CBB Workshop: Introduction to Survival Analysis

with Practical Applications in R

Rickard Strandberg & Davide Valentini
Center for Bioinformatics and Biostatistics

2024-03-19

CBB offers

- **Mentoring of PhD students and postdocs**
- Seminar series and workshops
- Support in project planning and grant applications
- Joint employments including co-financing
- Networking opportunities to discuss bioinformatical and biostatistical methods and problems
  - → **Drop-in on Thursdays 13:00-15:00 in Neo, room Protein**
- Please see www.ki.se/cbb for more information, or contact us at cbb@ki.se

Agenda

- 1st Session 13:00–14:00
  → The basics of survival analysis
- Coffee Break 14:00–14:30
- 2nd Session 14:30–16:00
  → Practical application in R
  → examples, graphics, and exercises

Introduction to Survival Analysis

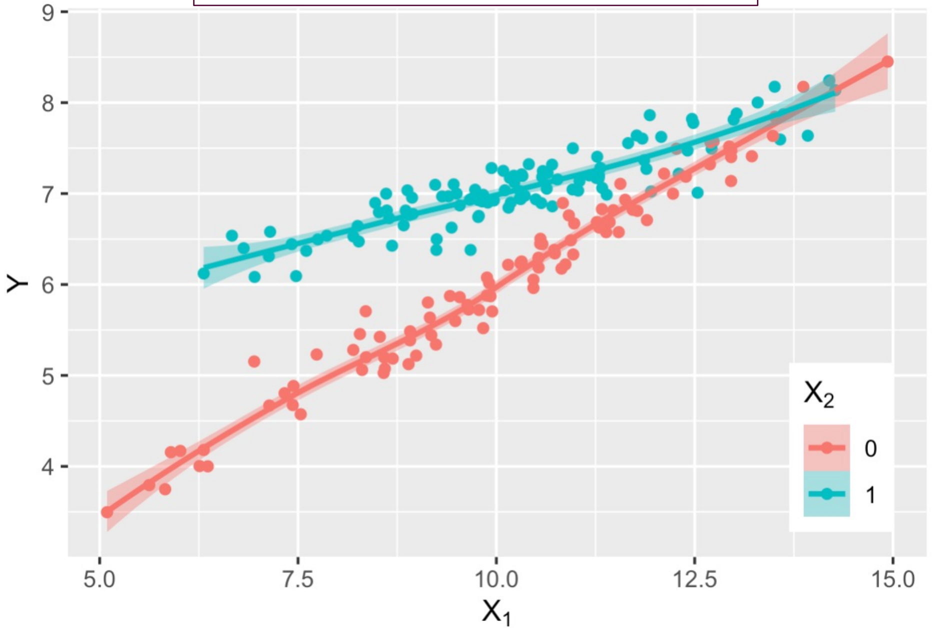
Outline

- The most important concepts of survival analysis
  1. The type of data we use
  2. The survival function & the cumulative incidence function
  3. The hazard rate function
  4. The Cox proportional hazards regression model

Our Typical Statistical Analysis

- Typical data & research question relationship:
  - → We measure some **Outcome variable (Y)**
  - → We measure a set of **Covariates (X_1, X_2, …)**
    - Exposure/risk factor variables potentially associated with the outcome
  - → We perform some statistical analysis—**Regression Model**

$$\hat{Y} = 1 + 0.5X_1 + 4X_2 - 0.3X_1X_2$$

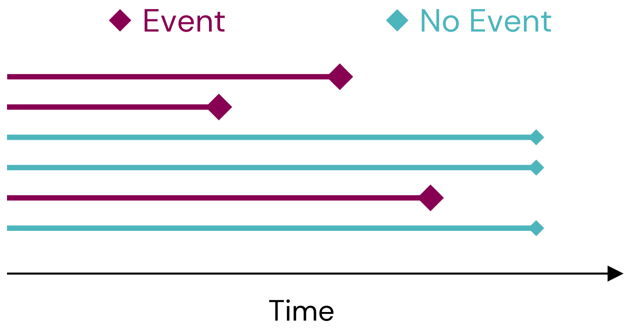


What is Survival Analysis?

- Survival Analysis is a special family of statistical analysis
- **The outcome is time**
- We measure the time until a particular event
  - → Death (Mortality)
  - → Diagnosis of a particular disease (Incidence)
  - → Progression to a stage of disease
  - → etc.

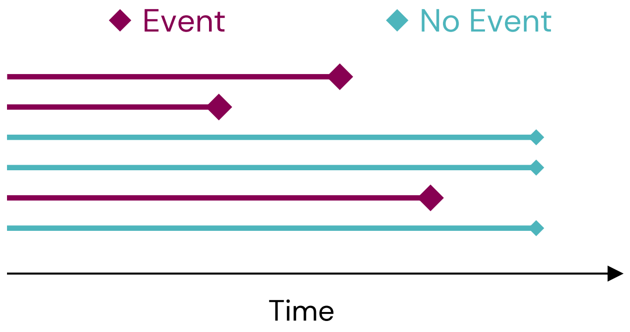
Time-to-Event Data

- Time-to-event outcome:
  1. The time duration we followed each subject
  2. Whether or not they experienced the event



Time-to-Event Data

- Time-to-event outcome:
  1. The time duration we followed each subject
  2. Whether or not they experienced the event

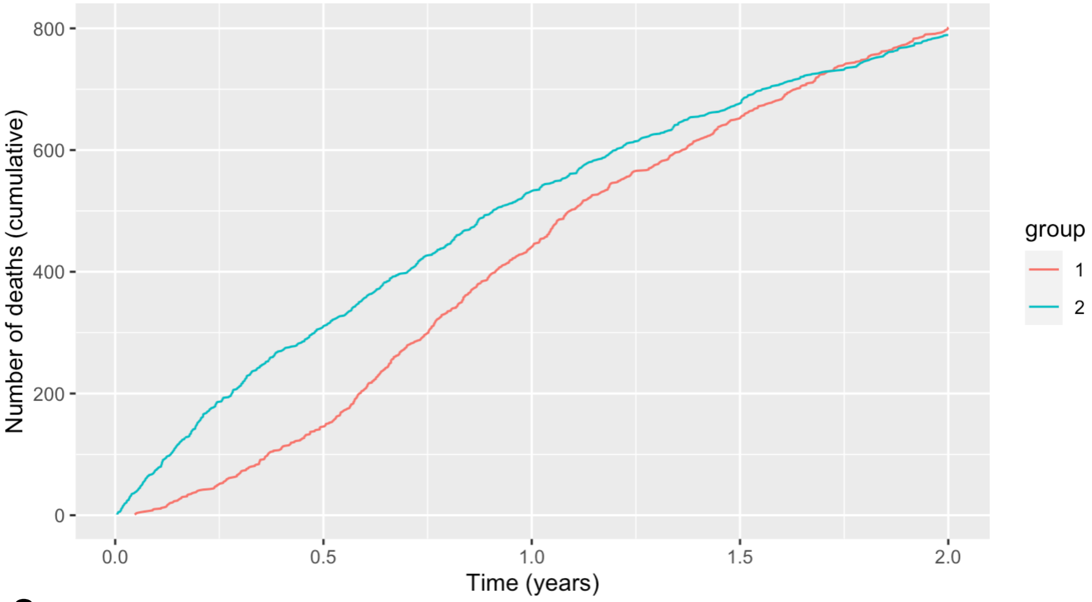


For example:

id	time	event	age	sex	bmi	masld
23	6.08	0	32	female	35.2	no
24	6.79	0	43	female	24.5	no
25	3.65	0	39	male	24.3	no
27	0.59	1	62	female	30.2	no
32	11.67	0	45	male	22.1	no
33	2.03	1	58	male	26.7	no
36	1.13	0	53	female	24.8	no
37	7.59	0	47	male	21.9	no
38	9.11	0	54	female	22.1	no
40	4.67	1	61	female	28.4	no
41	6.10	1	45	female	26.3	no

Why two-part outcome?

- Can't we just look at the final outcome?



- <u>Example:</u>
  Mortality over 2 years comparing 2 groups

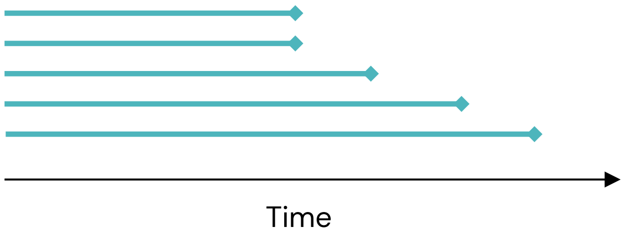
- Not just **what** happens, but also **when**

- Who experiences more events and/or events sooner?

Censoring

- We don’t observe events for everyone
  → We end the follow-up at some point
  → (Given enough time, would everyone have the event?)

