An introduction to Survival Analysis

Noushin Fahimfar, MD, MPH, PhD candidate in Epidemiology
Mohammad Ali Mansournia, MD, PhD in Epidemiology, A. Professor
Maryam Sharafkhah, MSc, PhD candidate in Biostatistics

- Most epidemiological cohort studies look at the incidence of relatively rare events, where incidence rate could be assumed not to vary or to vary only gradually with time (or age)

- These methods are less appropriate where:

  - There is an interest in quantifying the time elapsed from entry to the event. For example:

    Time to remission of disease following treatment (median time to remission)

  - Incidence rates vary rapidly over time. For example:

    Time to death following surgery for cancer

What is Survival Analysis?

- Survival analysis is a collection of statistical analysis techniques where the outcome is **time** to an **event** occurs.

- ***Survival time***: Is the time from a predetermined start point until the occurrence of the event of interest (end point event).

Initial event:

From the onset of disease or the start of treatment (in RCT)

Entry to the study

From fixed age point

From a particular date (first exposed to a carcinogen)

Censored observation

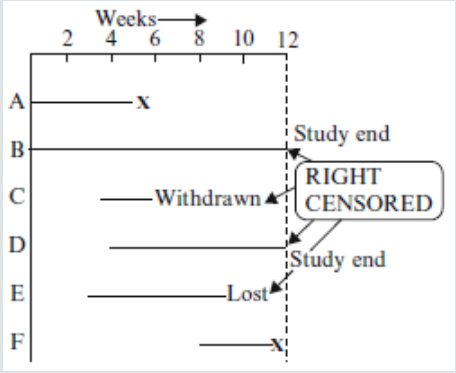
- Time to event in some cases may be censored.
- That is for some subjects the follow-up may not be complete and the event is not observed to happen.
  - Study ends
  - Loss to follow-up
  - Death from other causes

In censored data, the actual survival time is not known.

