
Max Planck Odense Center on the Biodemography of Aging

University of Southern Denmark

# Introduction to Survival Analysis

Virginnia Zarulli, Adam Lenart

MaxO - SDU

# Survival Analysis

- What is survival analysis?
  - Statistical techniques to analyze time-to-event data.
- How is it different from other techniques?
  - It concentrates on duration and explicitly models whether the event happens.
- What types of answers can it provide?
  - Hazard ratios.
  - Time ratios.
  - Survival probabilities.

# Sample research questions

- What is the impact of chemotherapy on relapse-free survival in patients with osteosarcoma?
- Do statins reduce the risk of major vascular events?
- Can risk-adjusted therapy of acute lymphoblastic leukemia improve survival?

# Data requirement

Survival analysis require information at least on

- duration (years, weeks, seconds...)
- event (death, relapse, ...)

and typically on

- other covariates (treatment, sex, BMI, etc.).

Response variable: combination of data concerning the duration and data containing the event indicator (1 when the event takes place and 0 when not).

# Example data

Observation of a cohort of cancer patients for 12 months since tumor diagnosis. The event is death. Half of them are treated with a new drug, the other half with the usual treatment:

| id | duration | status | treatment |
|---|---|---|---|
| 111 | 12 | 0 | new |
| 112 | 10.6 | 1 | new |
| 113 | 3.8 | 1 | old |
| 114 | 9 | 1 | new |
| 115 | 6.1 | 0 | old |
| 116 | 11.9 | 1 | old |
| 117 | 1.5 | 1 | new |
| 118 | 5 | 0 | new |

- The information about the process is evaluated **conditionally** on the available information at any specific moment.
- At every time unit the risk set changes and the estimation is based on those who are still alive (have not experienced the event yet) at that moment.

| id | duration | status | treatment |
|---|---|---|---|
| 111 | 12 | 0 | new |
| 112 | 10.6 | 1 | new |
| 113 | 3.8 | 1 | old |
| 114 | 9 | 1 | new |
| 115 | 6.1 | 0 | old |
| 116 | 11.9 | 1 | old |
| 117 | 1.5 | 1 | new |
| 118 | 5 | 0 | new |

- Time 0 - risk set: 8 individuals.
- Time 1.5 - 1 death; death rate: 1/8; risk set from now on will be 7.
- Time 3.8 - 1 death; death rate 1/7; risk set from now on will be 6.
- and so on...

The most important characteristics of duration data are the presence of **censoring** and **truncation**:

- right censoring:
  when at the end of the observation period the subject has not experienced the event yet;
- left censoring:
  when the subject has already experienced the event before the observation started;
- interval censoring:
  when we know that the event has taken place within an interval, but we don't know the precise time/age at which this happened.

The most important characteristics of duration data is the presence of **censoring** and **truncation**:

- left truncation:
  when the subjects have been at risk of the event before the observation started but we don't know anything else of that period;
- right truncation:
  when individuals enter the sample only if the event has happened to them.

