Lecture 8

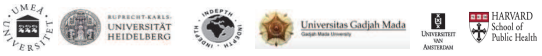
Longitudinal studies and survival analysis

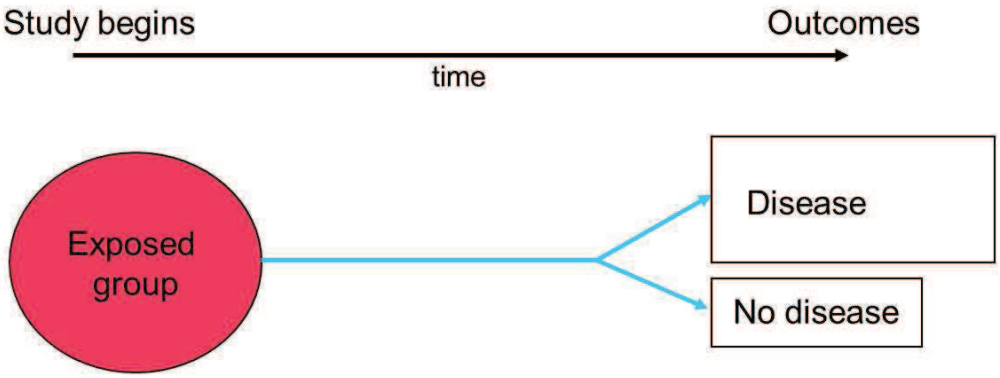
Outline

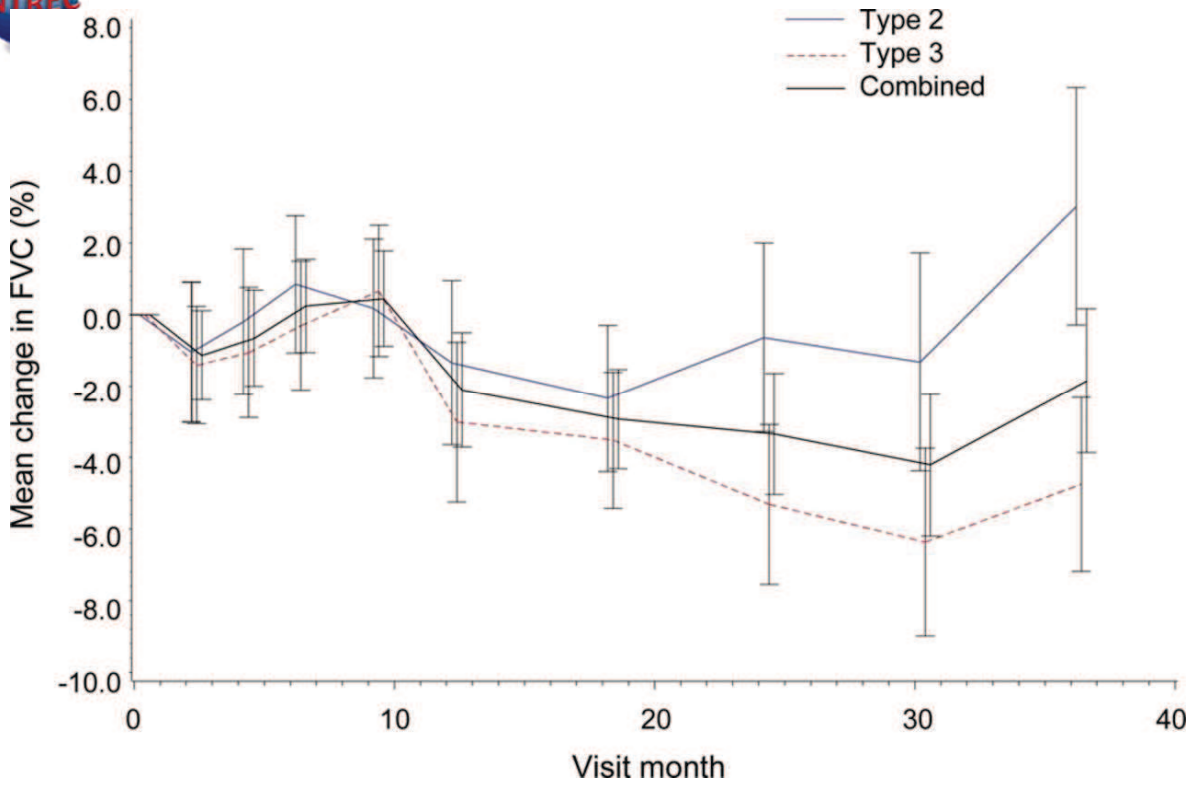
- Benefits and challenges of longitudinal studies
- Objectives of survival analysis
- Censored data
- Survival distributions
- Kaplan-Meier Curves
- Cox regression

What are longitudinal studies?

- Measures collected repeatedly on the same individuals over time.
- In epidemiology, a *cohort* study is a longitudinal study.
- A *cohort* is a set of individuals sharing a common characteristic (e.g. same baseline age) or experience in a particular time period.