Censoring

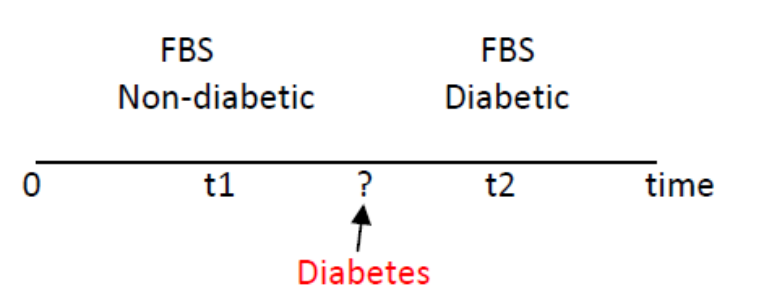
- **Right- censored**

True survival time is equal to or greater than observed survival time



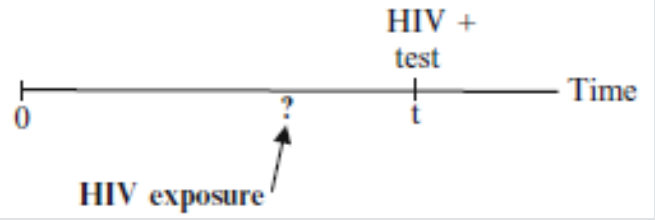
- **Interval censored**

True survival is within a known time interval

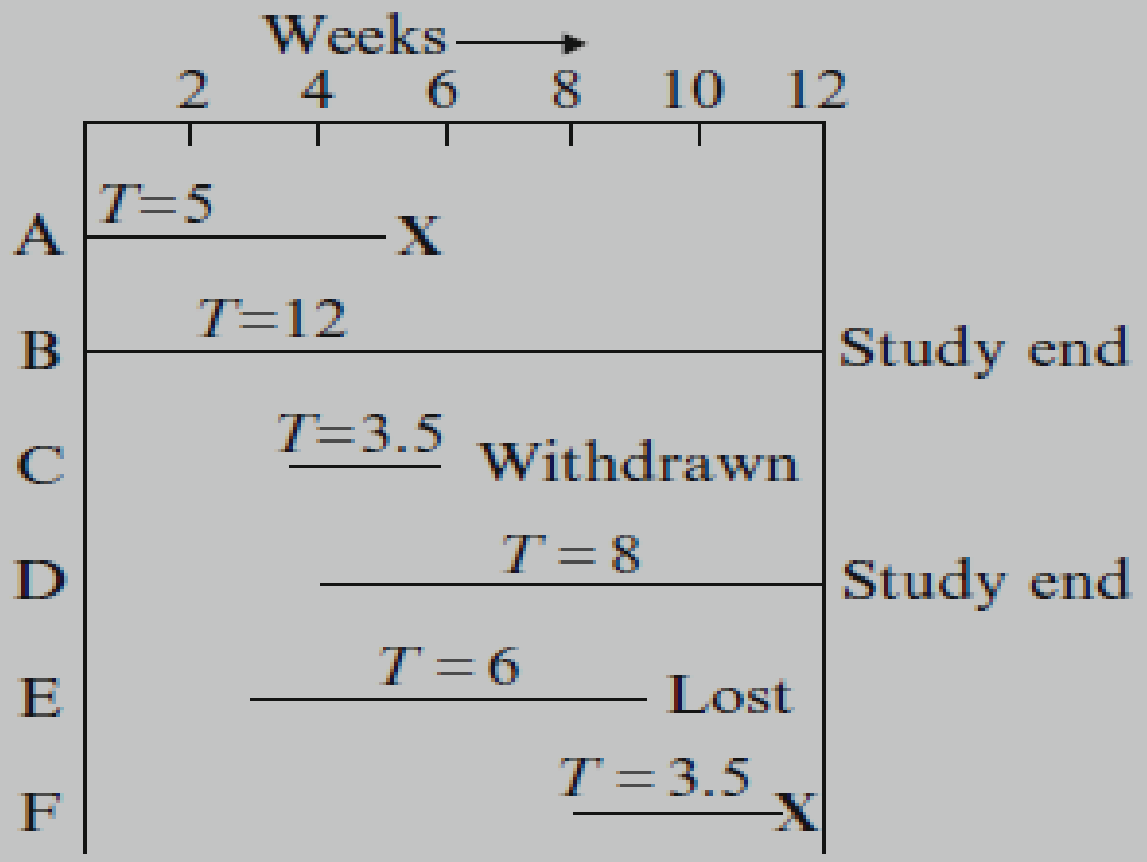


- **Left- censored**

True survival time is less than or equal to the observed survival time



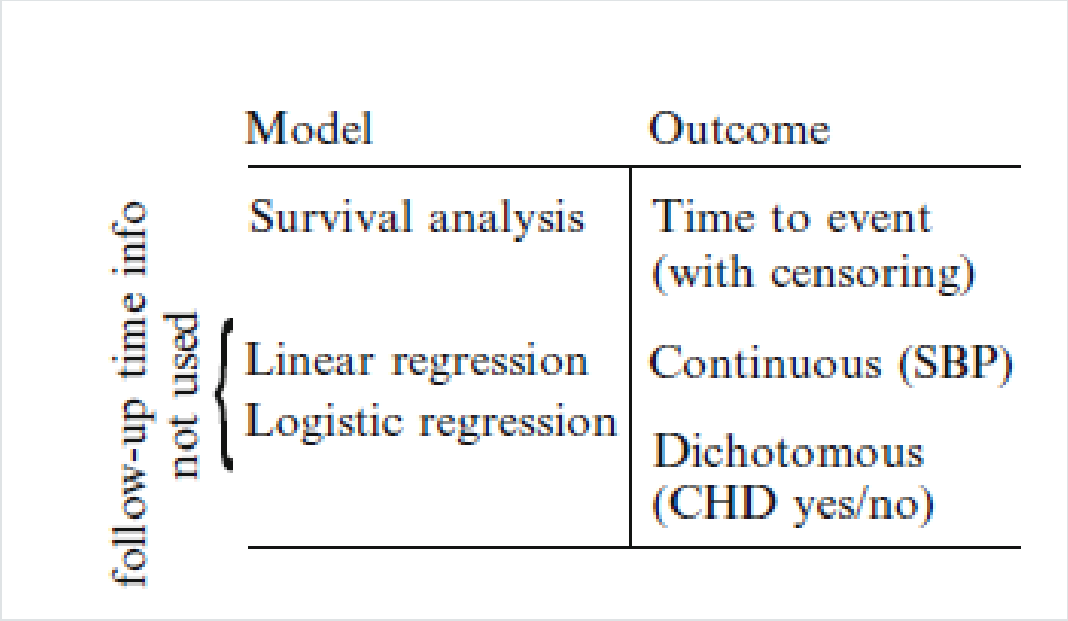
EXAMPLE



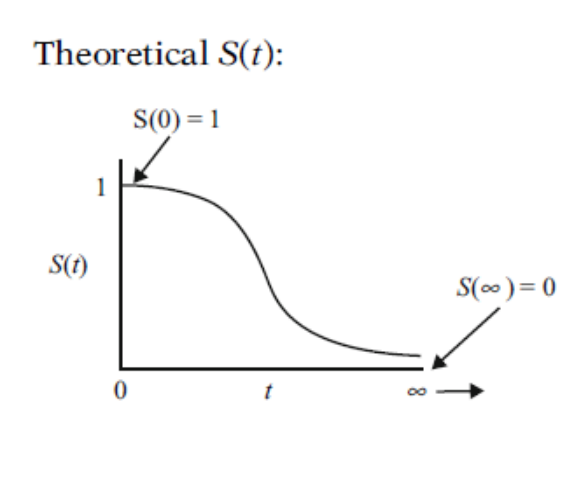
X ⟹ Event occurs

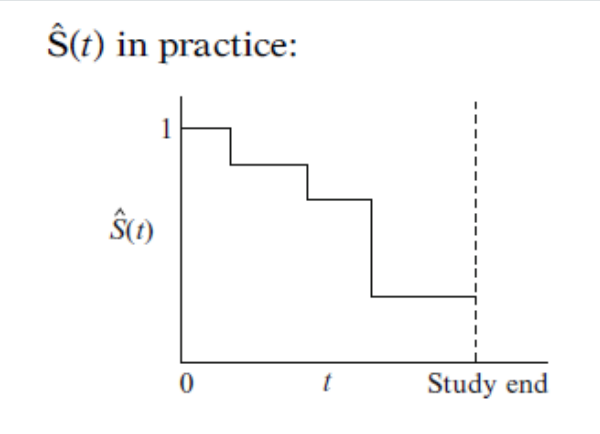
Why survival analysis ?

- Censoring (time of event not observed)
- Unequal follow-up time



- The survival function at time t, denoted S(t), is the probability of being event free at time t; equivalently, the probability that the survival time is greater than t.





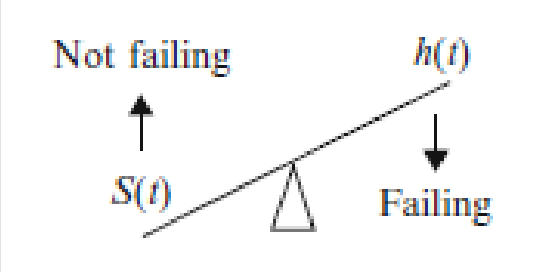
They are non-increasing ; that is ,they head downward as *t* increases;
At time t=0, $S(t) = S(0) = 1;$

at the start of the study, since no one has gotten the event yet,
the probability of surviving past time 0 is one;
At time $t = \infty, S(t) = S(\infty) = 0;$

theoretically, if the study period increased without limit,
eventually nobody would survive, so the survivor curve must eventually fall to zero.

The hazard function , denoted by h(t) , gives the **instantaneous potential** per unit time for the event to occur given that the individual has survived up to time *t.*

In contrast to the survivor function, which focuses on not failing, the hazard function focuses on failing, that is, on the event occurring.