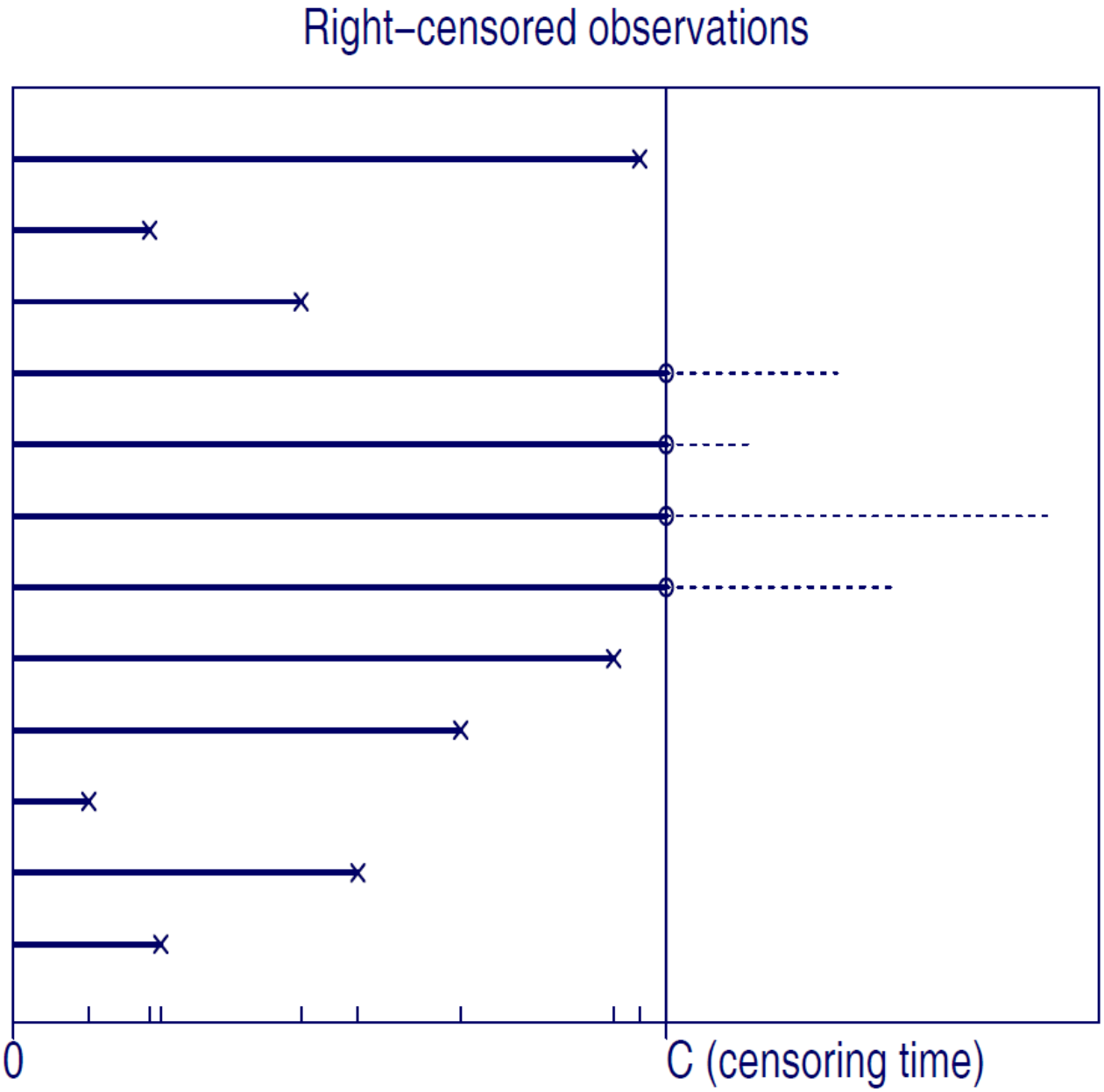
Typically, survival data is right censored, often left truncated and sometimes interval censored. Left censoring and right truncation occurs rarely.