- **“Censored at time *t*”** = event occurs sometime after *t*
  → *i.e. we didn’t observe it yet*
  → *a.k.a. “lost to follow-up”*

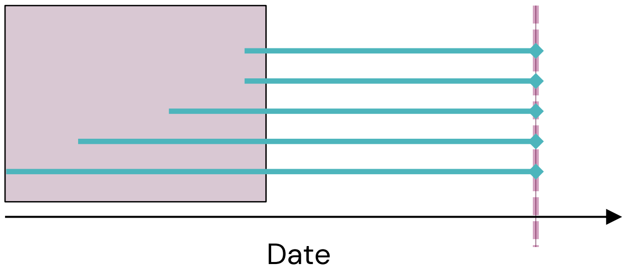


Censoring

- We don't observe events for everyone
  - → We end the follow-up at some point
  - → (Given enough time, would everyone have the event?)

- **"Censored at time *t*"** = event occurs sometime after *t*
  - → *i.e. we didn't observe it <u>yet</u>*
  - → *a.k.a. "lost to follow-up"*

- Often due to inclusion/recruitment over time and with a fixed end date



Now What?

- **What do we want to know?**

- In a typical non-survival analysis:

1. Summary statistics (mean/median, proportions) including
   → standard errors (confidence intervals)
   → preliminary comparisons between groups (graphically or with statistical tests)
2. Proper regression model
   → Multivariate model
   → Effect sizes (mean differences, odds ratios, etc.) and confidence intervals
   → Statistical tests for association with the outcome

The Survival Function

The Survival Function

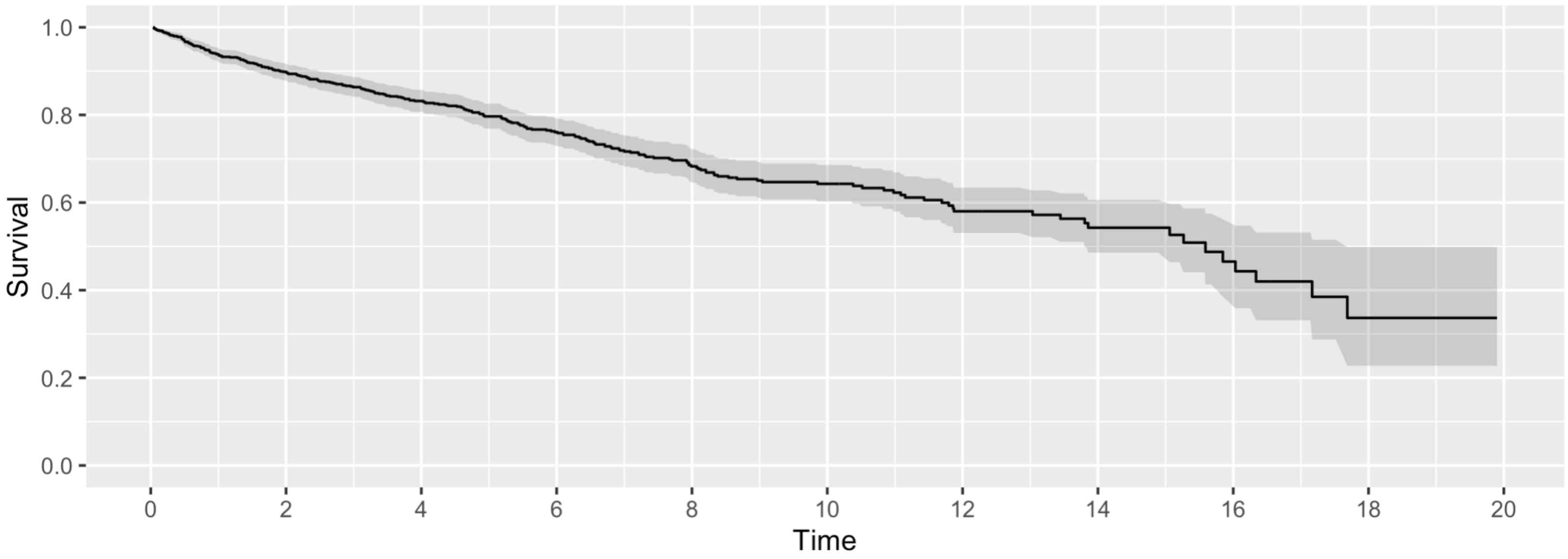
- If we measure the time to event, a reasonable question might be:
- *"What is the probability of not having the event by time t?"*
  or
  "What proportion of the population will not have the event by time t?"

- The survival function S(t): The probability of being event-free by time *t*
- It's a probability which decreases over time
- $S(0) = 1 \qquad if \quad t_1 < t_2 \quad then \quad S(t_1) \geq S(t_2)$

- *Use the time-to-event data to estimate:*

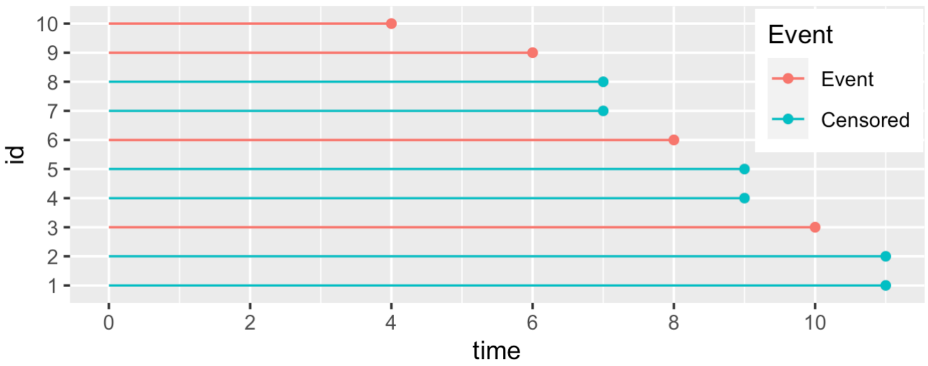
Kaplan-Meier Estimate

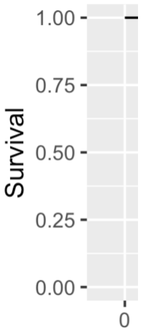
- Kaplan-Meier estimate of the survival function (with 95% confidence int.)



rickard.strandberg@ki.se

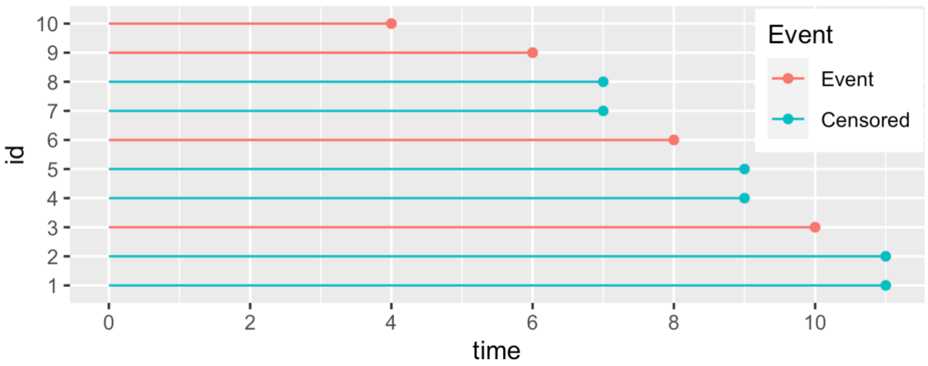
Calculating the Survival Function

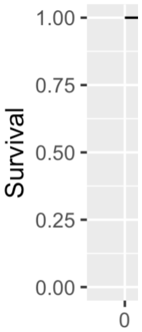




Survival calculation:

Calculating the Survival Function

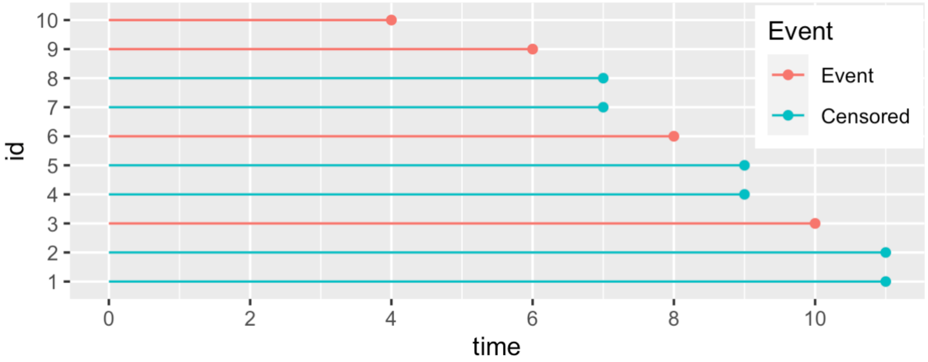


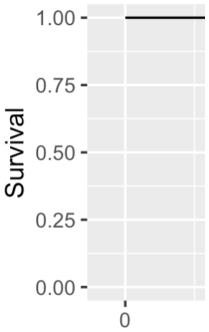


Survival calculation:

- $S(0) = \frac{10}{10} = 1$

Calculating the Survival Function

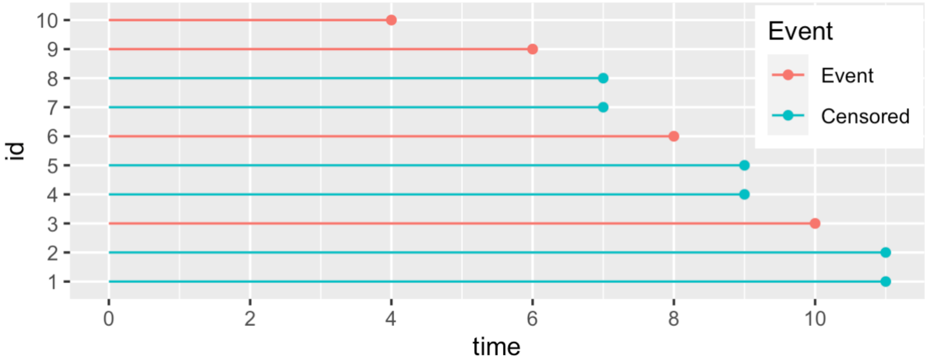


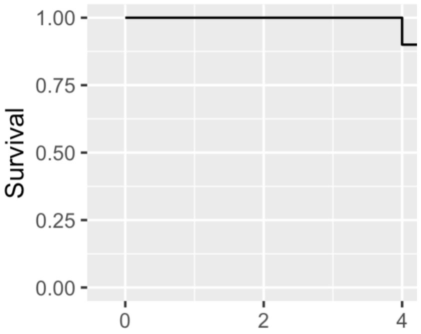


Survival calculation:

- $S(0) = \frac{10}{10} = 1$
- $S(1) = \frac{10}{10} = 1$

Calculating the Survival Function

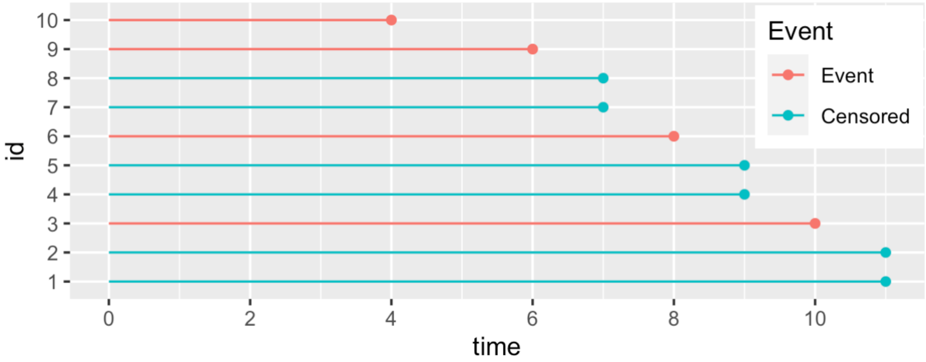


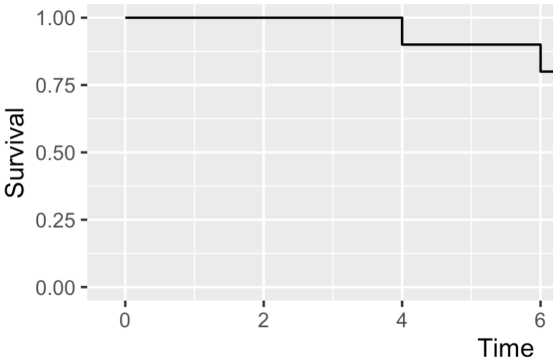


Survival calculation:

- $S(0) = \frac{10}{10} = 1$
- $S(1) = \frac{10}{10} = 1$
- $S(4) = \frac{9}{10} = 0.9$

Calculating the Survival Function

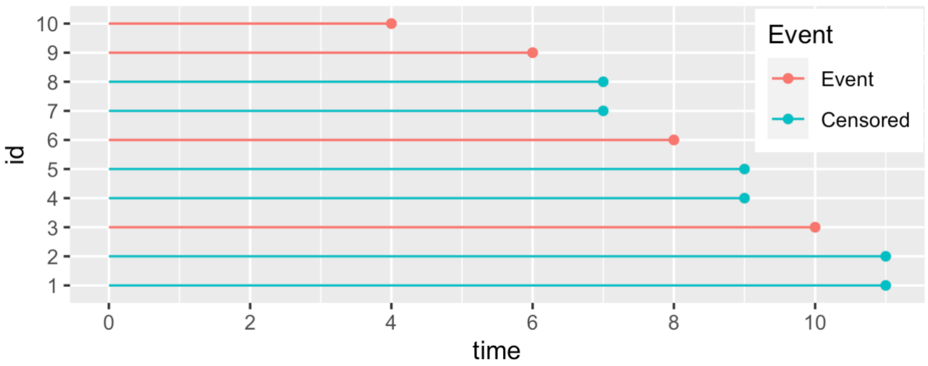


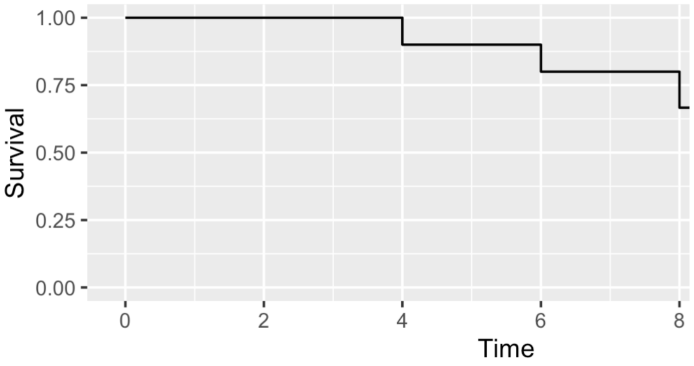


Survival calculation:

- $S(0) = \frac{10}{10} = 1$
- $S(1) = \frac{10}{10} = 1$
- $S(4) = \frac{9}{10} = 0.9$
- $S(6) = S(4) \cdot \frac{8}{9} = 0.9 \cdot 0.89 = 0.8$

Calculating the Survival Function

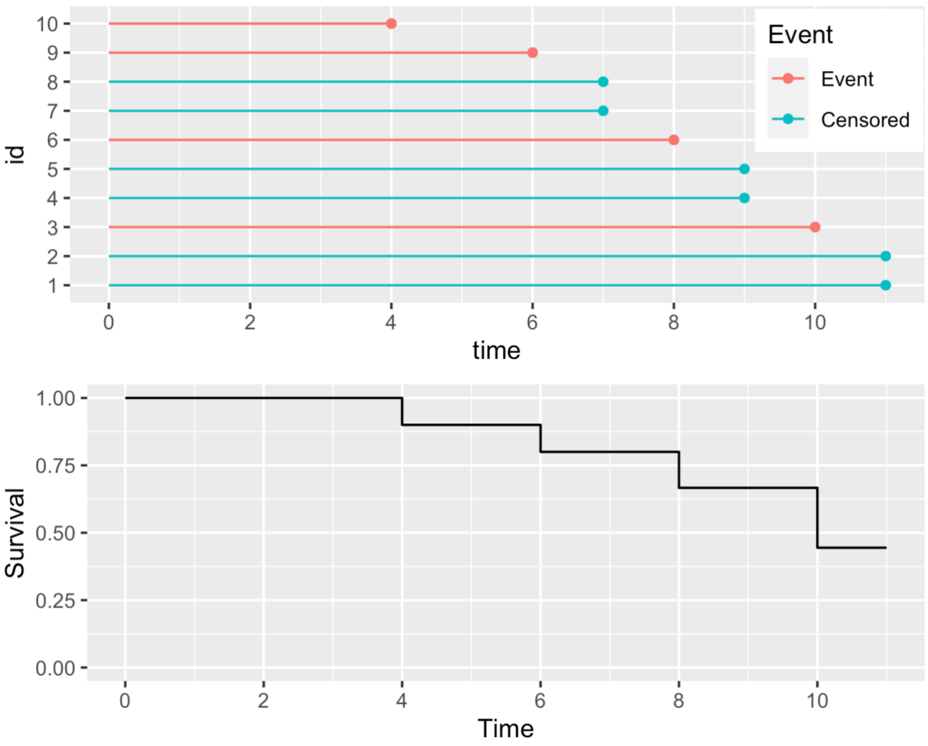




Survival calculation:

- $S(0) = \frac{10}{10} = 1$
- $S(1) = \frac{10}{10} = 1$
- $S(4) = \frac{9}{10} = 0.9$
- $S(6) = S(4) \cdot \frac{8}{9} = 0.9 \cdot 0.89 = 0.8$
- $S(8) = S(6) \cdot \frac{5}{6} = 0.8 \cdot 0.83 = 0.67$

Calculating the Survival Function

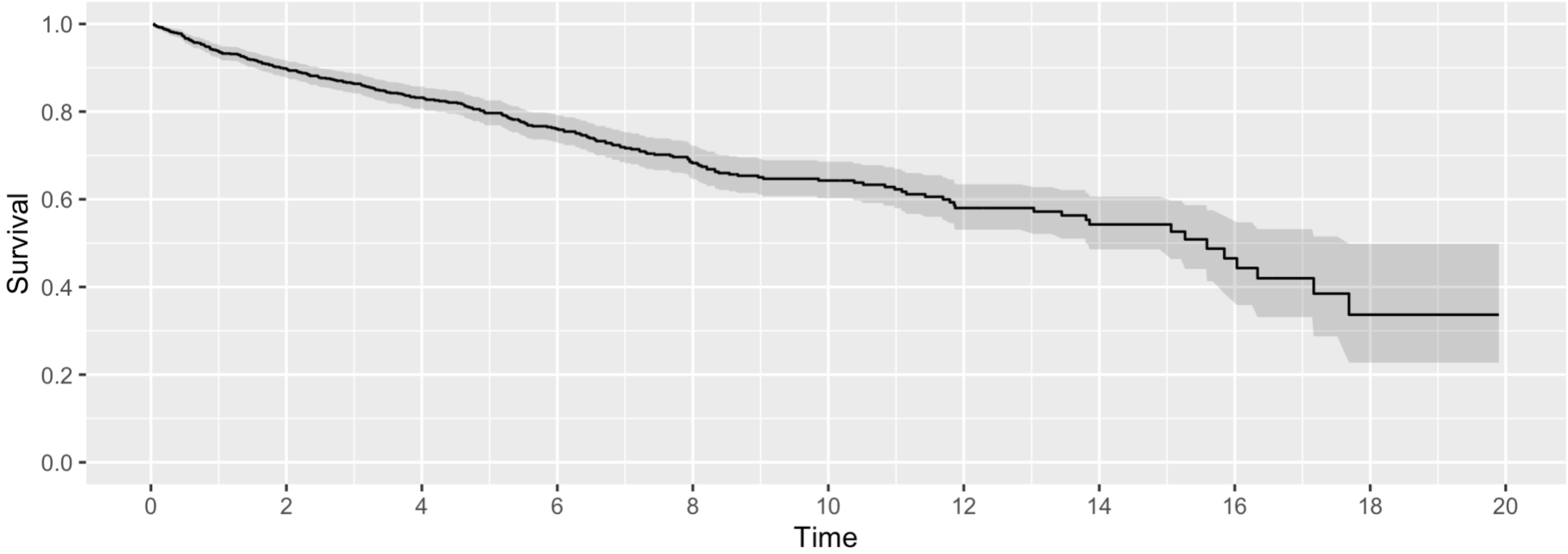


Survival calculation:

- $S(0) = \frac{10}{10} = 1$
- $S(1) = \frac{10}{10} = 1$
- $S(4) = \frac{9}{10} = 0.9$
- $S(6) = S(4) \cdot \frac{8}{9} = 0.9 \cdot 0.89 = 0.8$
- $S(8) = S(6) \cdot \frac{5}{6} = 0.8 \cdot 0.83 = 0.67$
- $S(10) = S(8) \cdot \frac{2}{3} = 0.67 \cdot 0.67 = 0.44$

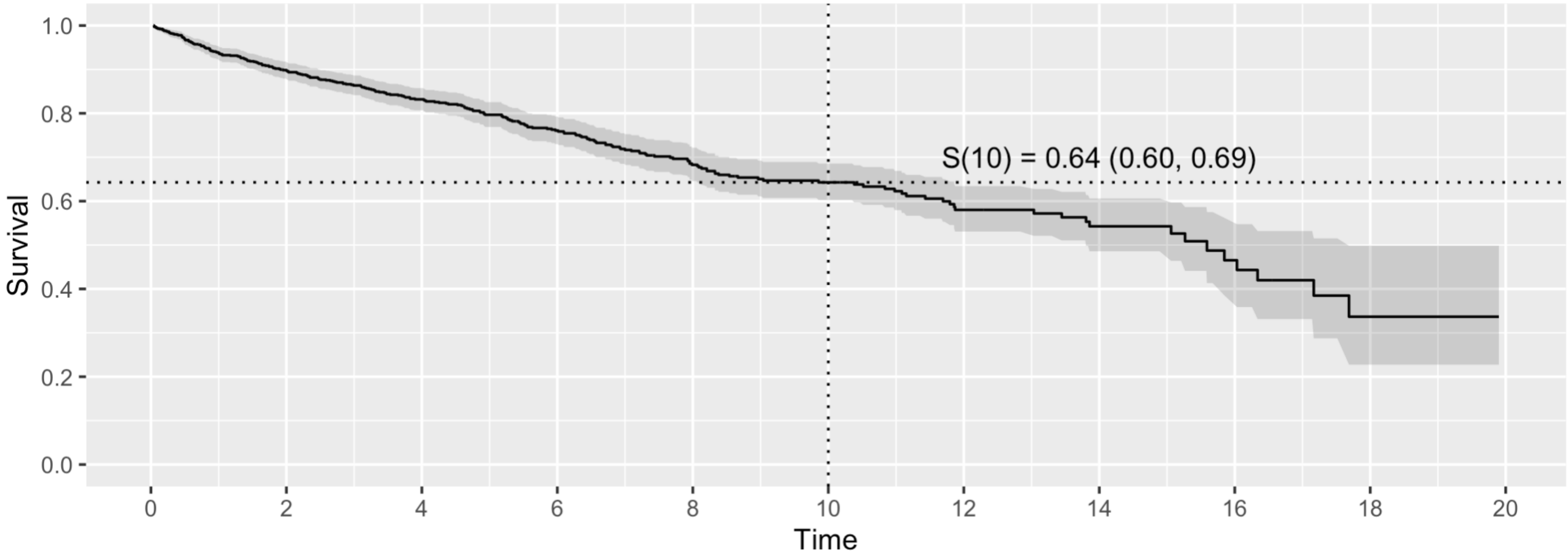
Kaplan-Meier Estimate

- Kaplan-Meier estimate of the survival function (with 95% confidence int.)



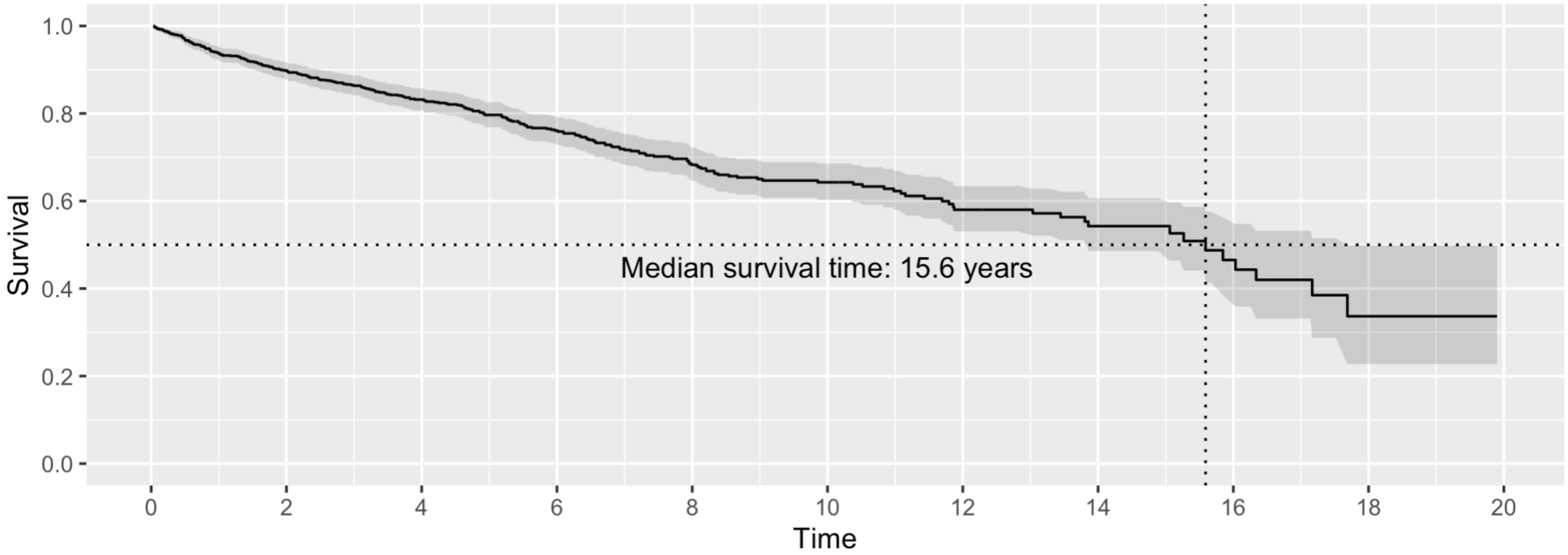
Kaplan-Meier Estimate (2)

- Kaplan-Meier estimate of the survival function (with 95% confidence int.)