Benefits of longitudinal studies

1. Incident events are recorded
2. Prospective collection of exposure data
3. Temporal order of exposures and outcomes is observed
4. Measurement of individual change in outcomes
5. Provides better handling of missing data
6. Provides a better estimate of lifetime prevalence
7. Permits the study of risk and disease progression

Challenges of longitudinal studies

1. Participant follow-up (missing data)
2. Analysis of correlated data
3. Time-varying covariates

Clinical longitudinal studies

Diagnosis

- *What is the chance that my patient will develop this disease?*

Prognosis

- *How long will it take for this patient to recover?*

Response to treatment

- *Which treatment is best for this patient?*

Objectives of survival analysis

- **Estimate time-to-event for a group of individuals**, such as time until second heart-attack for a group of patients.
- **To compare time-to-event between two or more groups**, such as treated vs. placebo patients in a randomized controlled trial.
- **To assess the relationship of co-variables to time-to-event**, such as: does weight, insulin resistance, or cholesterol influence survival time of heart-attack patients?

Why use survival analysis?

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

  -- ignores censoring

- Why not compare proportion of events in your groups using risk/odds ratios or logistic regression?

  -- ignores time

Survival Analysis: Terms

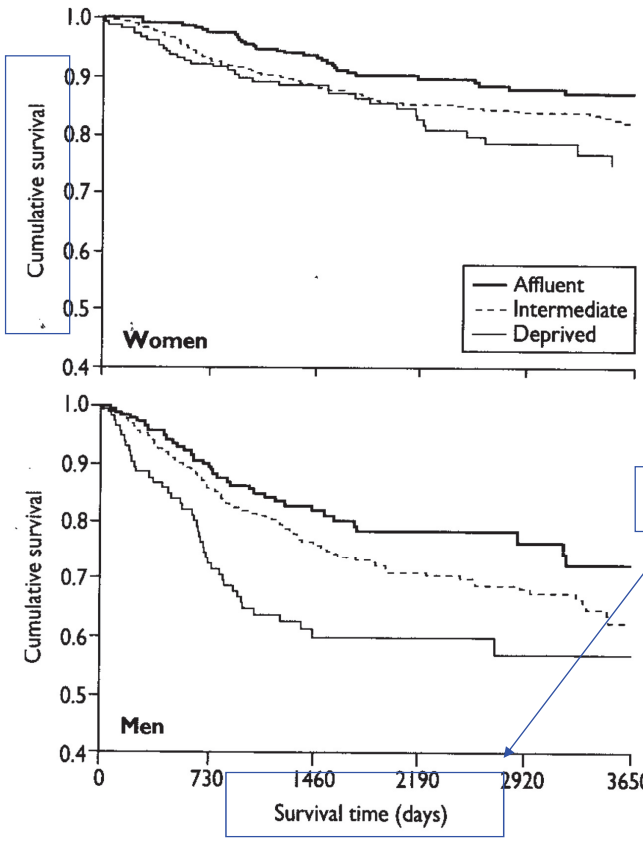
- Time-to-event: The time from entry into a study until a subject has a particular outcome

- Censoring: Subjects are said to be censored if they are lost to follow up or drop out of the study, or if the study ends before they die or have an outcome of interest. They are counted as alive or disease-free for the time they were enrolled in the study.
  - If dropout is related to both outcome and treatment, dropouts may bias the results

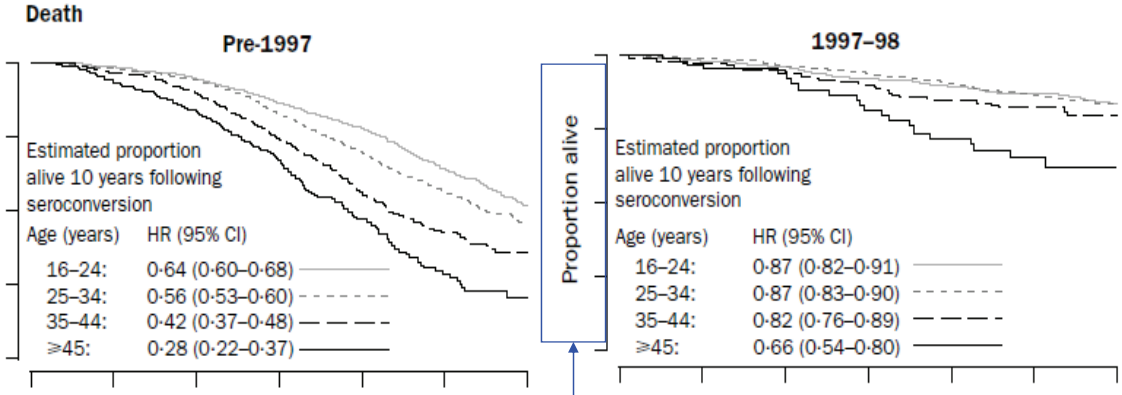
Regression vs. Survival Analysis

Technique	Output
Linear regression	Linear changes
Logistic regression	Odds ratios
Survival analysis	Hazard ratios