Because of censoring and truncation we cannot

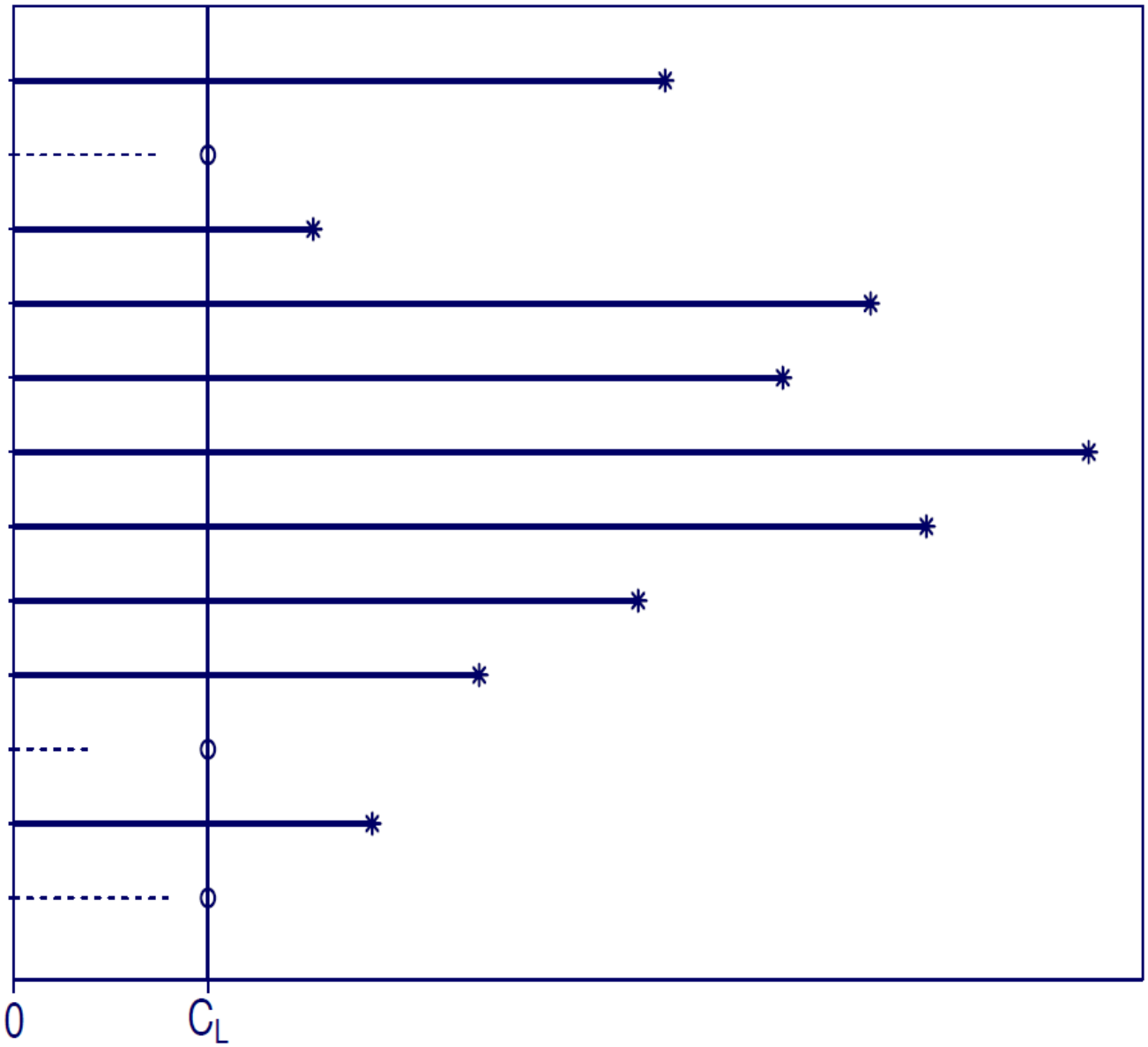
- calculate mean, standard deviation as usual
- use t-test, ANOVA, linear regression, logistic regression, etc.