$h(t)$ is a rate: 0 to ∞

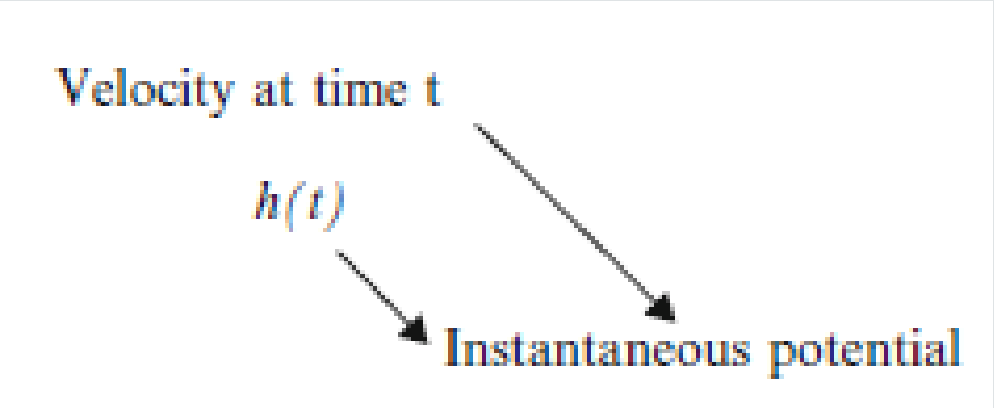
What we mean by instantaneous potential ?

- Consider the concept of velocity.

- *You are driving with a speed of 60 mph , what does this figure mean?*
  - If you continue to drive this way in the next hour, with the speedometer exactly on 60, you would cover 60 miles. This reading gives the potential, at the moment you have looked at your speedometer, for how many miles you will travel in the next hour.

  - Because you may slow down or speed up or even stop during the next hour, the 60-mph speedometer reading does not tell you the number of miles you really will cover in the next hour.
  - *The instrument gives your instantaneous potential or velocity*.

- Similar to the idea of velocity, a hazard function h(t) gives the instantaneous potential at time t for getting an event, like death or some disease of interest, given survival up to time t.



Types of survival analysis

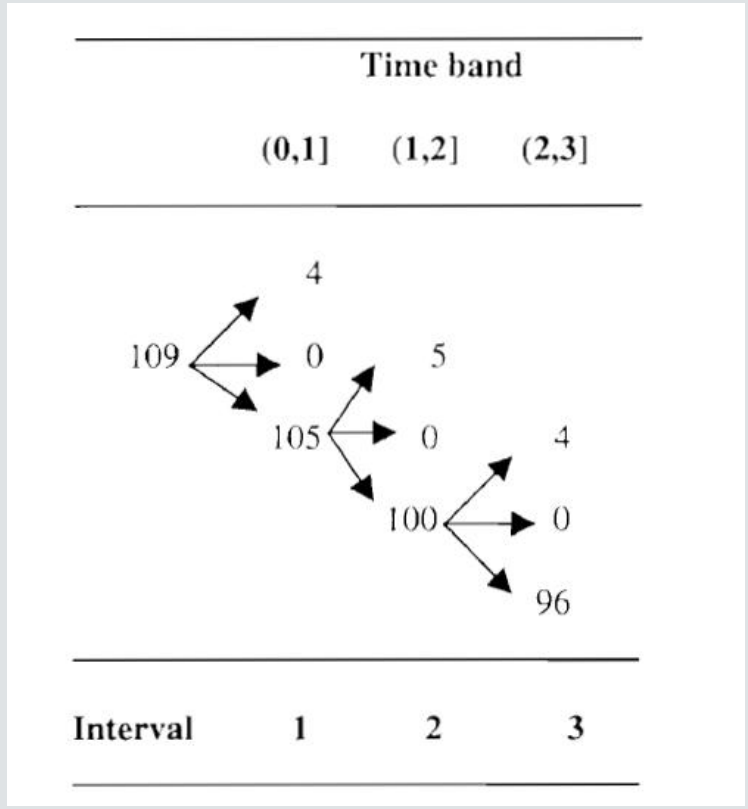
1. Non-parametric method
2. Semi-parametric method
3. Parametric method

Kaplan-Meier Method

- One of the main objectives in survival analysis is to obtain an estimate of the survival experience of the population .
- In most cohort study we have precise information on the time of death or censoring of every individual. It would seem preferable to use the whole of this information when calculating survival curves.
- KM is a kind of non-parametric methods to estimate and graph survival curves in the presence of censored cases.
- It shows probabilities of survival, given survival time and failure status information on a sample of subjects.