Median survival time

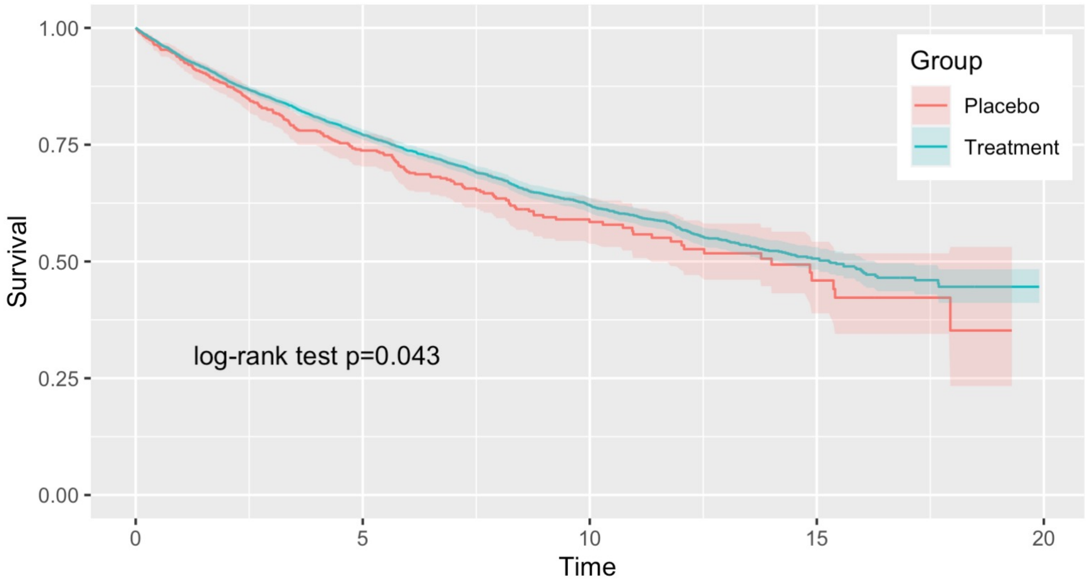
- Another interesting statistic could be the **Median Survival Time**:
  → The time until 50% have experienced the event.
  → For which *t* is *S(t)=0.5?*



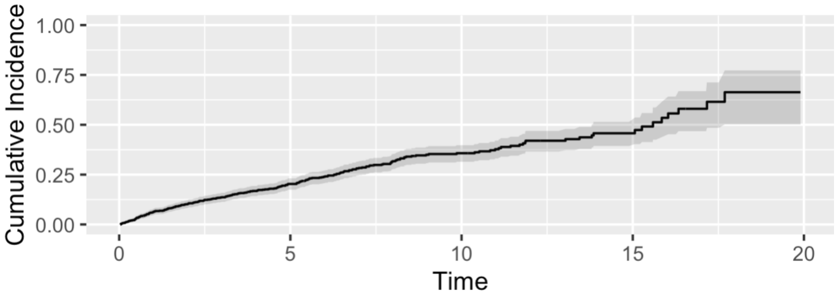
Comparing Survival Curves

- We can compare survival curves between groups

- We can do a statistical test—**the log-rank test**–for difference between the curves



The Cumulative Incidence Function



- *"The risk"*
- $1 - S(t)$

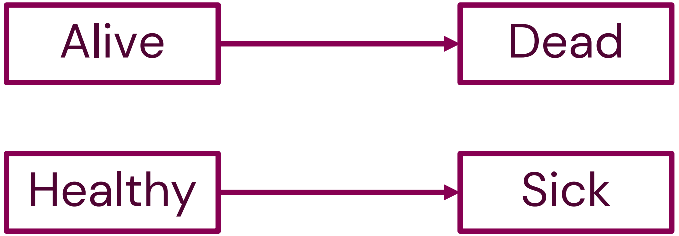
- The proportion who **have had the event** by *t*
- or equivalently the probability of having the event by time *t*

- Study Survival or Cumulative Incidence depending on context
  - → Which more interesting: the event-havers or the event-free?
  - → Cancer researchers favor survival (often survival after cancer dx)
  - → Hepatology favors cumulative incidence (who is at greater risk of dx)

The Hazard Rate

Different View: Transition Between States

- Subject starts in one state and at some timepoint moves to a different state



- We study the transition (the arrow)

- $S(t)$ is the proportion still in the 1st state at time t

The Hazard Rate

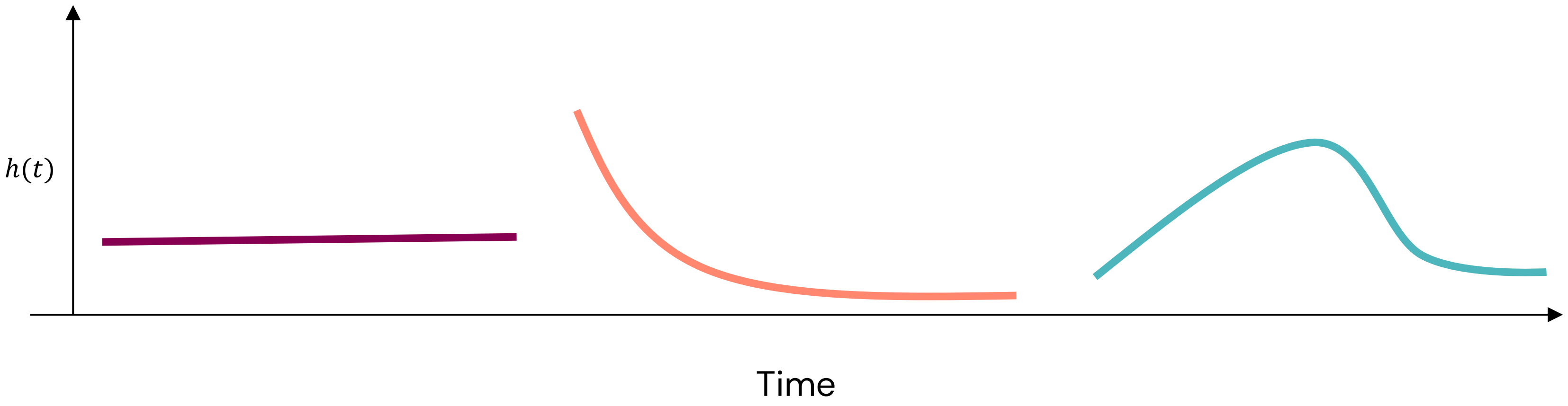
- The rate of transition is determined by the hazard rate
- Hazard rate function = $h(t)$

$$\boxed{\text{Healthy}} \xrightarrow{\boxed{h(t)}} \boxed{\text{Sick}}$$

- The hazard is a velocity
- $h(t) \geq 0$
- Higher hazard rate means both more events over time and events sooner

Hazard Function Examples

- The Hazard function can look very different depending on the mechanics of the transition



The Cumulative Hazard

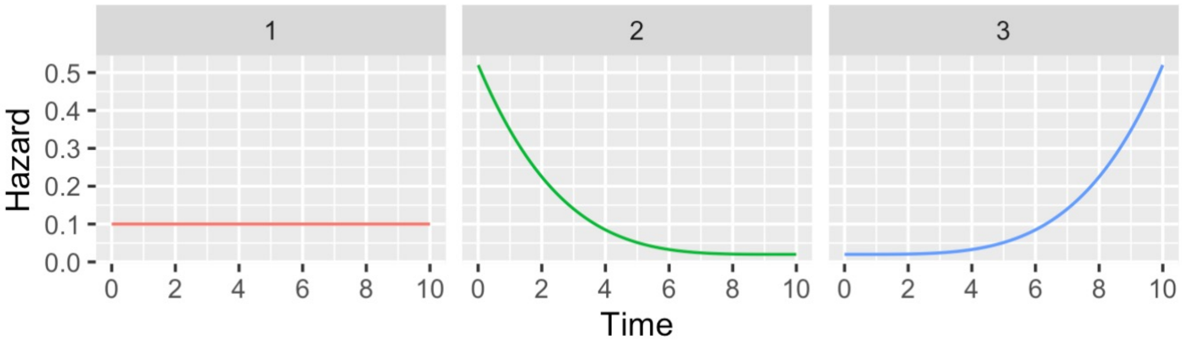
- The total accumulated hazard from *0* to *t*:

$$H(t) = \int_0^t h(u)du$$

- If the hazard is the velocity, the cumulative hazard is the total distance travelled
- How much hazard were you exposed to?
  - → A lot over a short time?
  - → A little over a long time?