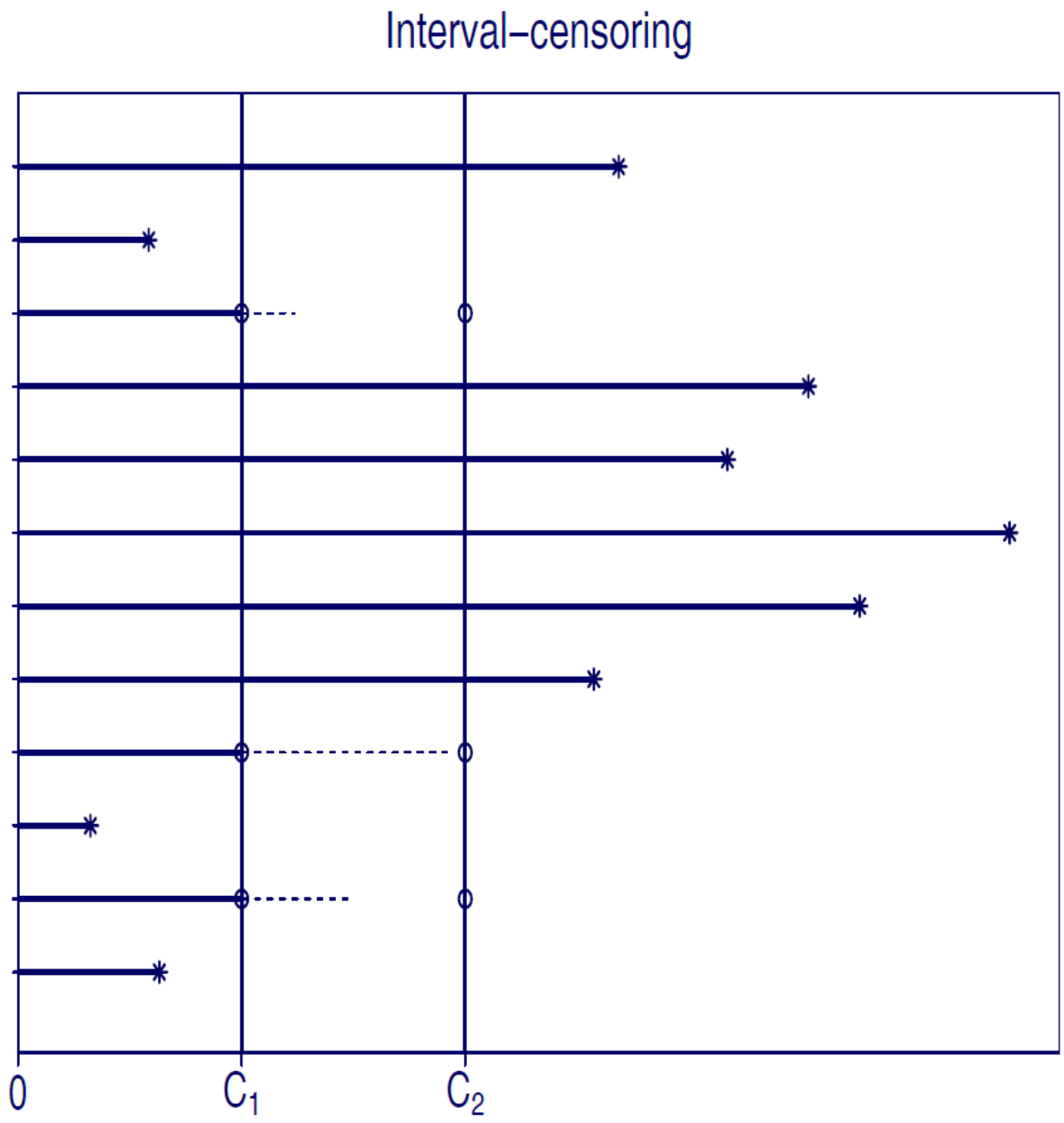
Why?

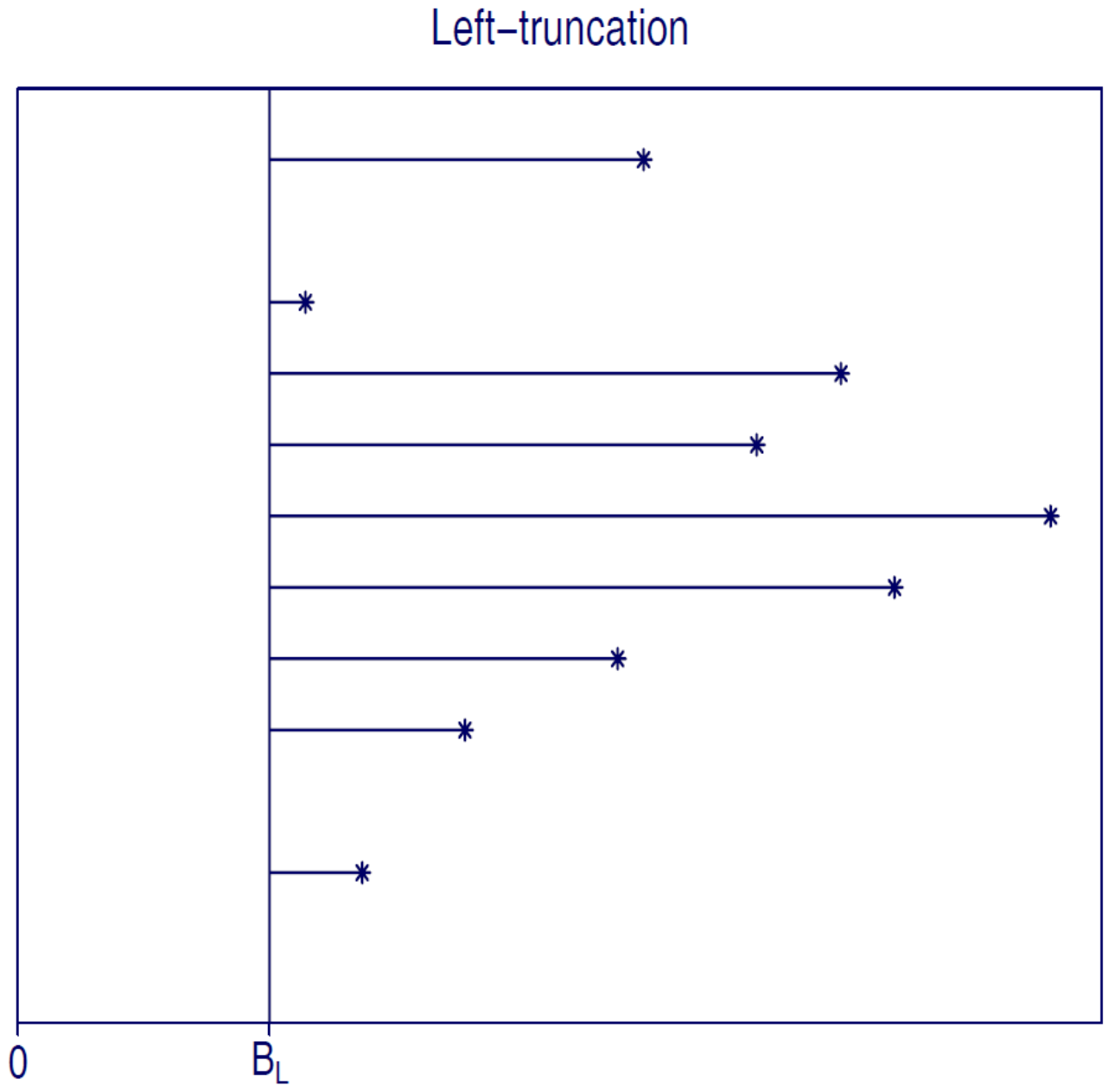
- Bias due to ignoring censoring or truncation, e.g., censoring as missing data underestimates survival.

