SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

COURSE NAME : 19CS407-DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit II – Statistics and Prescriptive Analytics

Topic : Survival Analysis

Introduction to Survival Data Analysis

Overview

- What is survival analysis?
  - Terminology and data structure
- Kaplan-Meier methods (non-parametric)
- Cox proportional hazards regression model (semi-parametric)
- Competing Risk

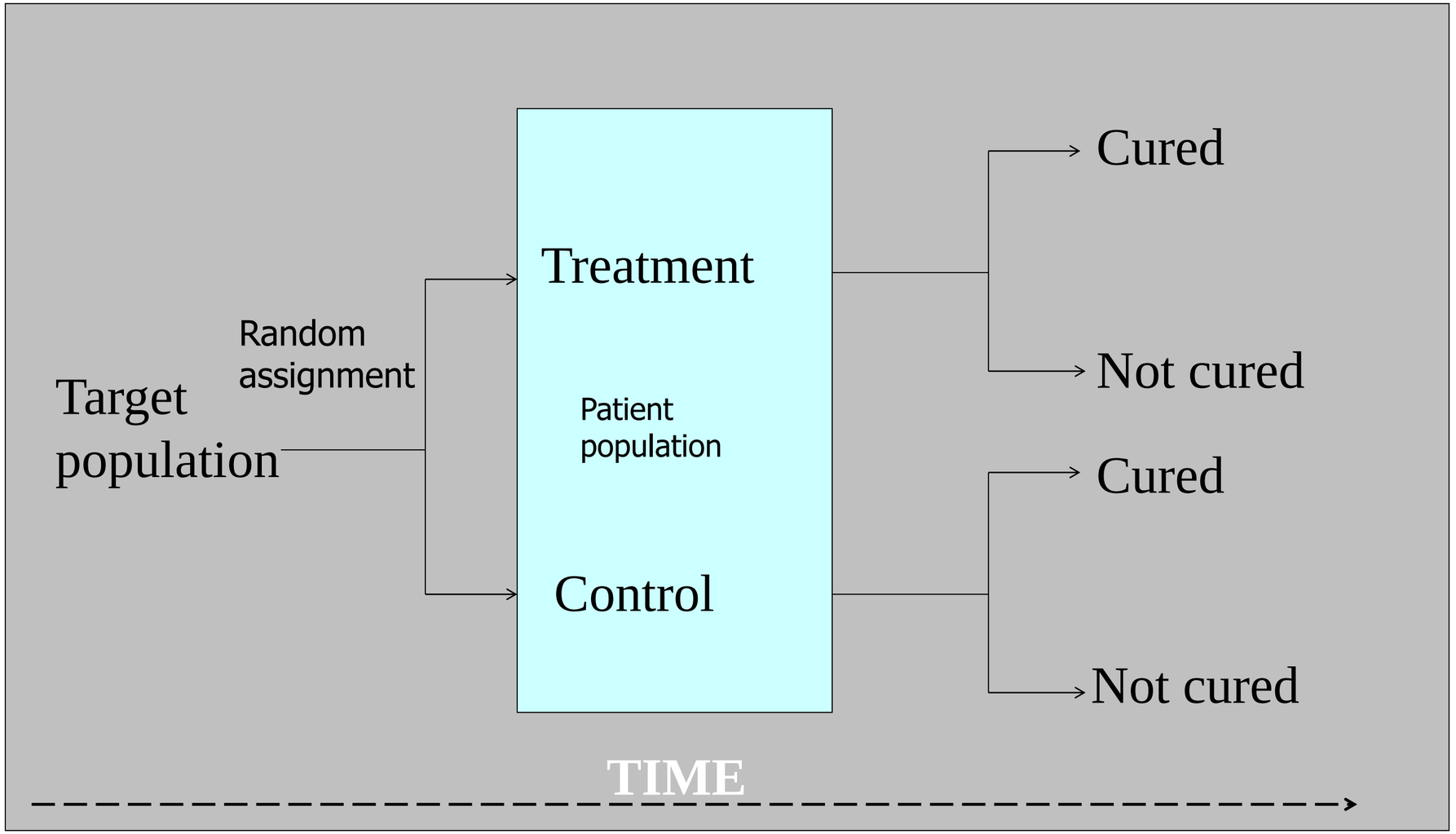
What is Survival Analysis

- **Time to Event:** In many studies, the primary endpoint is the time from entering a study until a subject has a particular event occurs.
- **Medical Research:**
  - Time to death
  - Time to relapse of a disease
  - Time re-hospitialization
- **Engineering, business, etc:**
  - Engineer measures the time until failure of a product or component (mean time to failure, MTTF)
  - Credit card company measures the length of time people keep using the credit card

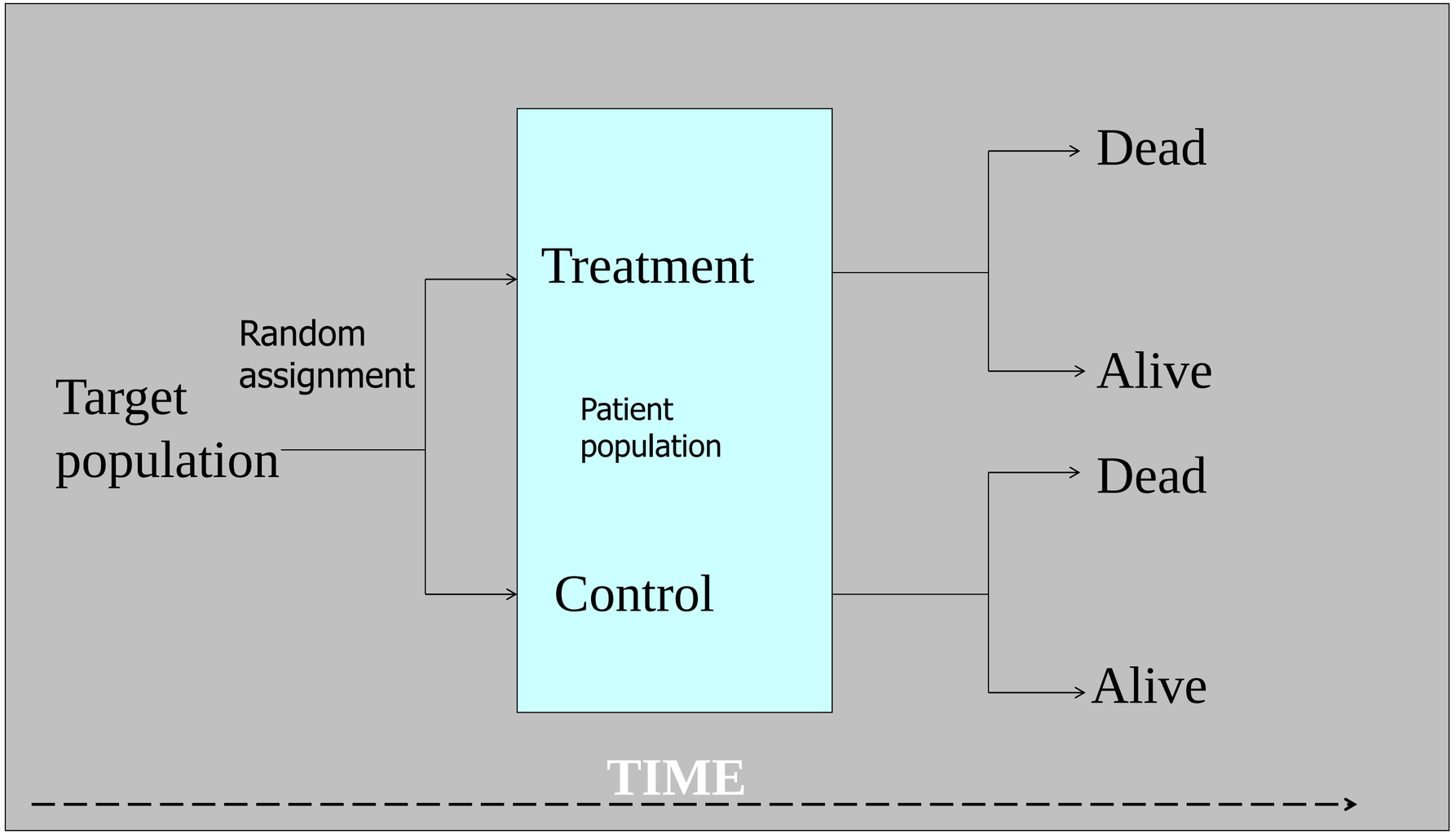
What is Survival Analysis

- Kind of survival studies
  - Clinical trials
  - Prospective cohort studies
  - Retrospective cohort studies