Incidence and thickness of primary tumours and survival of patients with cutaneous malignant melanoma in relation to socioeconomic status. MacKie et al. B*MJ* 1996; 312: 1125

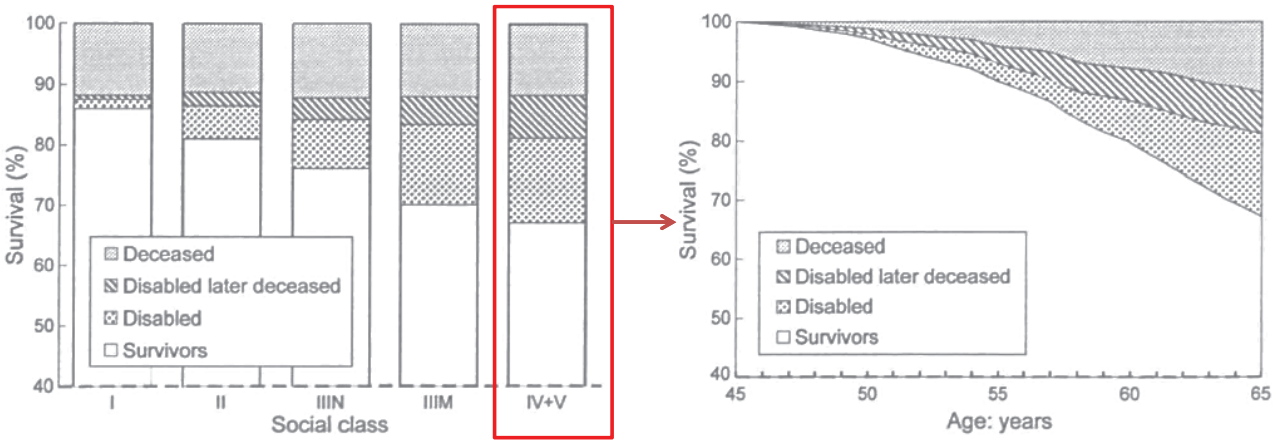


Determinants of survival following HIV-1 seroconversion after the introduction of HAART. CASCADE Collaboration Lancet 2003; 362:1267-74



Proportion alive = cumulative survival

Permanent work incapacity, mortality and survival without work incapacity among occupations and social classes: a cohort study of ageing men in Geneva. Guberan et al (1998) International Journal of Epidemiology 27:1026-32



Data Structure: survival analysis

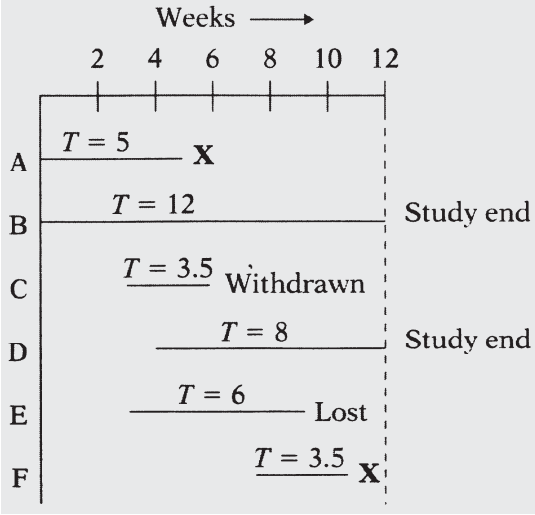
Two-variable outcome :

- Time variable: *t* = time at last disease-free observation or time at event
- Censoring variable: *c* =1 if had the event; *c* =0 no event by time *t*

Censored Data

- Some patients may still be alive or in remission at the end of the study period
- The exact survival times of these subjects are unknown
- These are called *censored observation* or *censored times* and can also occur when individuals are lost to follow-up after a period of study

Types of censoring



- Subject does not experience event of interest
- Incomplete follow-up
  - Lost to follow-up
  - Withdraws from study
  - Dies (if not being studied)
- **Left** or **right** censored