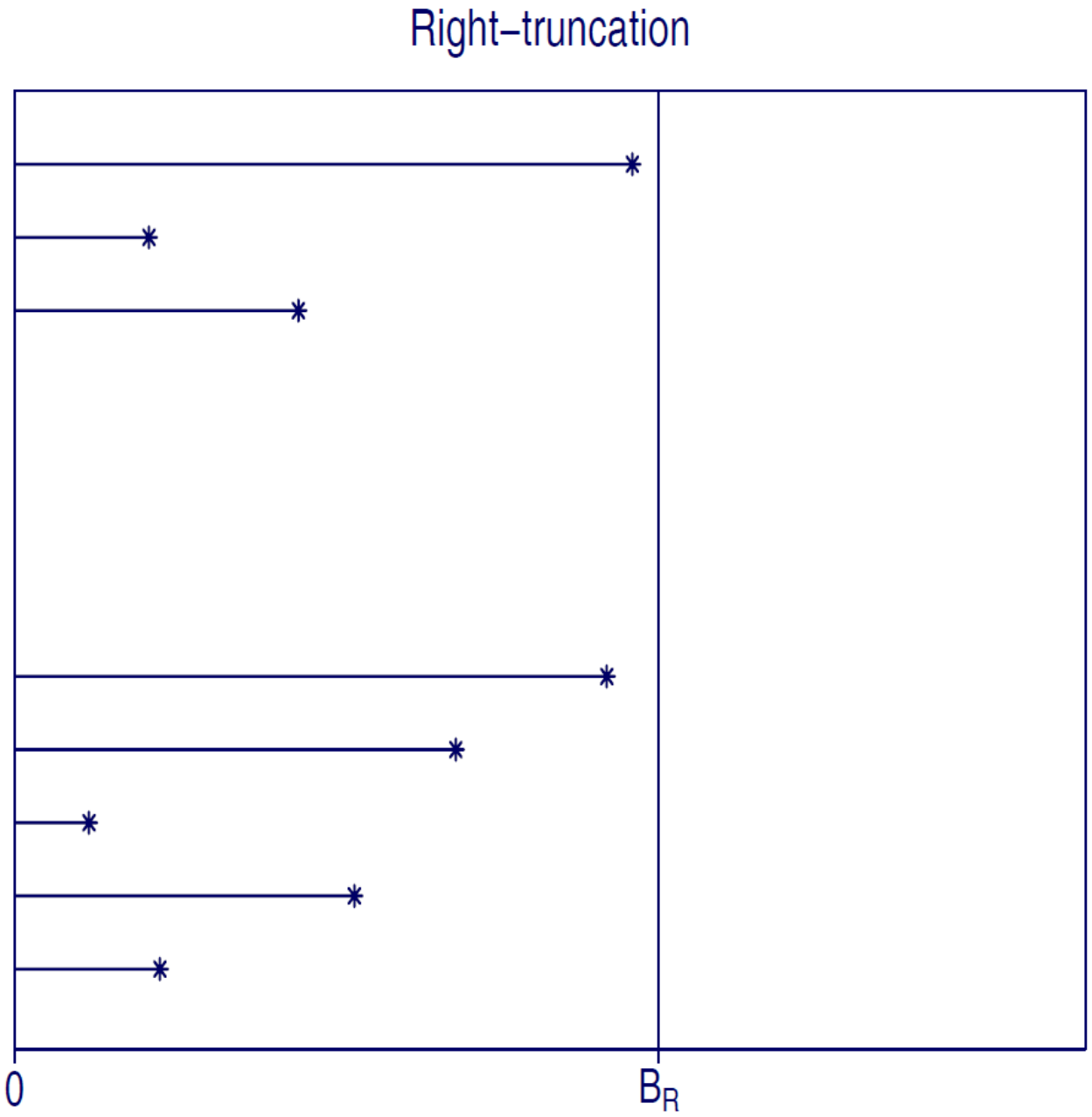

Right−censored observations
0
C (censoring time)

## Left–censoring




Interval−censoring
0
$C_1$
$C_2$


Left−truncation
0
$B_L$


Right−truncation
0
$B_R$

# Quick exercise

Which observation schemes do the following belong to?

A study wishes to test a new antibiotic treatment on patients with tuberculosis. Survival time is counted from the start of treatment for 6 months. The event of interest is cure from tuberculosis.

- Patient is cured from tuberculosis within 4 months.
- Patient still shows symptoms after 6 months.
- Patient becomes convinced that homeopathic remedies offer a better chance of cure than the antibiotic treatment and quits the study.

# Quick exercise

Which observation schemes do the following belong to?

A study wishes to test a new antibiotic treatment on patients with tuberculosis. Survival time is counted from the start of treatment for 6 months. The event of interest is cure from tuberculosis.

- Patient is cured from tuberculosis within 4 months.
  - Not-censored
- Patient still shows symptoms after 6 months.
  - Right censored
- Patient becomes convinced that homeopathic remedies offer a better chance of cure than the antibiotic treatment and quits the study.
  - Right censored

# Quick exercise

Which observation schemes do the following belong to?

Scientists find a way to identify the causes of lung cancer. They wish to estimate the risk of developing lung cancer due to exposure to radon. Detectors measure the radon levels in all houses in the study area from January 1, 2015 until December 31, 2024. The study offers lung cancer diagnosis frequently as well.

- A person is not diagnosed with lung cancer due to radon until the end of the study.
- A person is diagnosed with lung cancer due to radon on May 1, 2016.
- A person is diagnosed with lung cancer due to smoking on September 1, 2018.

# Quick exercise

Which observation schemes do the following belong to?

A study wishes to estimate the risk of developing lung cancer due to exposure to radon.

- A person is not diagnosed with lung cancer due to radon until the end of the study.
  - Left truncated and right censored.
- A person is diagnosed with lung cancer due to radon on May 1, 2016.
  - Left truncated.
- A person is diagnosed with lung cancer due to smoking on September 1, 2018.
  - Left truncated and right censored.

# Quick exercise

Which observation schemes do the following belong to?

- A known population of drug addicts is tested every half-year for HIV infection.
- Estimating the incubation period of ebola. Only infected people are observed.
- Estimating time to conception since intention to conceive. A woman is already found pregnant at the start of the study.

# Quick exercise

Which observation schemes do the following belong to?

- A known population of drug addicts is tested every half-year for HIV infection.
  - Interval censored
- Estimating the incubation period of ebola. Only infected people are observed.
  - Right truncated
- Estimating time to conception since intention to conceive. A woman is already found pregnant at the start of the study.
  - Left censored

Main goals of survival analysis:

- to analyze $h(t)$ (hazard function), $S(t)$ (survival function) or one of the other related functions (for more formal details see the appendix).
- to measure the effect of covariates on one of these functions

$$h(t_i, x_i) = h_0(t)e^{x_i\beta}$$

How?