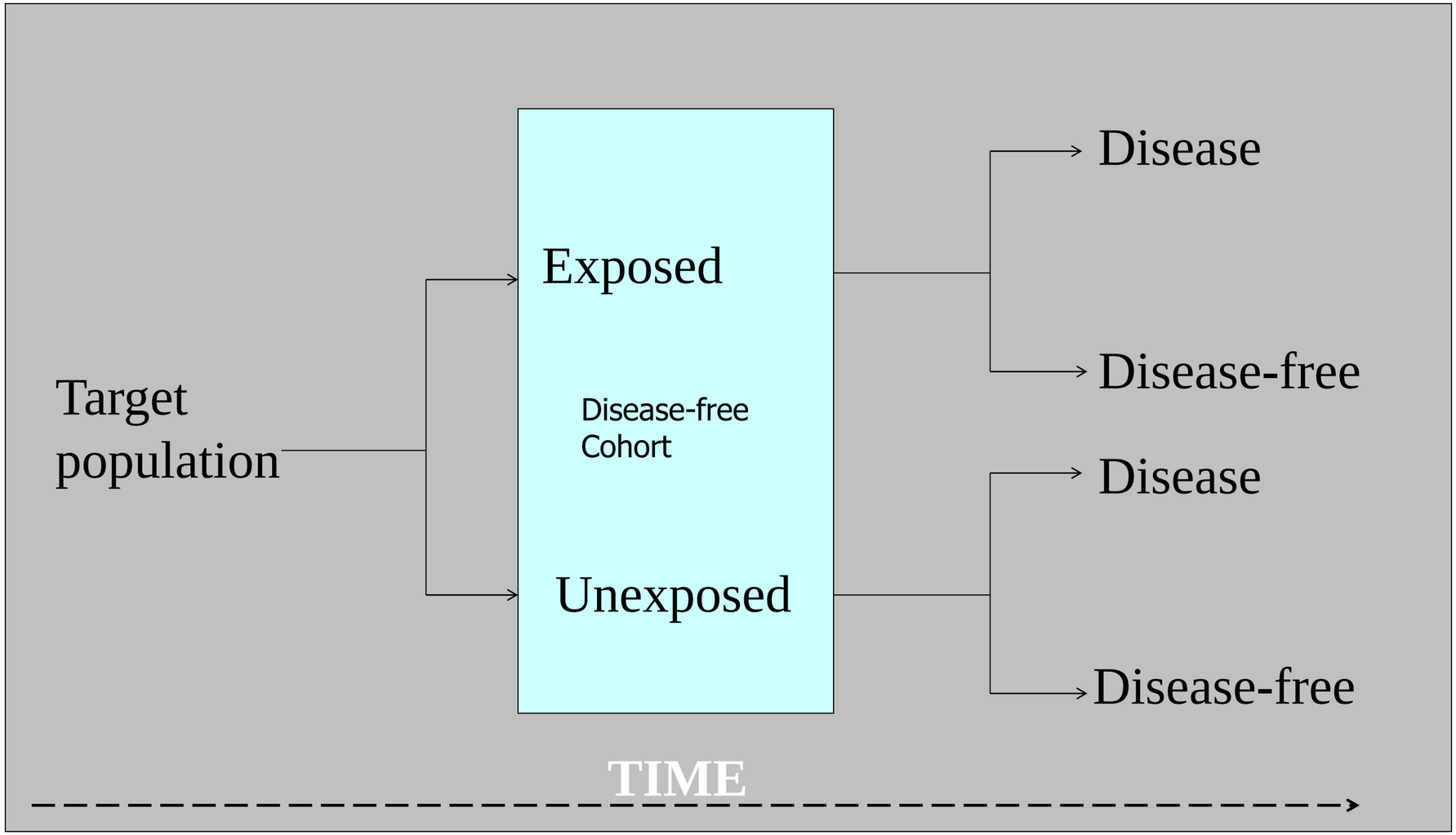
Randomized Clinical Trial (RCT)



Randomized Clinical Trial (RCT)



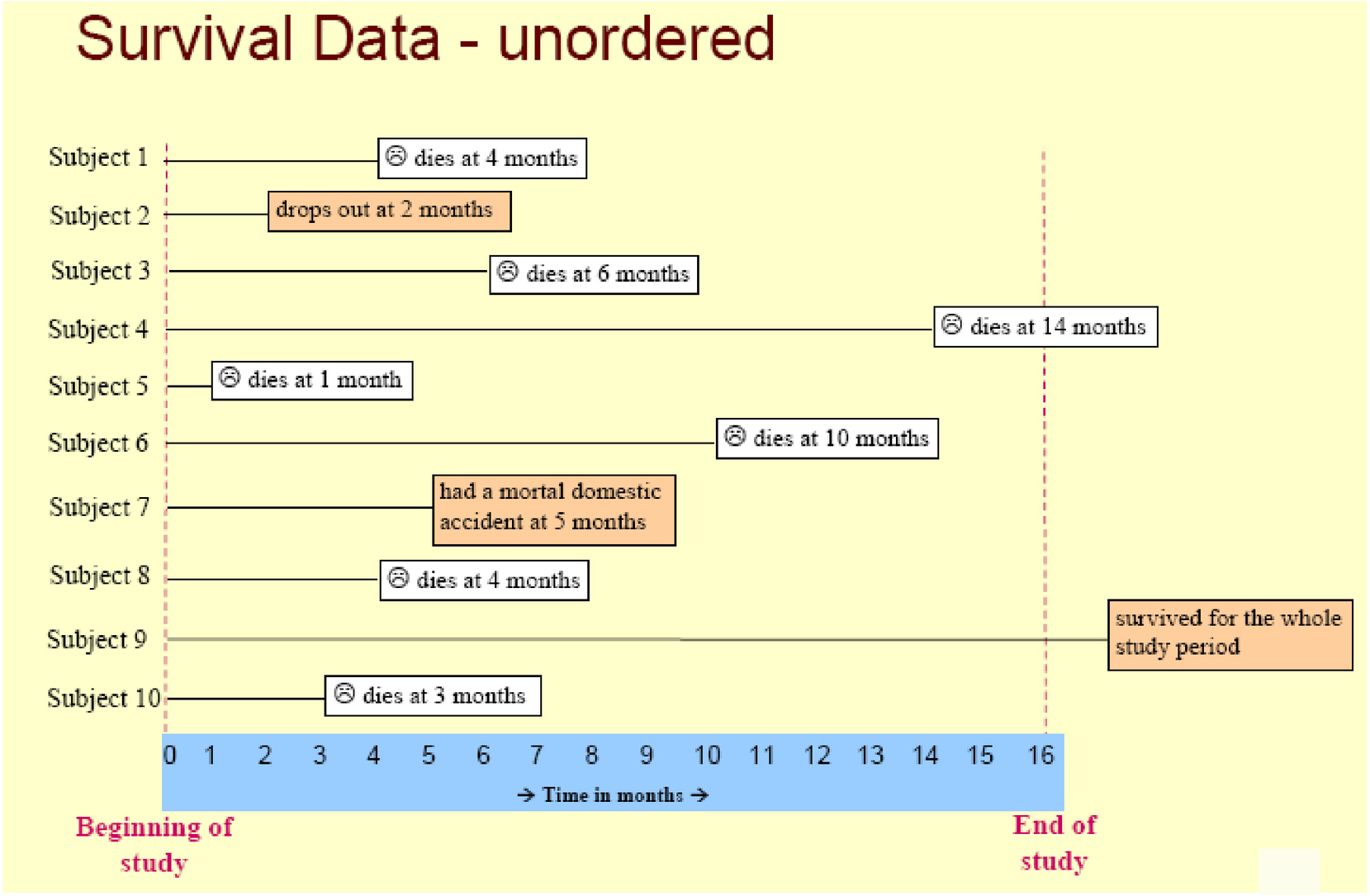
Cohort Study (Prospective/Retrospective)