Kaplan Meier method:

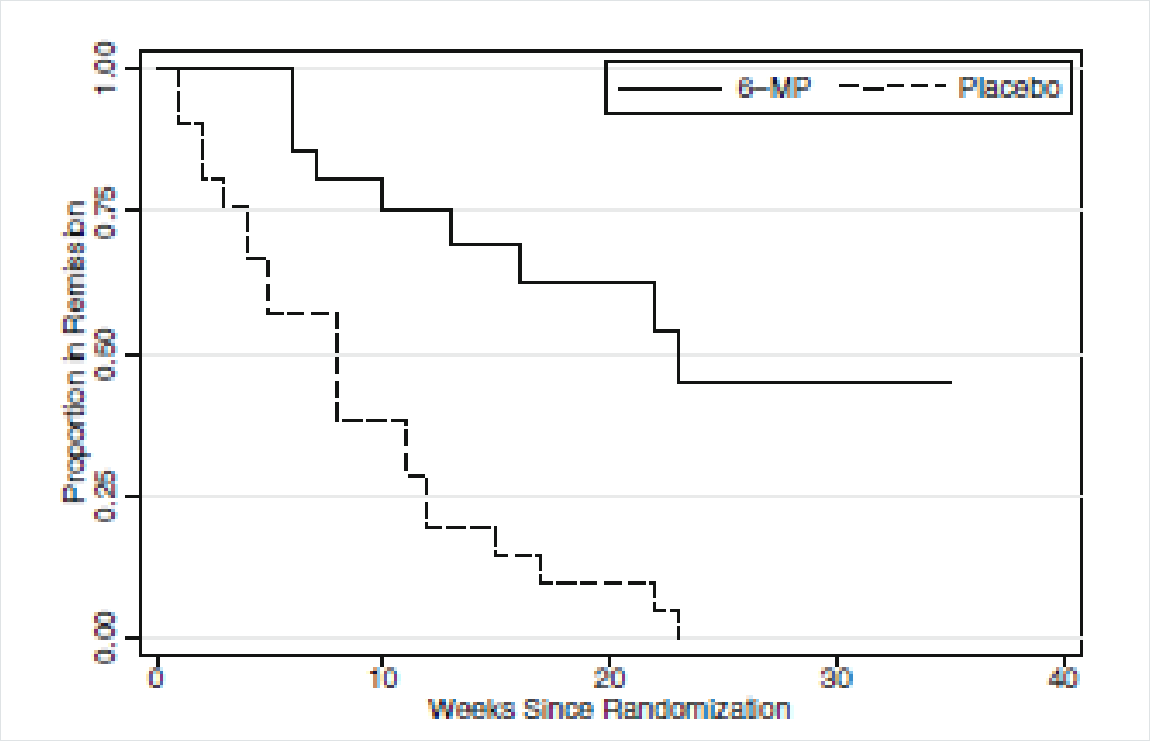
Time Band (yrs)	Deaths from any cause	Censored observations
(0,1]	4	0
(1,2]	5	0
(2,3]	4	0
(3,4]	5	0
(4,5]	7	1
(5,6]	2	4
(6,7]	2	5
(7,8]	7	12
(8,9]	3	27
(9,+)	1	20
All	40	69



Interval i	Time band	Prob(death|interval)	Prob(surviving| interval)
1	(0,1]	4/109= 0.037	0.963
2	(1,2]	5/105= 0.048	0.952
3	(2,3]	4/100= 0.040	0.960