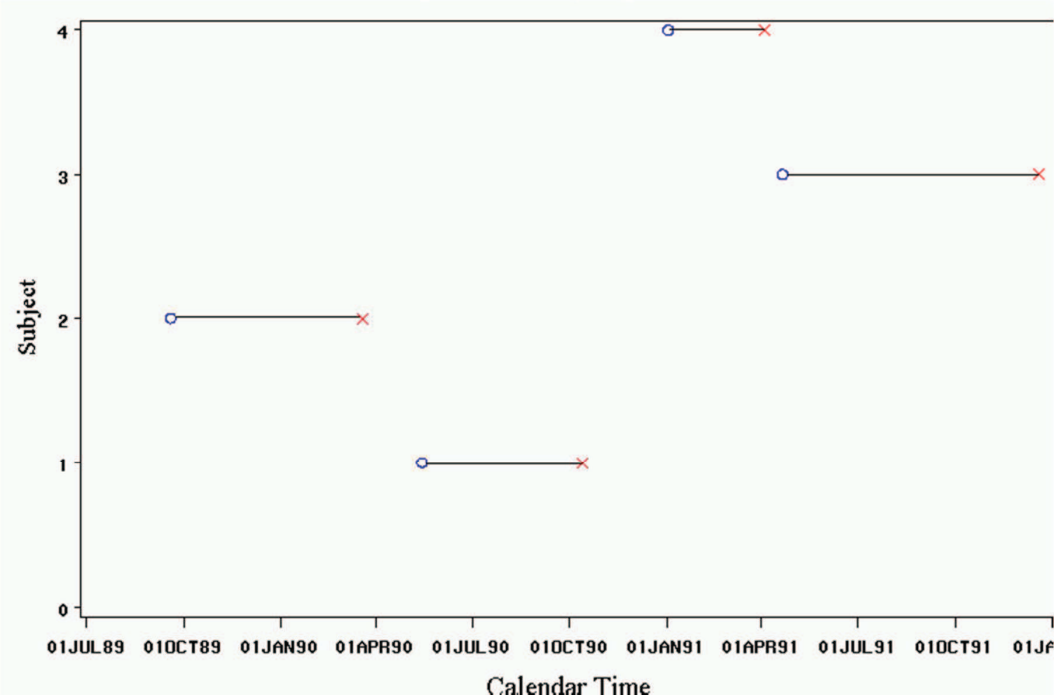
Right Censoring (T>t)

Common examples

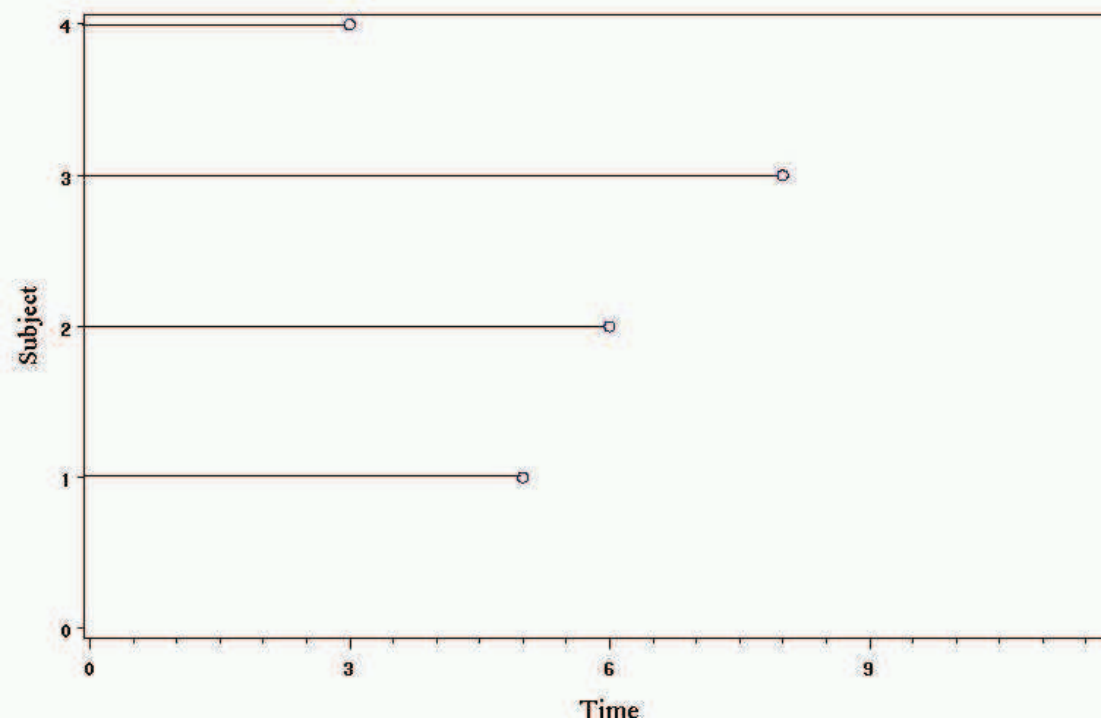
- Termination of the study
- Death due to a cause that is not the event of interest
- Loss to follow-up

We know that subject survived at least to time *t*.