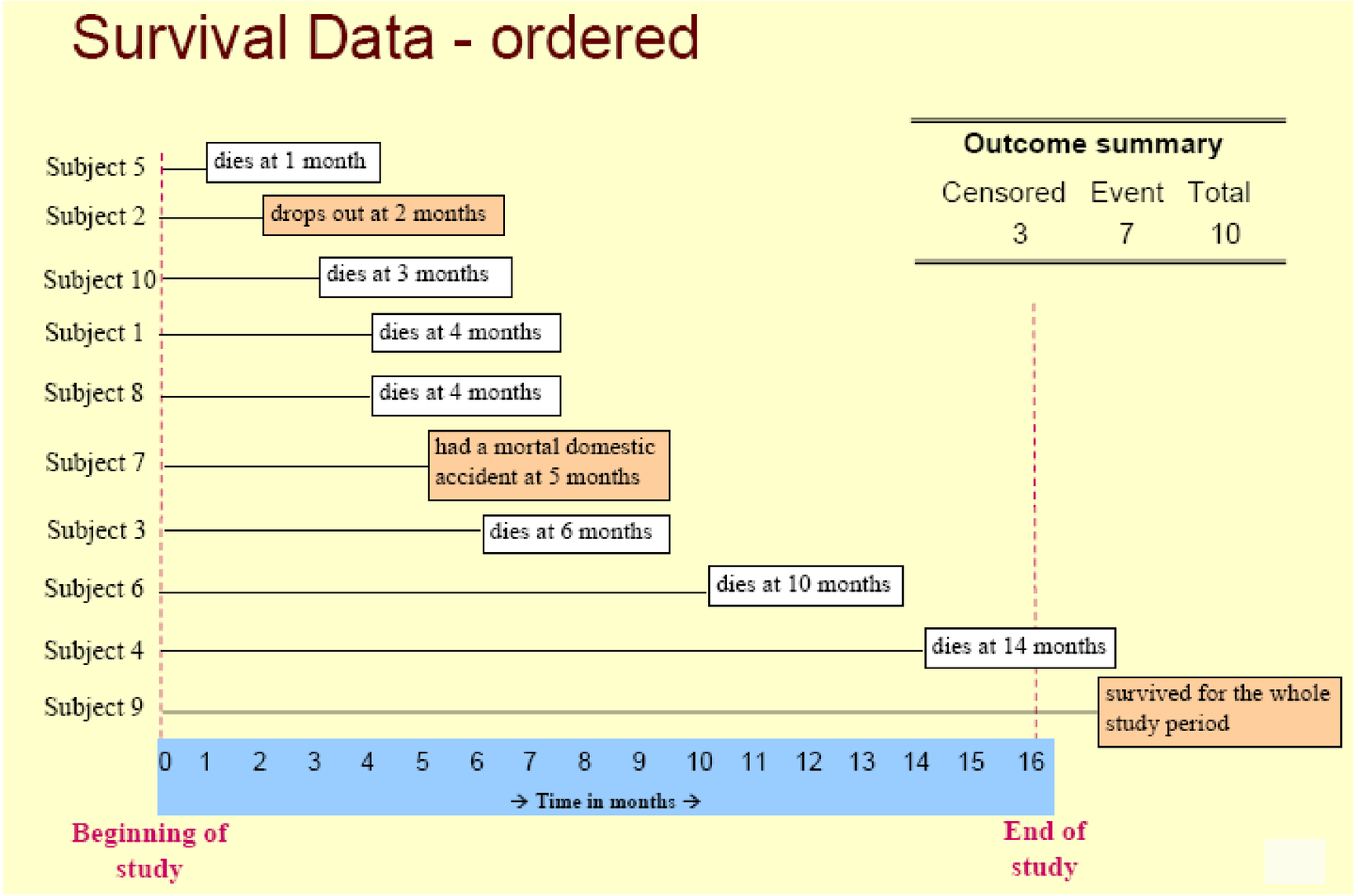
Time to Event: Example

- 10 patients with squamous cell carcinoma are recruited to receive specific treatment.
- The objective is to investigate the survival probability of the patients under this treatment.
  - Event of interest: death
- They were followed up to 16 months.
  - Duration of the study: 16 months
  - Time scale: months
- Consider right censored observations.

Time to Event: Example



Time to Event: Example



Censoring

- Data are typically subject to censoring when the event does not occur within the study observation time
- Survival data are characterized by incomplete observation: **Censoring.**
  T-test or ANOVA cannot be used because of the censored data.
- Most common is **right censoring**. Subject leaves the study before an event occurs.
  - the study ends
  - the individual withdrew from the study/lost to follow-up
  - the individual died from **other causes**
  - the individual is ineligible for research because of other reasons

Time to Event

- Non-negative, $T \geq 0$
- To correctly collect a time to event, we need:
  - An unambiguous time origin
  - A time scale (day, month, year)
  - Definition of the event of interest

Example of Data Structure

Subject ID	Survival Time	Event (0=no/1=yes)	Gender	Treatment Group (0=Placebo/1=Treatment)
1	4	1	F	0
2	2	0	F	0
3	6	1	M	1
4	14	1	F	1
5	1	1	M	1
6	10	1	M	1
7	5	0	F	0
8	4	1	M	0
9	16	0	F	1
10	3	1	M	0

Survival Analysis

- Survival analysis is concerned with studying the time between entry to a study and a subsequent event.
  - Also called “time to event analysis”

- Survival analysis attempts to answer questions such as:
  - Which fraction of a population will survive past a certain time ?
  - At what rate will they fail ?
  - At what rate will they present the event ?
  - How do particular factors benefit or affect the probability of survival ?

Survival Analysis

- Objectives
  - To estimate time to event for a group of individuals.
  - To compare time to event between two or more groups.
  - To assess the relationship between explanatory variables and time to event.

Survival Analysis – advantages

- Why not compare mean time to event between groups using a t-test or linear regression?
  - Ignores censoring

- Why not compare proportion of events in your groups using logistic regression?
  - Ignores time
  - Ignores censoring

Survival analysis accounts for censored observations as well as time to event.

Survival Analysis – methods

- Non-parametric estimation
  - Within-group survival: **Kaplan-Meier**
  - Between-group comparison: **Log-rank Test**

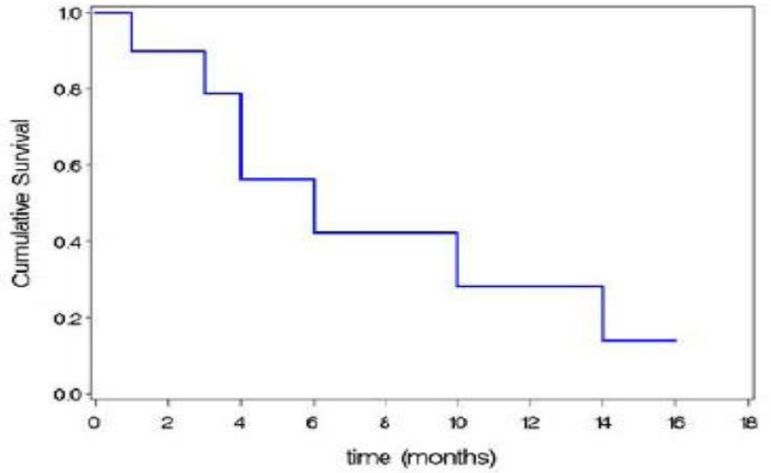
- Semi-parametric estimation model
  - **Cox proportional hazard model** (allows **explanatory** variables)
    - Hazard: The event of interest. Usually it is believed to be harmful, e.g. death, relapse of a disease, re-hospitalization, failure of the product or part, etc