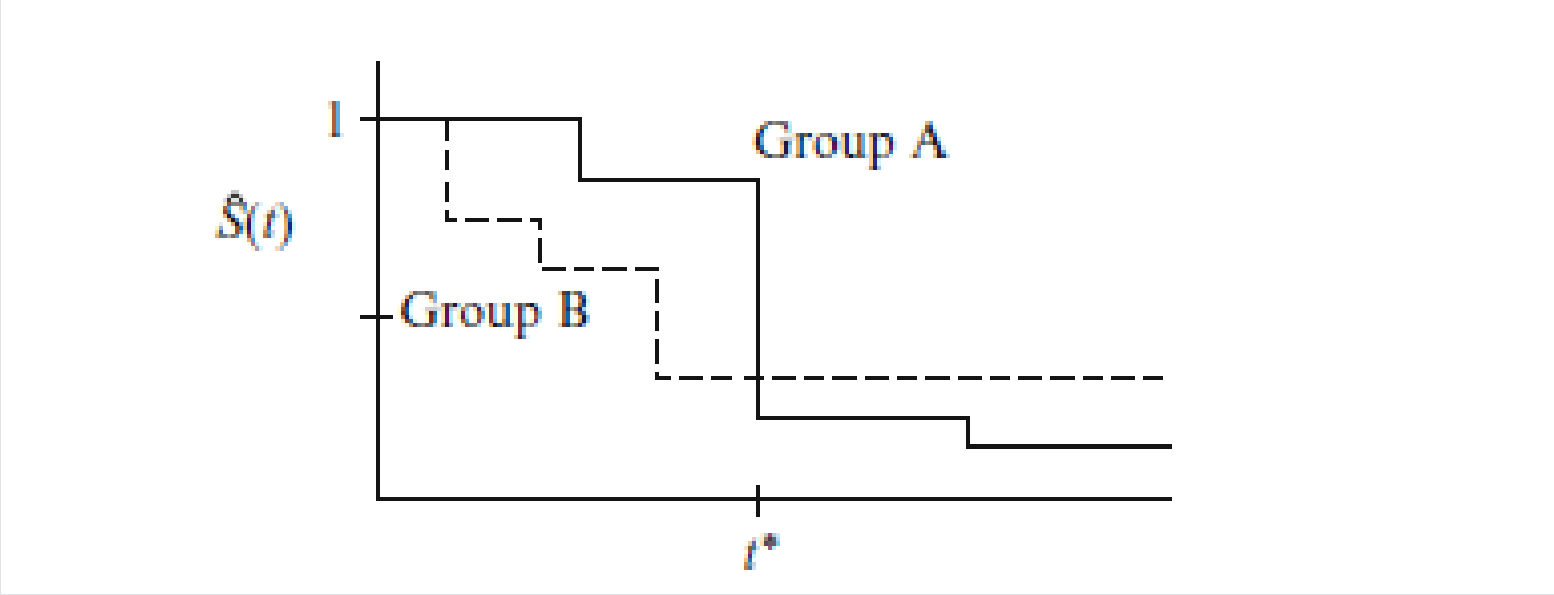
$$S(3) = 0.963 \times 0.952 \times 0.960 = 0.88$$

Interpretation of Kaplan-Meier curve: an example



Survival curves by treatment for leukemia patients

Example:



- Which group has a better survival prognosis **before** time t ?
- Which group has a better survival prognosis **after** time t ?
- Which group has a longer median survival time?