Left censoring



Right censoring



Introduction to survival distributions

- T the event time for an individual, is a random variable having a probability distribution.
- Different models for survival data are distinguished by different choice of distribution for T

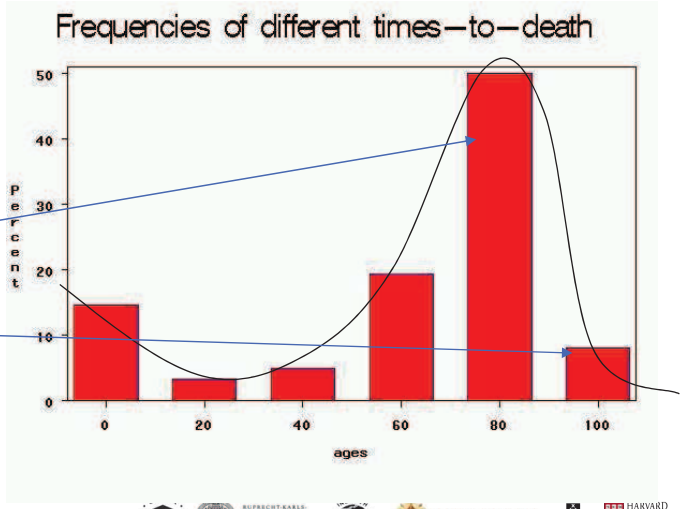
Probability density function: f(t)

In the case of human longevity, T is unlikely to follow a normal distribution, because the probability of death is not highest in the middle ages, but at the beginning and end of life.

Hypothetical data:

People have a high chance of dying in their 70's and 80's;

BUT they have a smaller chance of dying in their 90's and 100's, because few people make it long enough to die at these ages.



Probability density function: f(t)

The probability of the failure time occurring at exactly time t (out of the whole range of possible t's).

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}$$

Survival function: 1-F(t)

The goal of survival analysis is to estimate and compare survival experiences of different groups.

Survival experience is described by the cumulative survival function:

$$S(t) = 1 - P(T \leq t) = 1 - F(t)$$

Example: If t=100 years, S(t=100) = probability of surviving beyond 100 years.

Survival Function or Curve

- Let *T* denote the survival time
- S(t) = P(surviving longer than time t)
  = P(T > t)
- *The function S(t) is also known as the cumulative survival function. $0 \leq S(t) \leq 1$*
- $\hat{S}(t)$=*number of patients surviving longer than t / total number of patients in the study*

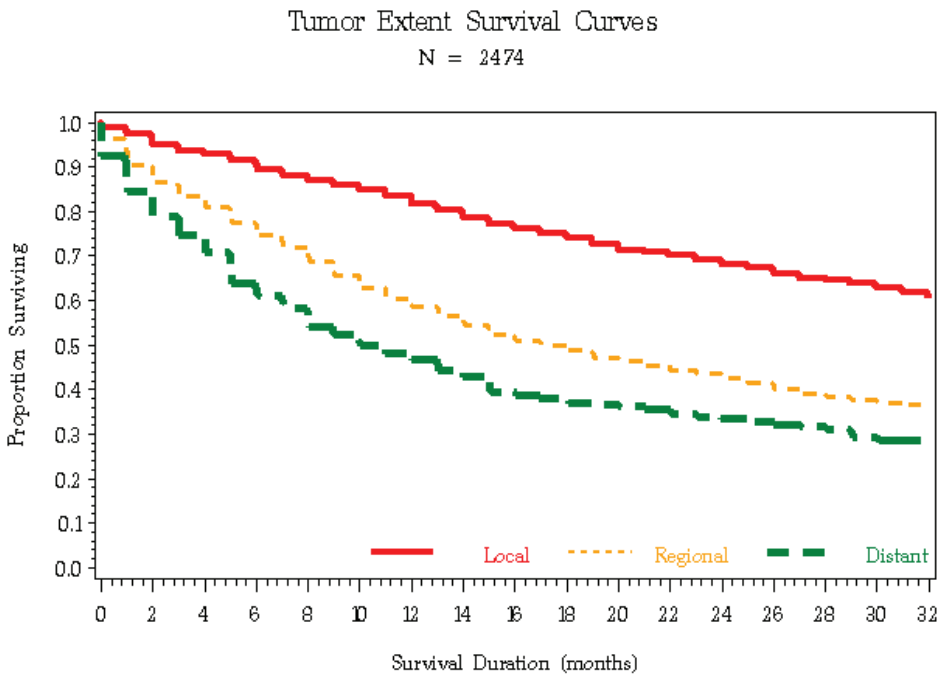
Kaplan-Meier

Non-parametric estimate of the survival function:

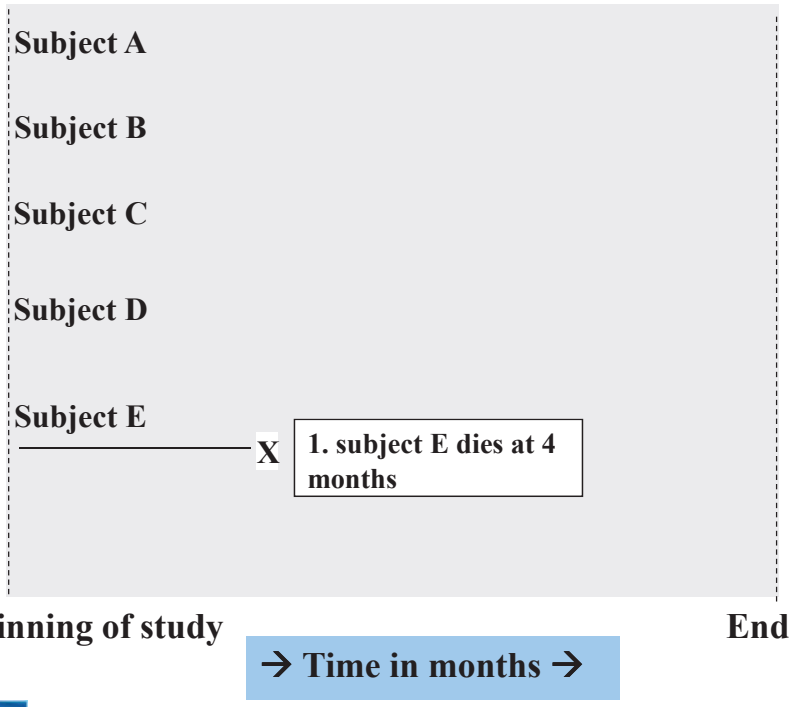
Simply, the empirical probability of surviving past certain times in the sample (taking into account censoring).

Time t_i	# at risk	# events	$\hat{S}$
0	20	0	1.00
5	20	2	[1-(2/20)]*1.00=0.90
6	18	0	[1-(0/18)]*0.90=0.90
10	15	1	[1-(1/15)]*0.90=0.84
13	14	2	(1-(2/14)]*0.84=0.72

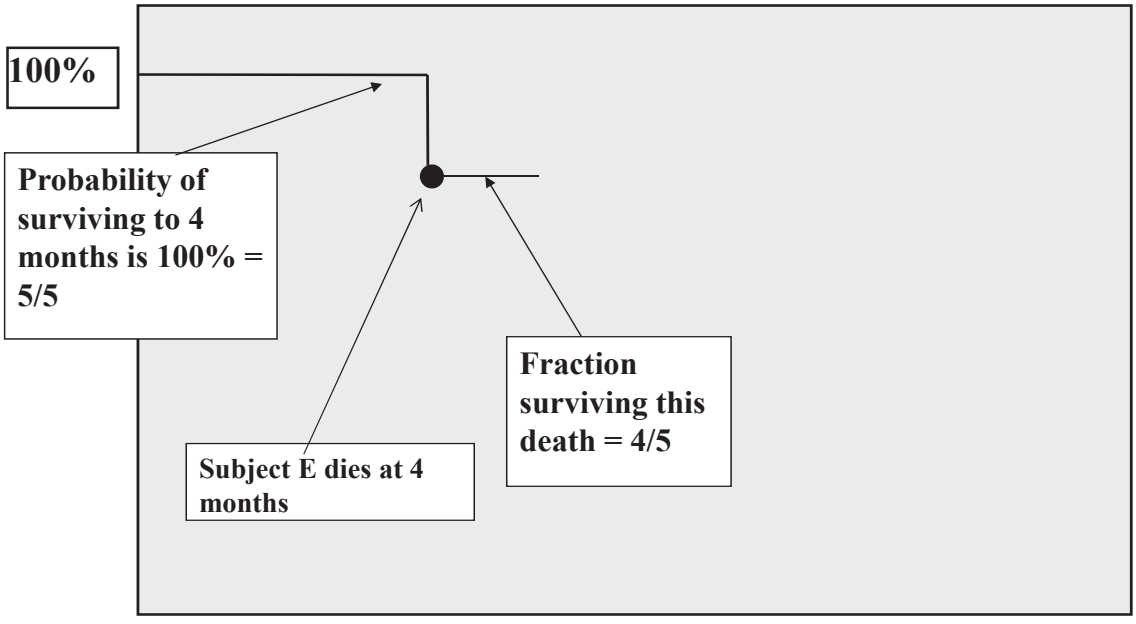
Kaplan-Meier Curve



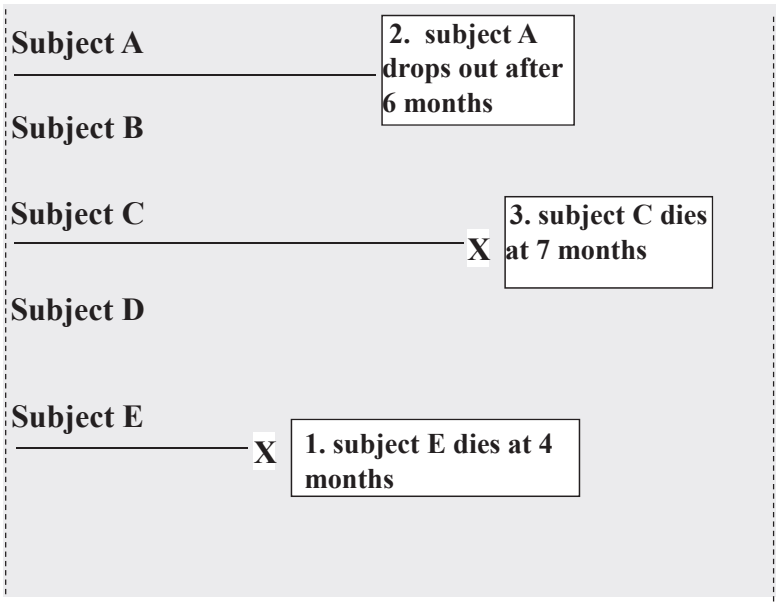
Survival Data (right-censored)