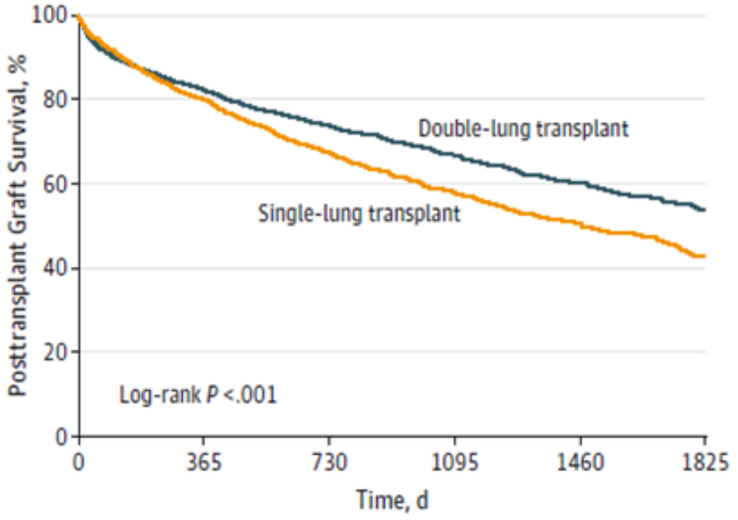
- Parametric models: Exponential, Weibull distribution, etc.
  (won't cover today)

Kaplan-Meier Survival Method

- Non-parametric estimate of survival probability
- Commonly used to describe survival-ship of a study population
- Intuitive graphical presentation

- Cumulative survival characteristics
- Estimation of median survival time
- Commonly used to compare two study population





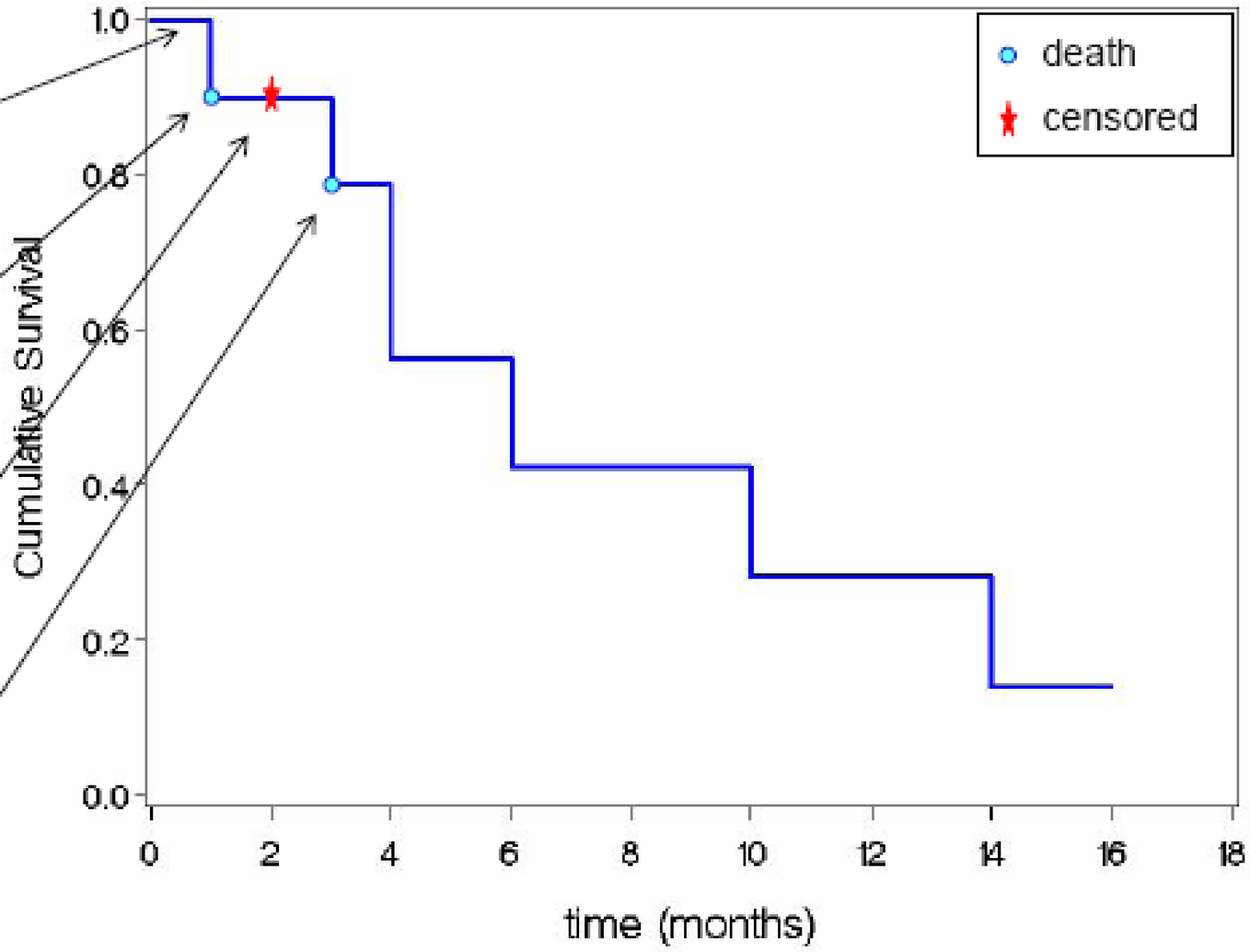
Kaplan-Meier Plot

N° subjects **at risk** for death =10
Fraction surviving before 1 month =10/10 ⇒ 100%

Subject dies at 1 month
Hazard = 1/10
Fraction surviving =9/10

Subject drops out of the study at 2 months.
Subjects at risk for death = 8

Subject dies at 3 months
Hazard = 1/8
Fraction surviving =7/8



Kaplan-Meier Survival Probability

Survival Function:

$$\hat{S}_0(t) = \prod_{j=1}^{t}\left(1 - \frac{E_j}{E_j + S_j}\right)$$

Survival Probability at

1-month = 1-1/10 = 0.9

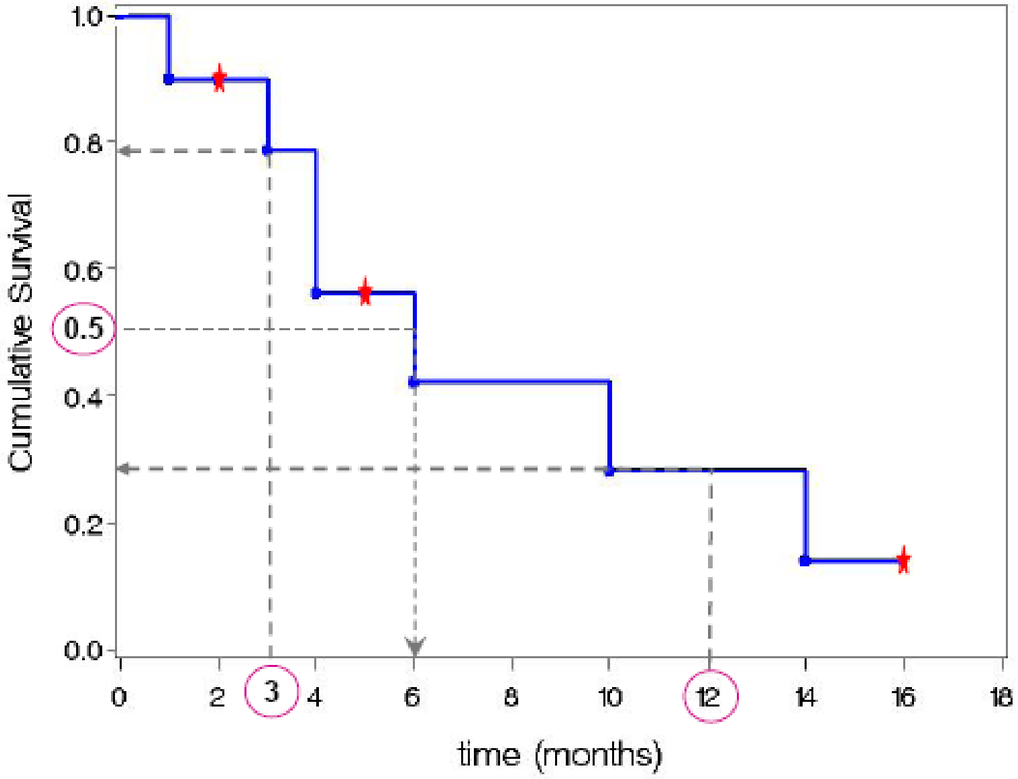
3-month = (1-1/10)*(1-1/8)=0.788

4-month = (1-1/10)*(1-1/8)*(1-2/7) = 0.56

1-year survival rate = 28%

Median survival time= 6 months

The K-M curve takes a step down when there is an event.



Comparison of groups-Logrank Test

- Logrank Test :
  - For comparison of survival distributions between groups
  - The groups are defined by categorical covariates. Can be more than 2 groups.

    e.g. Therapy : treatment, placebo

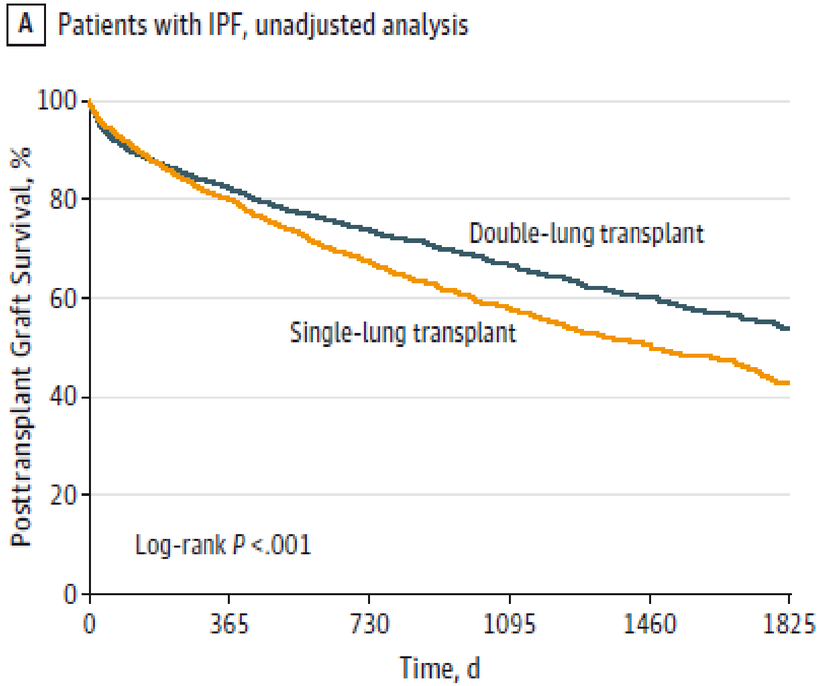
    Gender : male, female

    Age group : ≤ 40, ≥ 40

- Bad performance when two survival curves are crossing.
  - The logrank test has better performance under the assumption of **proportional hazards**.

Proportional hazards: The hazard functions for any two individuals at any point in time are proportional, and does not change with time *t*.

Logrank Test-Example



No. at risk						
Single-lung transplant	2010	1313	890	601	400	239
Double-lung transplant	2124	1373	945	646	424	240

Comparison of post-transplant death and graft failure probability in IPF patients with lung transplantation

Comparison group:

single transplantation

vs.

double transplantation

Logrank test:

p-value < 0.001

Single- vs Double-Lung Transplantation in Patients With Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary Fibrosis Since the Implementation of Lung Allocation Based on Medical Need
JAMA. 2015;313(9):936-948. doi:10.1001/jama.2015.1175

Log-rank Test- Another example

Comparison of composite of cardiovascular death, MI, or severe recurrent ischemia in patients with acute MI:

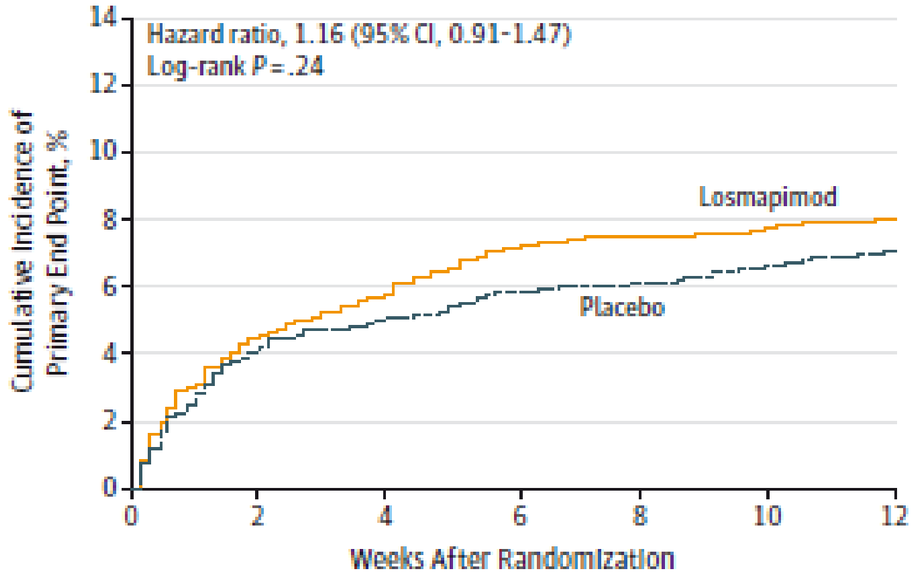
Comparison group:

Losmapimod vs. Placebo

Logrank test:

p-value =0.24

Figure 2. Kaplan-Meier Curves for the Primary End Point



Effect of Losmapimod on Cardiovascular Outcomes in Patients With Acute MI *JAMA*. 2016;315(15):1591-71599

Cox Regression Survival Model

- Allows for prognostic factors.
- Explore the relationship between survival and explanatory variables.
- Models and compares the hazards for different groups/factors (explanatory variables).
- Important assumption:
  - Survival curves with proportional hazards

    (risk of an event at different time points).

Cox Regression Survival Model

- $h(t,X) = h_0 \exp(\beta X)$

  $= h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$

- Hazard ratio:

$$\frac{h_1(t)}{h_0(t)} = \exp(\beta)$$

- Constant, does not depend on time
- Proportional hazard over time

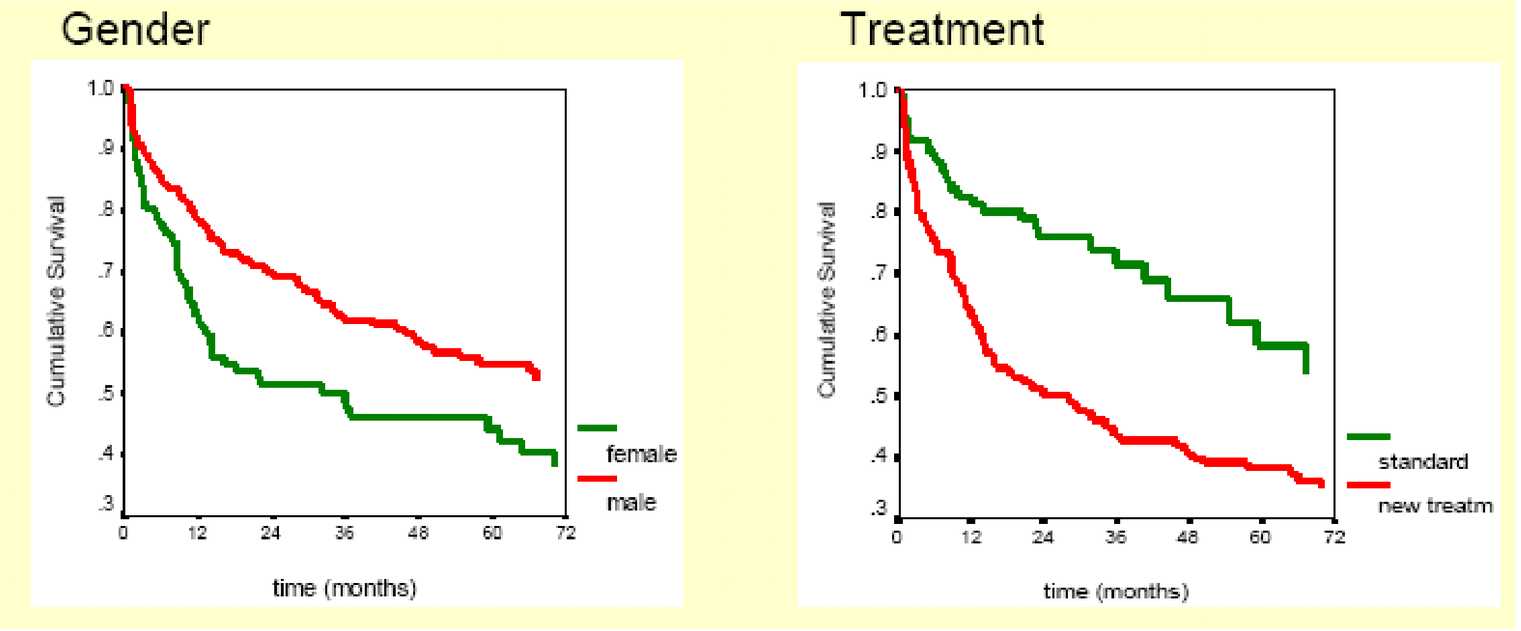
exp(β): indicates how large (small) is the hazard in one group with respect to the hazard in the reference group

Cox Regression Survival Model – Example

- HIV-patients receiving 2 types of treatment.
- The objective is to investigate the survival probability of the patients, by gender and by treatment.
  - Event of interest: death
  - Covariates: gender (male, female), treatment (new, standard)

- They are followed up to 6 years
  - Duration of the study: 6 years (72 months)
  - Time scale: months
- Consider right censored observations

Cox-Regression model: exploring covariates



- ✓ Curves do not cross each other
- ✓ Proportional hazards…

} Good candidates for covariates

Cox-Regression model: exploring covariates

- $h(t,X) = h_0 \exp(\beta X)$

  $= h_0(t) \exp(\beta_1 \text{gender} + \beta_2 \text{Treatment})$
- Reference group: Female, StandardTreatment

- Fitted model:

  $h(t, X) = h_0(t) \exp(-0.51 * \text{Male} + 0.69 * \text{NewTreatment})$

  $\exp(-0.51) = 0.6$ $\exp(0.69) = 2.0$

Males have larger probabilities of survival than females

Patients receiving new treatment having lower survival probabilities than patients with standard treatment

Cox-Regression: Example

Figure 3. Hazard Ratios for the Primary End Point in Prespecified Subgroups of Interest at 12 Weeks After Randomization

Subgroup	Total No.		Events, No.		12-wk Kaplan-Meier Rate, %		Hazard Ratio (95% CI)
	Losmapimod	Placebo	Losmapimod	Placebo	Losmapimod	Placebo	
Qualifying diagnosis							
Non-ST-elevation MI	1299	1325	111	90	8.6	6.8	1.27 (0.96-1.68)
ST-elevation MI	432	433	28	33	6.5	7.6	0.84 (0.51-1.40)
Chronic kidney disease							
Yes	121	123	25	15	20.7	12.2	1.73 (0.91-3.28)
No	1609	1635	113	108	7.0	6.6	1.07 (0.82-1.39)
Prior MI							
Yes	425	426	48	40	11.4	9.4	1.22 (0.80-1.85)
No	1306	1332	91	83	7.0	6.2	1.12 (0.84-1.51)
Prior PCI or CABG surgery							
Yes	466	464	50	42	10.8	9.1	1.21 (0.80-1.82)
No	1264	1294	88	81	7.0	6.3	1.12 (0.83-1.51)
Prior heart failure							
Yes	206	216	28	21	13.6	9.8	1.43 (0.81-2.52)
No	1524	1542	110	102	7.2	6.6	1.10 (0.84-1.44)
Planned treatment strategy							
Invasive	1575	1603	115	105	7.3	6.6	1.12 (0.86-1.46)
Conservative	156	155	24	18	15.4	11.6	1.34 (0.73-2.48)
No. of predictors of cardiovascular risk							
0	40	40	3	2	7.5	5.0	1.51 (0.25-9.02)
1	912	900	59	46	6.5	5.1	1.27 (0.87-1.87)
2	466	518	36	39	7.8	7.6	1.03 (0.66-1.62)
≥3	313	300	41	36	13.2	12.1	1.11 (0.71-1.73)
Overall	1731	1758	139	123	8.1	7.0	1.16 (0.91-1.47)

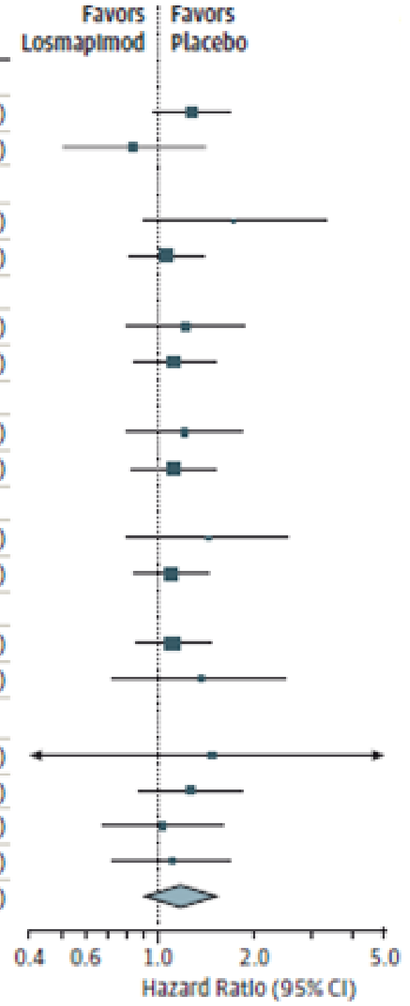


Figure 3 (truncated) - JAMA. 2016;315(15):1591-71599

Competing Risks

"Competing risks are said to be present when a patient is at risk of more than one mutually exclusive event, such as death from different causes, and the occurrence of one of these will prevent any other event from ever happening."

