- Fit a modified Cox model that includes a time-dependent variable which measures the interaction of exposure with time

$$h(t) = h_0(t) \times \exp(\beta_1 X + \beta_2 X \times g(t))$$

$$g(t) = \begin{cases} 1 \; if \; t \geq 4 \\ 0 \; if \; t < 4 \end{cases}$$ (Heaviside function)

```
Call:
coxph(formula = Surv(time = obst, event = event) ~ tx1 + tt(tx1),
    data = dkm1, tt = function(x, t, ...) x * (t >= 4))

           coef exp(coef) se(coef)      z      p
tx1     -0.4362    0.6465   0.3257 -1.339 0.1805
tt(tx1) -0.9341    0.3929   0.4926 -1.896 0.0579

Likelihood ratio test=17.81  on 2 df, p=0.0001359
n= 200, number of events= 77
```

HR before 4 months: exp(-0.43)=0.65
HR after 4 months: exp(-0.43-0.93)=0.25

How can we do if PH assumption is violated?

- Stratified Cox model
  - Stratification of a covariate that does not satisfy the PH assumption
  - Ex. Sex doesn't satisfy PH

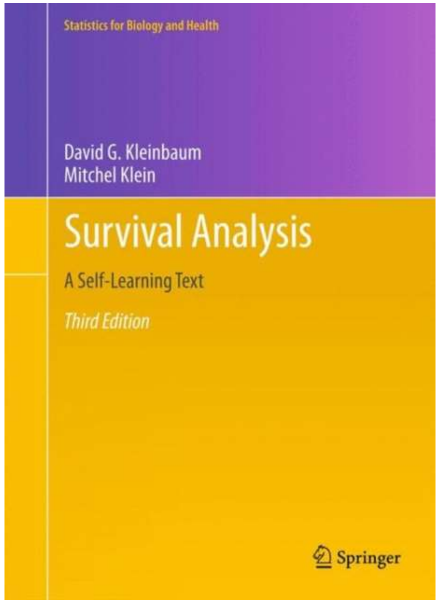
$$h(t) = h_{0i}(t) \times \exp(\beta_1 X)$$

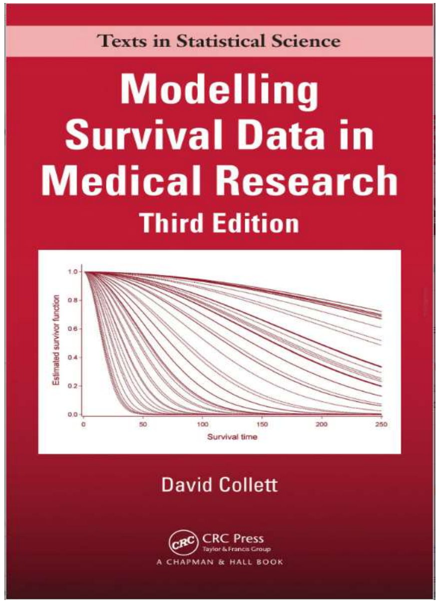
$h_{01}(t)$ for males and $h_{02}(t)$ for females but they share the same β_1 for treatment

Note: can't obtain a hazard ratio for the effect of sex adjusted for treatment

- Flexible parametric survival model
  - 5/12 kl 12-13 Icke-proportionella hazarder i Coxmodeller, Christel Häggström

Suggested references





- Kleinbaum, & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (3 ed.) Springer Nature. https://doi.org/10.1007/978-1-4419-6646-9
- Collett. (2015). *Modelling survival data in medical research* (3 ed.). Chapman & Hall/CRC.

Kliniska Studier Sverige Forum Norr

Canvas page:
https://www.canvas.umu.se/courses/2600

Hösten 2023

- 19/9 kl 12-13 SPSS-handhavande
- 4/10 kl 12-13 Bortom linjäritet: En introduktion till splines
- 26/10 kl 12-13 Standardiserade epidemiologiska mått
- 14/11 kl 12-13 Enkätkonstruktion
- 5/12 kl 12-13 Icke-proportionella hazarder i Coxmodeller