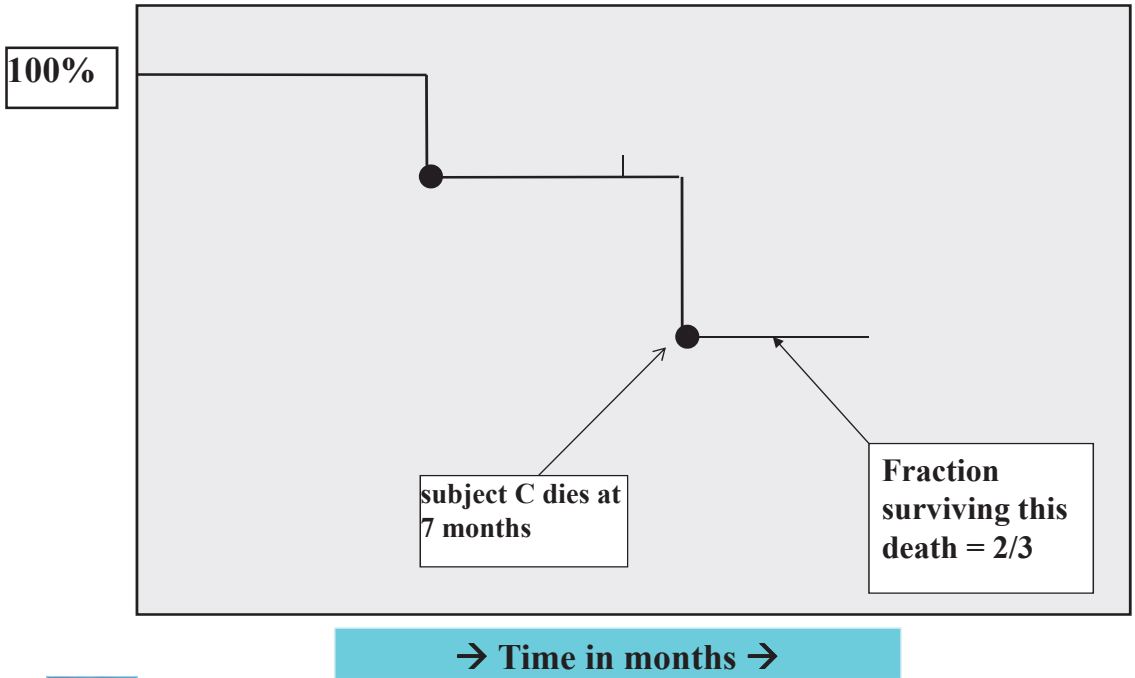
Corresponding Kaplan-Meier Curve



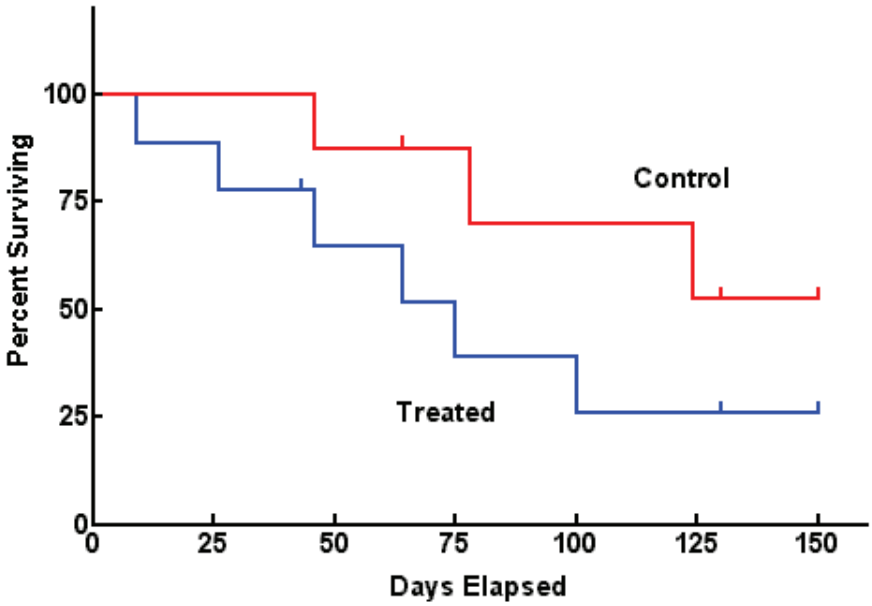
Survival Data



Corresponding Kaplan-Meier Curve



Comparing 2 groups



Use log-rank test to test the null hypothesis of no difference between survival functions of the two groups