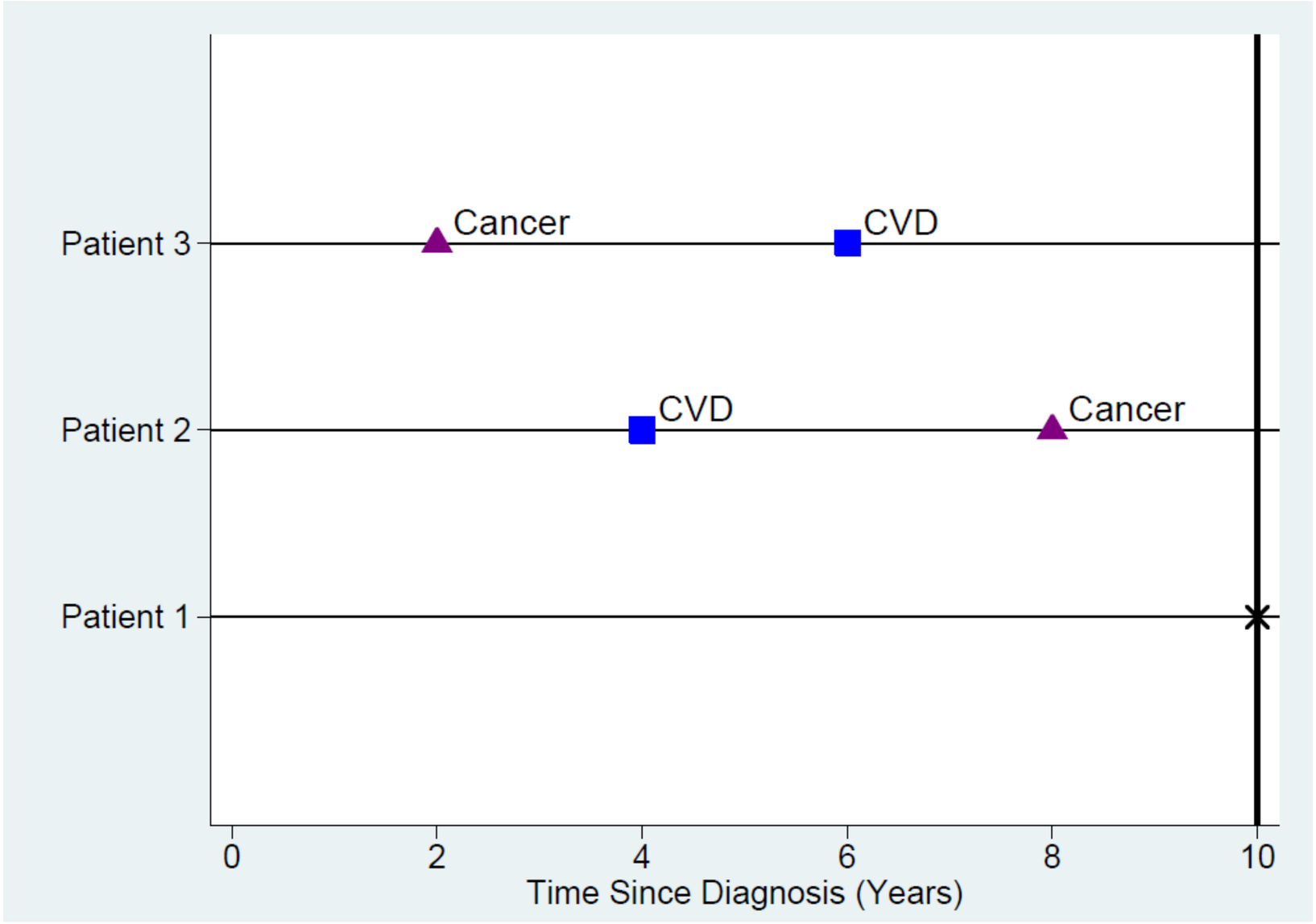
Gichangi & Vach (2005)

Competing Risks

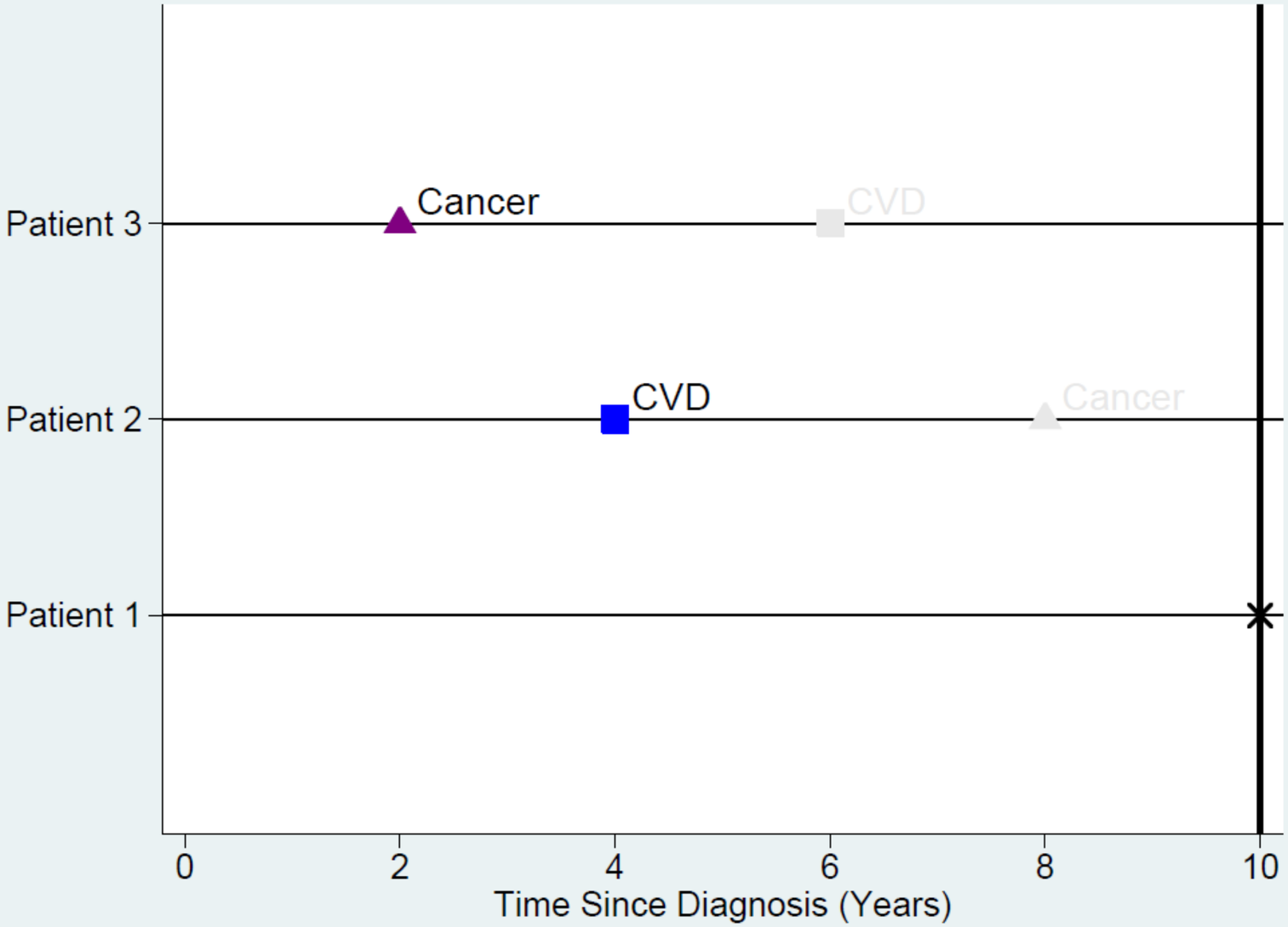
Examples:

- **Event of interest: Death.**
- **Cause of death was categorized into the following:**
  - Breast cancer
  - Heart disease (CVD)
  - Other causes

Competing Risks – Example



Competing Risks – Example



Competing Risks – Data format

Data

Patient ID	T (treatment)	Y1 (Cancer death)	Y2 (CVD death)	Time (years)
1	B	No	No	10
2	A	No	Yes	4
3	A	Yes	No	2

Competing Risks – Data format

Old method

PID	T	Y1	Y2	Time
1	1	0	0	10
2	0	0	1	4
3	0	1	0	2

For outcome (Y1, Y2): 0 = censored, 1 = event

Competing Risks – Data format

Competing Risk method

PID	T	Y	Time
1	1	0	10
2	0	2	4
3	0	1	2

For outcome (Y):

0 = censored

1 = event 1 (death from cancer, our primary event)

2 = event 2 (death from CVD, competing event)

Competing Risks – Key Concepts

- Cumulative incidence function (CIF)
  - The cumulative incidence function gives the proportion of patients at time *t* who have died from cause *k* accounting for the fact that patients can die from other causes.
- Cause-specific hazard (won't cover today)
  - The cause-specific hazard, $h_k(t)$, is the instantaneous risk of dying from a particular cause *k* given that the subject is still alive at time *t*.
- Subdistribution hazard (won't cover today)
  - The subdistribution hazard, $h_{ks}(t)$, is the instantaneous risk of dying from a particular cause *k* given that the subject has not died from cause *k*.