Hazard & Survival

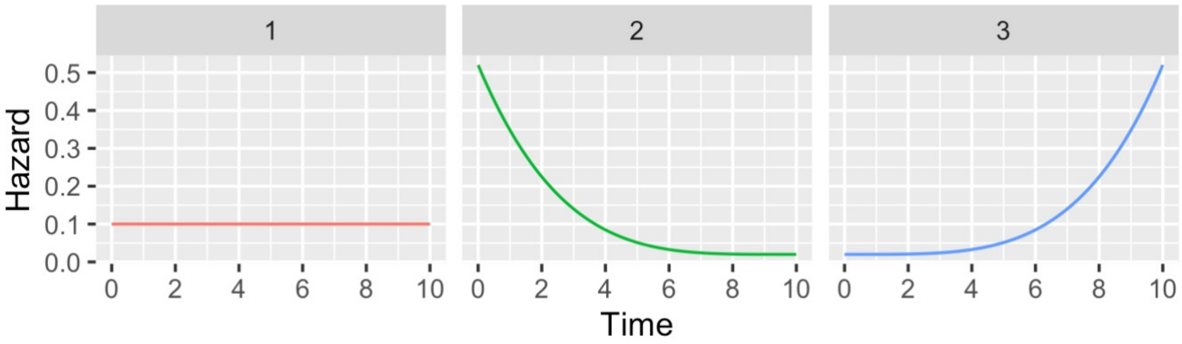
- How is the hazard function related to the survival function?

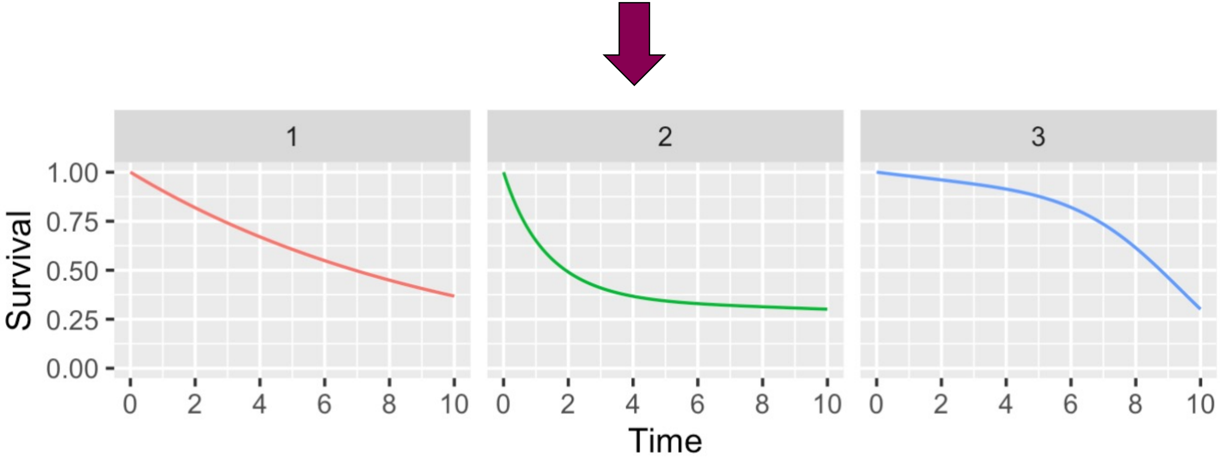


Hazard & Survival

- How is the hazard function related to the survival function?

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





Two Measures of Interest

- When doing survival analysis, we want to study two quantities:

1. The Survival (or Cumulative Incidence) function over time for summary and interesting comparisons
   → More practical scale
      - e.g. survival after 5 years
      - treatment vs. no treatment
      - by smoking and age group (4 comparisons)
2. The Hazard function
   → Useful scale for exposure/risk factors
   → Example: carcinogens for developing cancer

Cox Proportional Hazards Regression

Proportional Hazards

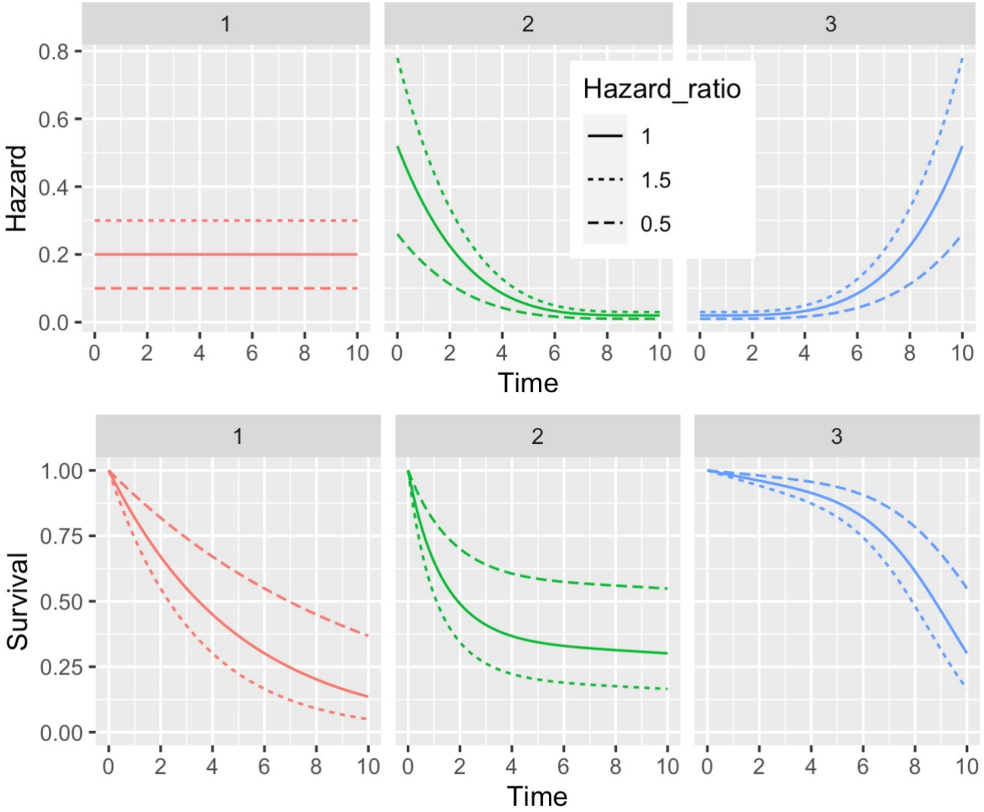
- How do we compare hazards $h_1(t)$ and $h_2(t)$?

- Hazard ratio: $\mathrm{HR(t)} = \frac{h_1(t)}{h_2(t)}$

- **Proportional hazards:**
  → If the hazard ratio is constant over time
  → $h_1(t) = HR \cdot h_2(t)$

Proportional Hazards

- How do we compare hazards $h_1(t)$ and $h_2(t)$?
- Hazard ratio: $\mathrm{HR(t)} = \frac{h_1(t)}{h_2(t)}$
- **Proportional hazards:**
  → If the hazard ratio is constant over time
  → $h_1(t) = HR \cdot h_2(t)$



Cox Proportional Hazards Regression Models

- The workhorse for inference in survival analysis

- All subjects have an unknown hazard function specific to them

- We impose a rule:
  - → There is a common underlying shape to all the hazards, $\boldsymbol{h_0(t)}$
  - → All subjects' hazards are then **proportional to $\boldsymbol{h_0(t)}$** according to their covariates/risk factors
- Subject *i*'s hazard = $h_0(t) \cdot HR_i$

Cox Proportional Hazards Regression Models (2)

- Subject *i*'s hazard = $h_0(t) \cdot HR_i$
- Cox regression is regression on $\ln(HR_i)$:

$$\ln(HR_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_m X_{mi}$$

- The β:s are estimated using the time-to-event outcome and the covariates
- $h_0(t)$ is not estimated! ("semi-parametric", "partial log-likelihood")

Interpreting Cox Models

$$\ln(HR_i) = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_m X_{mi}$$

- The β:s are "*log-hazard ratios*"
- **e^{β_j} is the hazard ratio from a 1 unit increase in covariate X_j**
  - → Binary (0/1): Change from treatment group 1 to treatment group 2
  - → Continuous: Change in BMI from 25 to 26
  - → Multiplicative, e.g. 2 unit change ⇒ $e^{\beta_j \cdot 2} = e^{\beta_j} \cdot e^{\beta_j}$, etc.