Limitations of Kaplan-Meier

- Mainly descriptive
- Doesn't control for covariates
- Requires categorical predictors
- Can't accommodate time-dependent variables

Survivor function and hazard function

- Survivor function, S(t) defines the probability of surviving longer than time *t*
  - this is what the Kaplan-Meier curves show.
  - Hazard function is the derivative of the survivor function over time h(t)=dS(t)/dt
- Survivor and hazard functions can be converted into each other

Hazard Function

- The hazard function *h(t)* of survival time *T* gives the *conditional failure rate*
- The hazard function is also known as the *instantaneous failure rate, force of mortality,* and *age-specific failure rate*
- *The hazard function gives the risk of failure per unit time during the aging process*

Hazard Function

- Event = death, scale = months since treatment
- "h(t) = 1% at t = 12 months"
- "At 1 year, patients are dying at a rate of 1% per month"
- "At 1 year the chance of dying in the following month is 1%"

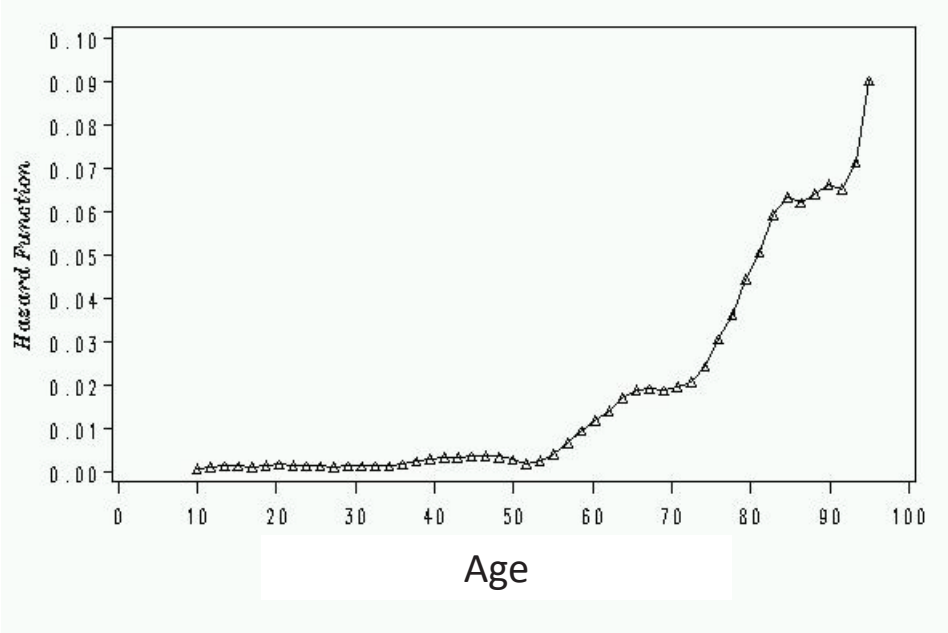
Hazard Function

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \le T < t + \Delta t / T \ge t)}{\Delta t}$$

In words: the probability that ***if you survive to t***, you will succumb to the event in the next instant.

$$\text{Hazard from density and survival}: h(t) = \frac{f(t)}{S(t)}$$

Hazard Function



Hazard rate is an instantaneous incidence rate.

Limit of Kaplan-Meier curves

- What happens when you have several covariates that you believe contribute to survival?
- Example
  - Smoking, hyperlipidemia, diabetes, hypertension, contribute to time to myocardial infarct
- Can use **stratified K-M curves** – for 2 or maybe 3 covariates
- Need another approach – multivariate **Cox proportional hazards model** is most common -- for many covariates
  - (think multivariate regression or logistic regression rather than a Student's t-test or the odds ratio from a 2 x 2 table)

Cox Proportional Hazards Model

- Add covariates to the model
- 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 - H_0: $\beta=0$

Limitations of Cox PH model

- Does not accommodate variables that change over time
  - Most variables (e.g. gender, ethnicity, or congenital condition) are constant
    - If necessary, one can program time-dependent variables
    - When might you want this?
- Baseline hazard function, $h_o(t)$, is never specified
  - You can estimate $h_o(t)$ accurately if you need to estimate S(t).

Summary

- Survival analyses quantifies **time to a single, dichotomous event**
- Handles **censored data** well
- Survival and hazard can be mathematically converted to each other
- **Kaplan-Meier survival curves** can be compared statistically and graphically
- **Cox proportional hazards models** help distinguish individual contributions of covariates on survival, provided certain assumptions are met.

SAGE

- Currently only cross-sectional data available
- Future releases of data will allow longitudinal and survival analysis
- Possibility for some important research studies on social determinants of health!