- non-parametric models (Kaplan-Meier)
- semi-parametric models (Cox models).
- parametric models (exponential, Weibull, Gompertz ...).

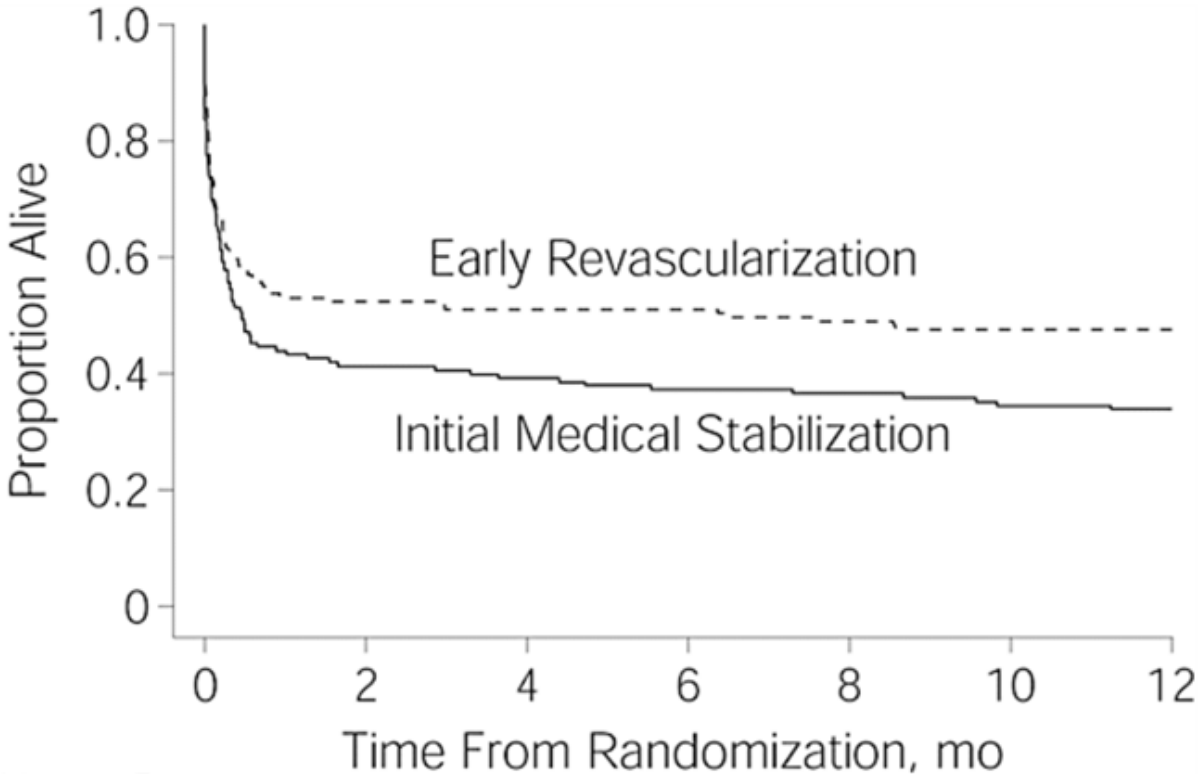
# Survival function

The survival function, $S(t)$ shows the probability of surviving longer than a given time $t$,

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

for duration $T$ and point in time $t$.

# Survival function




Hochman et al. 2001: *One-Year Survival Following Early Revascularization for Cardiogenic Shock*, JAMA. 2001; 285(2):190–192.

# Hazard function and cumulative hazard function

The hazard function, $h(t)$ indicates the (instanteneous) risk of death given having survived until time $t$.