- Like all other regressions we can get confidence intervals and p-values for each β

Pros and Cons of the Cox Model

- Pros:
  - → We don't need to model $h_0(t)$
  - → Very efficient compared to fully parametric alternatives
- Cons:
  - → Hinges on the proportional hazards assumption
  - → No time-varying effects

Example: Primary Biliary Cholangitis & Death

- PBC: Autoimmune disease destroying the bile ducts in the liver
- Causes fibrosis (scarring) $\Longrightarrow$ Cirrhosis $\Longrightarrow$ Liver failure $\Longrightarrow$ Death

- *What is your prognosis based on your current fibrosis stage (1–4)?*

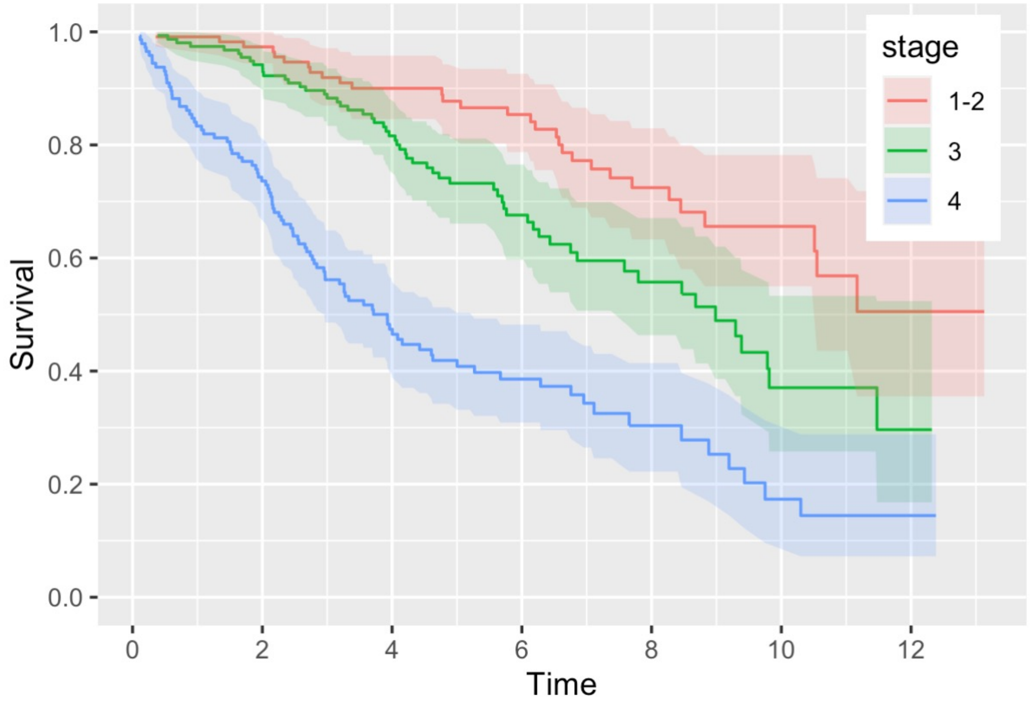
- 412 PBC patients from the Mayo clinic
- <u>Stage 1+2:</u> 113 <u>Stage 3:</u> 155 <u>Stage 4 (cirrhosis):</u> 144
- 182 died or underwent liver transplant over up to 12 years (median 5 years)

- data("pbc", package="survival") in R

Example: Primary Biliary Cholangitis & Death (2)

- Kaplan-Meier curves:

- Can do log-rank test, but…
  - → Age is potential confounder!
    - risk of death
    - fibrosis progression (time)
  - → Overall quantification of the difference
- Cox regression let's us do all this **and** test for differences between stages



Example: Primary Biliary Cholangitis & Death (3)

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \ \beta_3[stage\ 3] + \beta_4[stage\ 4]$

Estimates:	Hazard Ratios:	Significance tests:

Example: Primary Biliary Cholangitis & Death (3)

Stage 1–2:

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \boxed{\beta_3 \cdot 0 + \beta_4 \cdot 0}$

Estimates:	Hazard Ratios:	Significance tests:

Example: Primary Biliary Cholangitis & Death (3)

Stage 3:

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \boxed{\beta_3 \cdot 1 + \beta_4 \cdot 0}$

Estimates:	Hazard Ratios:	Significance tests:

Example: Primary Biliary Cholangitis & Death (3)

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \boxed{\beta_3 \cdot 0 + \beta_4 \cdot 1}$

Stage 4:

Estimates:

Hazard Ratios:

Significance tests:

Example: Primary Biliary Cholangitis & Death (3)

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \beta_3[stage\ 3] + \beta_4[stage\ 4]$

Estimates:

	coef	exp(coef)
age	0.011160	1.011222
sexf	-0.276887	0.758140
stage3	0.588064	1.800500
stage4	1.463718	4.321997

Hazard Ratios:

Significance tests:

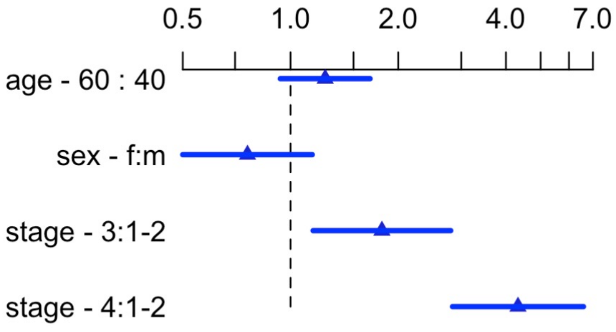
Example: Primary Biliary Cholangitis & Death (3)

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \beta_3[stage\ 3] + \beta_4[stage\ 4]$

Estimates:

	coef	exp(coef)
age	0.011160	1.011222
sexf	-0.276887	0.758140
stage3	0.588064	1.800500
stage4	1.463718	4.321997

Hazard Ratios:



Significance tests:

HR age 60 vs 40: $e^{0.011\cdot(60-40)} = 1.25$

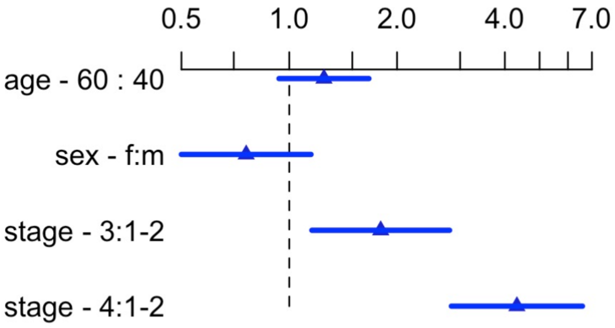
Example: Primary Biliary Cholangitis & Death (3)

- Cox model: $\ln(HR) = \beta_1[age] + \beta_2[sex\ (f)] + \beta_3[stage\ 3] + \beta_4[stage\ 4]$

Estimates:

	coef	exp(coef)
age	0.011160	1.011222
sexf	-0.276887	0.758140
stage3	0.588064	1.800500
stage4	1.463718	4.321997

Hazard Ratios:



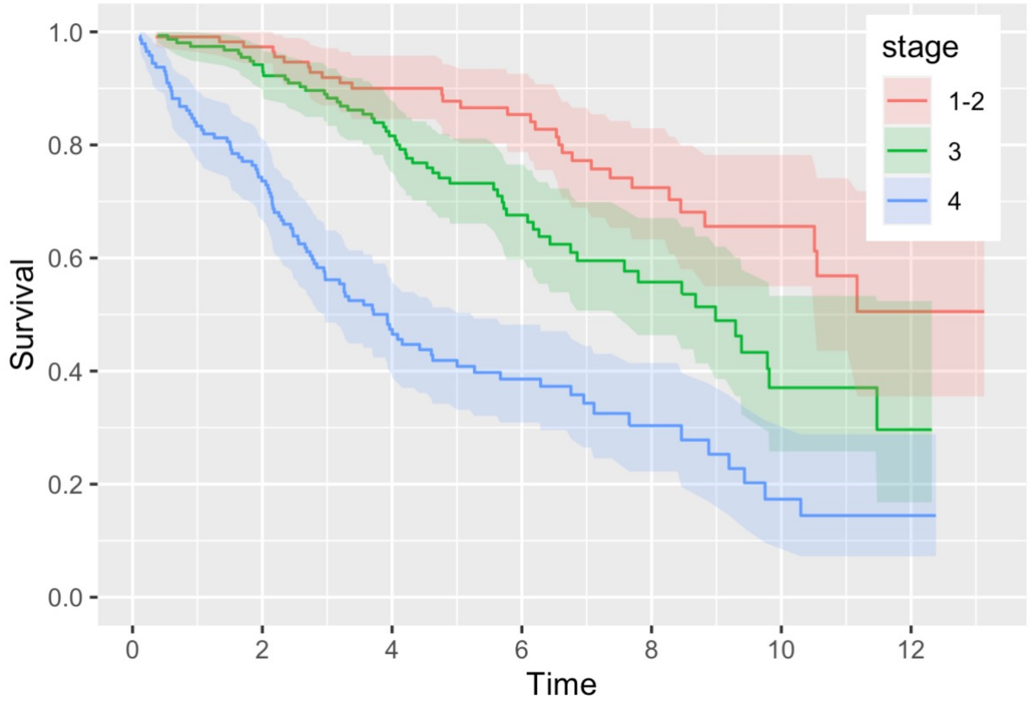
Significance tests:

Factor	Chi-Square	d.f.	P
age	2.29	1	0.1301
sex	1.70	1	0.1929
stage	55.57	2	<.0001
TOTAL	69.00	4	<.0001

HR age 60 vs 40: $e^{0.011\cdot(60-40)} = 1.25$

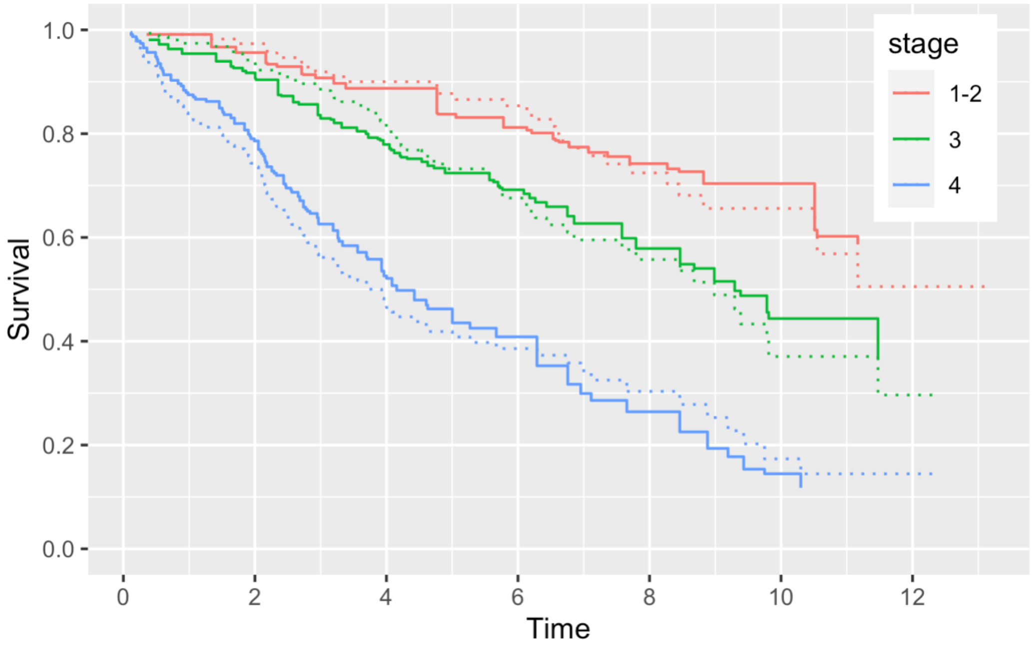
Checking the Proportional Hazards Assumption

- **Do the curves look like they have proportional hazards?**



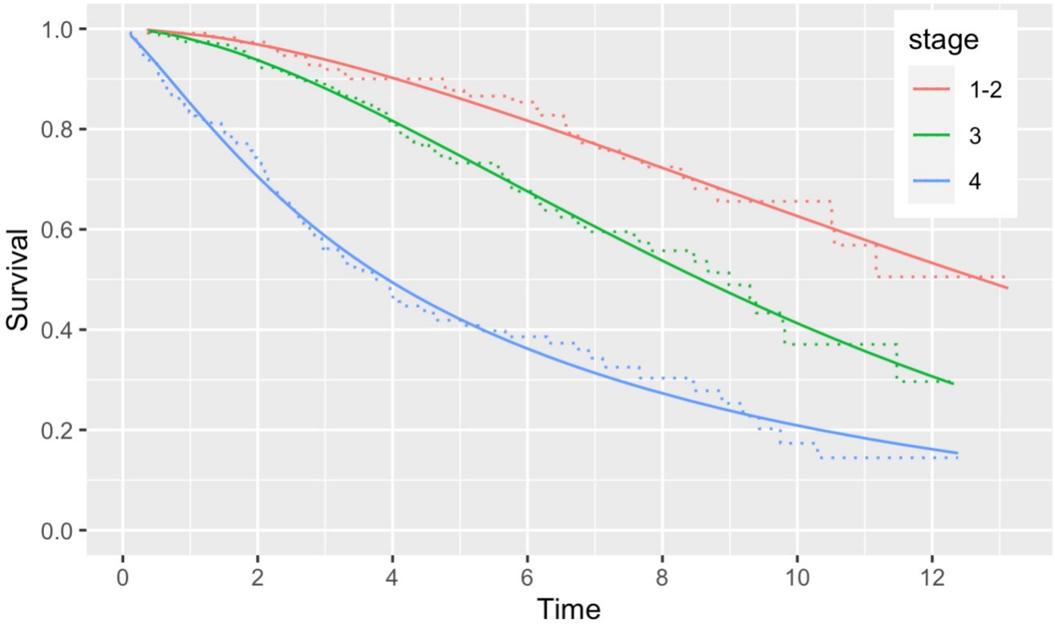
Checking the Proportional Hazards Assumption

- **Do the curves look like they have proportional hazards?**
- If they did, according to the Cox model:
- Important to check!
- What to do?
  → Flexible parametric survival model (advanced)



Checking the Proportional Hazards Assumption

- **Do the curves look like they have proportional hazards?**
- If they did, according to the ~~Cox model:~~ Flexible parametric model:
- Important to check!
- What to do?
  → Flexible parametric survival model (advanced)

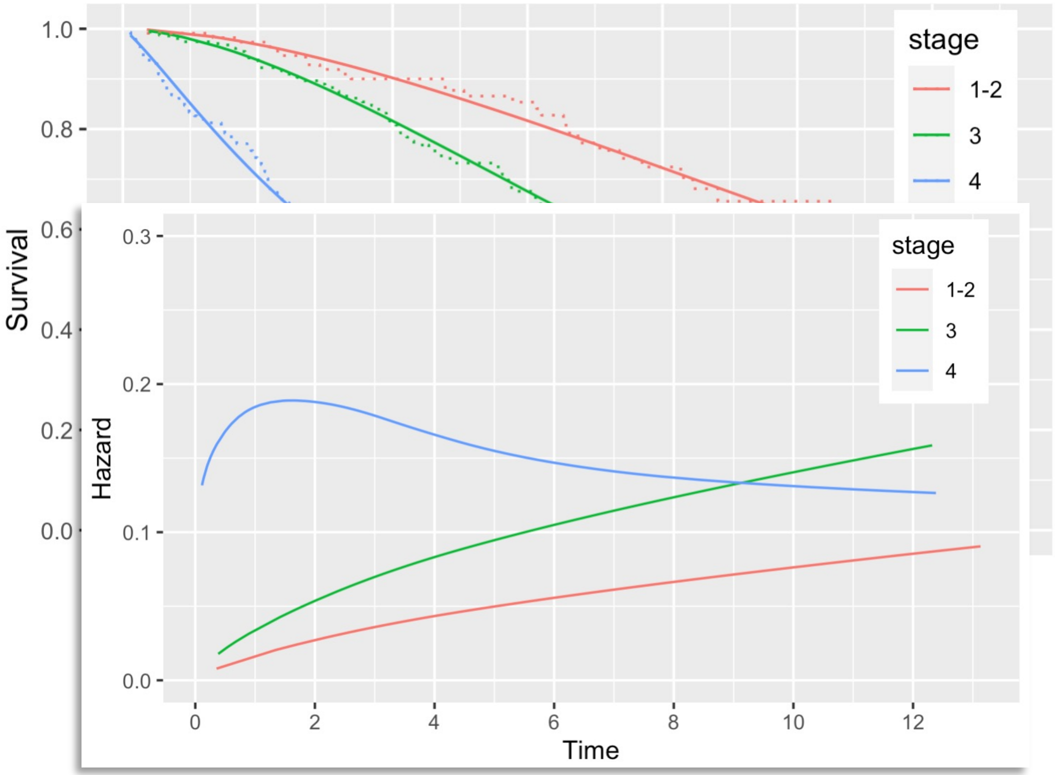


Checking the Proportional Hazards Assumption

- **Do the curves look like they have proportional hazards?**

- If they did, according to the ~~Cox model:~~ Flexible parametric model:
- Important to check!

- What to do?
  → Flexible parametric survival model (advanced)



Summary

- The most important concepts of survival analysis
  1. The type of data we use:
     - Time-to-event
  2. The survival function & the cumulative incidence function:
     - Summary statistics
  3. The hazard rate function
     - The mechanism behind the survival
  4. The Cox proportional hazards regression model
     - All the regression goodness!

KAROLINSKA INSTITUTET
ANNO 1810

Karolinska
Institutet