Cumulative Incidence Function (CIF)

The cumulative incidence function gives the proportion of patients at time *t* who have died from cause *k* accounting for the fact that patients can die from other causes.

Define:

- S_t = Number at risk at the end of period *t*
- E_t = Number of primary events in period *t*
- A_t = Number of competing events in period *t*

$$P(E = t \mid E \geq t) \approx \frac{E_t}{E_t + A_t + S_t}$$

Competing Risks

Cumulative Incidence Function (CIF)

S_t = Number at risk at the end of period t

E_t = Number of primary events in period t

A_t = Number of competing events in period t

Note:

$$P(E = t \mid E \geq t) \approx \frac{E_t}{E_t + A_t + S_t}$$

→ **Kaplan-Meier estimator does NOT work !**

$$P(E \geq t+1 \mid E \geq t) \neq 1 - \frac{E_t}{E_t + A_t + S_t}$$

Competing Risks

Cumulative Incidence Function (CIF)

Define the survival function as before (using **Kaplan-Meier**)

$$\hat{S}(t) = \prod_{j=1}^{t} \left(1 - \frac{A_j + E_j}{E_j + A_j + S_j}\right)$$

Competing risk method

$$\hat{S}_0(t) = \prod_{j=1}^{t} \left(1 - \frac{E_j}{E_j + S_j}\right)$$

Old method

Define the **CIF** (of the primary events) as

$$\hat{C}_E(t) = \sum_{j=1}^{t} \frac{E_j}{E_j + A_j + S_j} \hat{S}(j-1)$$

Competing risk method

$$\hat{C}(t) = \sum_{j=1}^{t} \frac{E_j}{E_j + S_j} \hat{S}_0(j-1)$$

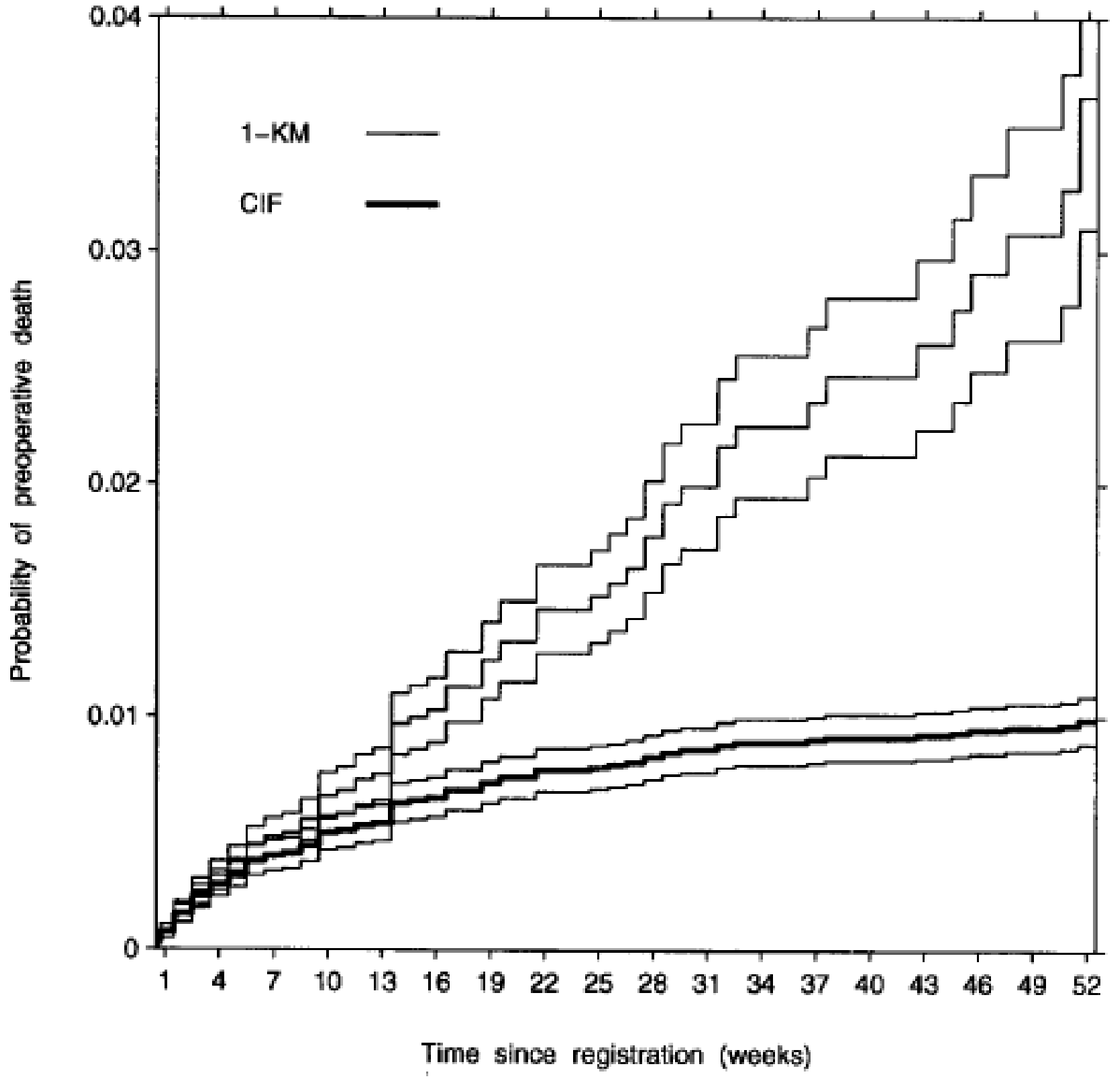
Old method

Competing Risks

CIF compared to **Kaplan-Meier**

$\hat{C}(t) > \hat{C}_E(t)$

So, Kaplan-Meier (old method) over-estimates the hazard rate



Competing Risks

- We won’t cover more statistical analysis models in the competing risks today, such as Cause-specific hazard model and Subdistribution hazard model.

- Keep in mind: When your study subjects are at risk of more than one mutually exclusive event such as death from different causes, Kaplan-Meier estimator may NOT be appropriate. Competing risk methods may be considered.

- Come to BCC for consulting !

Summary

- A common circumstance in working with survival data is that not all the individuals in a sample are observed until their respective times of “failure”. The incomplete observation of a times to “failure” is known as censoring;
- Kaplan-Meier method is a nonparametric technique that uses the exact survival time for each individual in a sample instead of group the times into intervals
- A nonparametric technique known as the logrank test is used to determine whether survival differs systematically between the groups.
- Cox proportional hazard regression model is a semi-parametric method to study the effect of different covariates on a time-to-event endpoint after adjusting for each other.
- Competing risks occur when a patient is at risk of more than one mutually exclusive event such as death from different causes. Kaplan-Meier estimator may NOT be appropriate. Competing risk methods may be considered.

References

TEXT BOOKS

1.João Moreira, Andre Carvalho, Tomás Horvath – "A General Introduction to Data Analytics" – Wiley - 2018

2.An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics. W. N. Venables, D.M. Smith and the R Development Core Team. Version 3.0.1 (2013-05-16). URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

References:

1**. Dean J, —***Big Data, Data Mining and Machine learning***,** Wiley publications, 2014.

2. Provost F and Fawcett T, —*Data Science for Business*, O'Reilly Media Inc, 2013.

3. Janert PK, —*Data Analysis with Open Source Tools*, O'Reilly Media Inc, 2011.

4. Weiss SM, Indurkhya N and Zhang T, —*Fundamentals of Predictive Text Mining,* Springer-Verlag London Limited, 2010.

5.**Marz N and Warren J**,- Big Data, Manning Publications,2015

Thank You