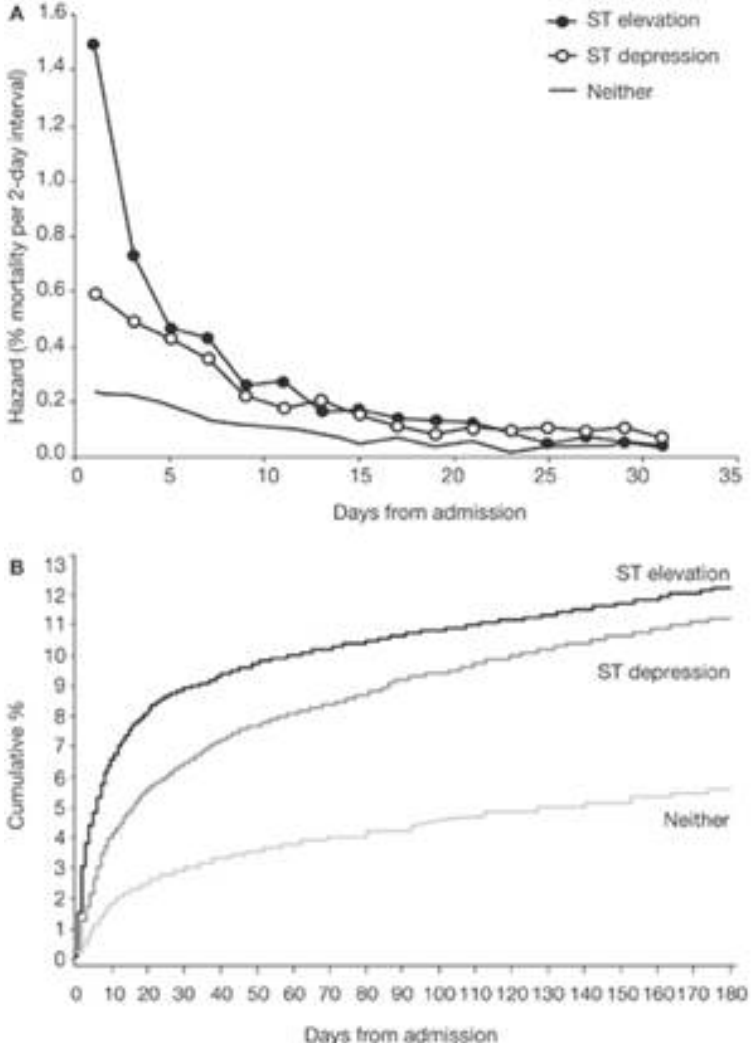
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Sometimes it is also called the death rate. Because it can be difficult to estimate it empirically, usually the cumulative hazard function, $H(t)$ is used:

$$H(t) = \int_0^t h(x)dx$$

It could be interpreted as "how many times a person would die (if she could die more than once) until time $t$".

# Hazard function and cumulative hazard function




Fox et al. 2008: *Time course of events in acute coronary syndromes*. Nat Clin Pract Cardiovasc Med (2008) 5, 580-589.

# Survival analysis methods

## Non-parametric models

- Exploratory analysis, descriptive results. Usually it is the first step of analysis.
- Can be done over the whole population or stratified by key categorical covariates.
- We can get empirical survival and cumulative hazard functions but we can't estimate the impact of the covariates.

# Survival analysis methods

## Semi-parametric models

- When we **do not** have any information on or care about how the risk of experiencing the analyzed event changes over time (baseline hazard).
- Only the influence of the covariates are of interest.
- We can get a rough estimate of the baseline hazard.

# Survival analysis methods

## Parametric models

- When we do have information on how the risk of experiencing the analyzed event changes over time.
- When we are interested in how the risk of experiencing the analyzed event changes over time.
- The influence of the covariates are also of interest.
- We can calculate hazard and survival functions.

# Summary

- Survival analysis is used to analyze time-to-event data.
- Information on censoring and truncation can be included. Ignoring censoring or truncation leads to biased estimates.
- Survival analysis can be either
  - non-parametric: descriptive results,
  - semi-parametric: estimating the effect of covariates,
  - parametric: estimating the effect of covariates and baseline hazard.

# Appendix

Given a positive random variable $T$ for the survival time ($T \geq 0$):

- probability density function: $f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$
- cumulative distribution function:
  $F(t) = P(T \leq t) = \int_0^t f(x)dx$
- survival function: $S(t) = 1 - F(t) = P(T \geq t) = \int_t^\infty f(x)dx$
- hazard function:
  $h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$
- cumulative hazard function: $H(t) = \int_0^t h(x)dx$

# Appendix

$$f(t) = -S'(t)$$

$$h(t) = \frac{f(t)}{S(t)}$$

$$S(t) = e^{-\int_0^t h(x)dx} = e^{-H(t)}$$

$$h(t) = -\frac{d}{dt} \ln S(t)$$

$$H(t) = -\ln S(t)$$