Summary statistics

1. Median survival: time when S(t) = 0.5

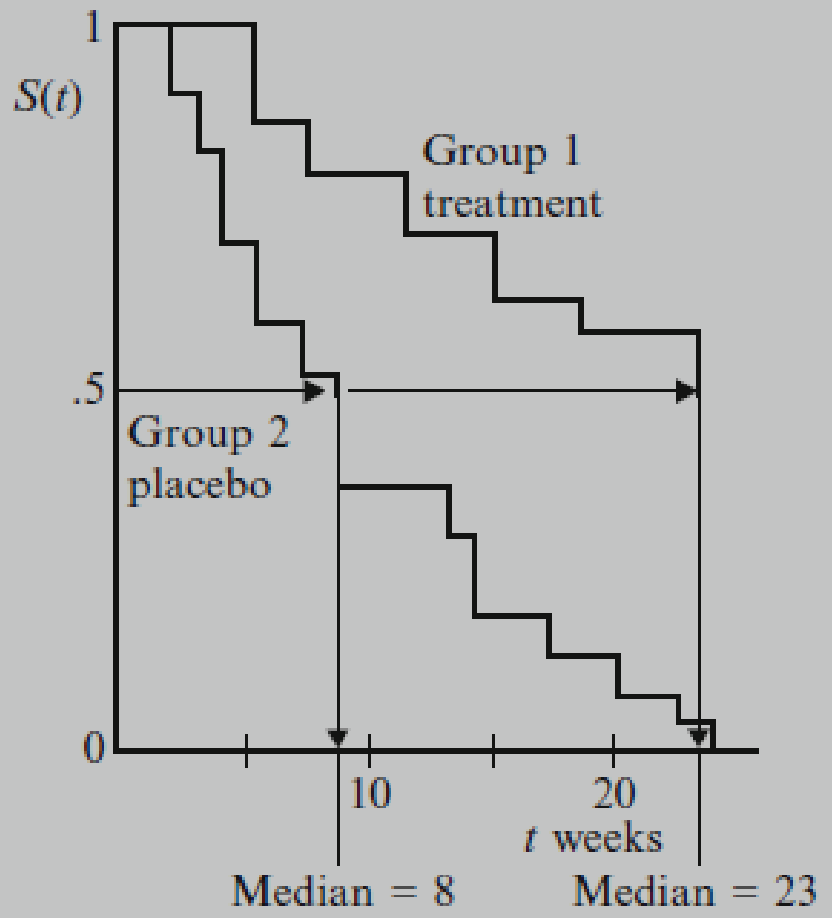
   * Must have enough data

2. Mean survival: area under the survival curve

3. Survival rate according to time to event: 5-year survival, 10-year survival***

Median (treatment) = 23 weeks
Median (placebo) = 8 weeks

EXAMPLE



Log-Rank test

- To test whether there is any real difference between the survival experience of two or more than two groups.

- A type of chi-square test

Proportional Hazards assumption

- The hazard for any individual is a fixed proportion of the hazard for any other individual
- Under the proportional hazards assumption, the hazard ratio does not vary with time. That is, HR(t)=HR
- This assumption is useful but not necessary for Cox proportional hazards models

Cox Proportional Hazards Model

- A semi-parametric model because even if the regression parameters (the betas) are known, the distribution of the outcome remains unknown. The baseline survival (or hazard) function is not specified in a Cox model
- A flexible tool for assessing the relationship of multiple predictors to a right-censored, time to event outcome.

  - Add covariates to the model
  - No need to stratify
  - Change in a prognostic factor → proportional change in the hazard (on the log scale)
  - Can test the effect of the prognostic factor as in linear regression - H0: $\beta=0$

Extended Cox models could be used when the hazard ratio for a predictor change with time.

- Including an interaction between that predictor and time
- Stratified Cox model

Time dependent models

- Time-dependent covariates (TDCs)
  - ✓ A time-dependent covariate in a Cox model is a predictor whose values may vary with time
  - ✓ More common when we have repeated measures; for example 3 times measurements of FBS

Recurrent events model

Event occurs more than once per subject over follow-up time.

- Two different approaches:

1. Recurrent events are treated as identical
   Counting Process (CP) approach

2. Recurrent events are not treated as identical: event order important or different disease categories
   Stratified Cox (SC) model approach

Parametric Survival Models:

- A parametric survival model is a model in which survival time (the outcome) is assumed to follow a known distribution. Many parametric models are acceleration failure time models in which survival time is modeled as a function of predictor variables

- Distributions commonly used for parametric survival models:
    - Weibull
    - Exponential
    - Log-logistic
    - Lognormal
    - Generalized gamma

When we use parametric models?

- These models are preferable for predictions when we know the exact distribution of time

Competing risks Survival analysis

- 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.

1. A person can die from lung cancer or from a stroke, but not from both (although he can have both lung cancer and atherosclerosis
   before he dies);
2. Patients with advanced-stage cancer may die after surgery before their hospital stay is long enough for them to get a hospital infection;

- There are some models which involve competing risk such as “Fine and Gray model”

Multi state models

- An event may be considered as a transition from one state to another and, therefore, multi-state models will often provide a relevant modeling framework for event history data
- Consider an event of a patient with two states and one possible transition from an “alive” state to a “dead” state
- The “alive” state may be divided into two or more transient states, each of which corresponding to a particular stage of the illness
- **Multi-state models can be used to model the movement of patients among the various states**

.

Example: Illness-death model without recovery