# Introduction to Survival Analysis in R

karin.isaev@gmail.com

November 24th, 2017

## What is survival analysis?

- Statistical approaches to investigate the time it takes for an event to occur
- Events may include death, onset of illness, recovery from illness (binary variables) or transition above or below the threshold of a continuous clinical variables (CD4 counts)
- Accommodates data from randomized clinical trial or cohort study designs

# What is survival analysis?

In cancer studies, typical research questions are like:

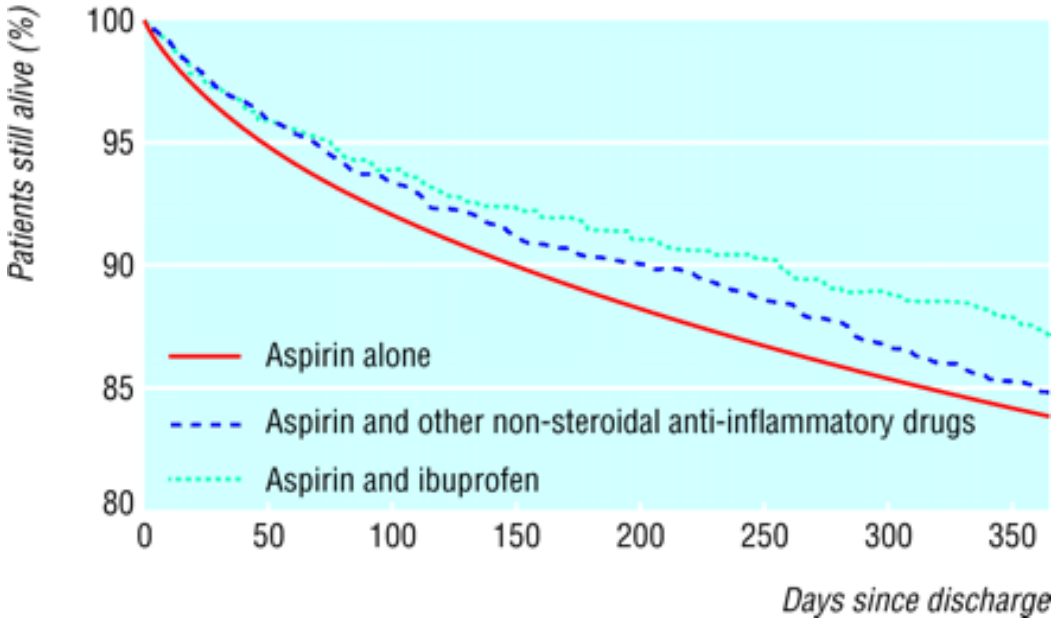
- What is the impact of certain clinical characteristics on patient's survival
- What is the probability that an individual survives 3 years?
- Are there differences in survival between groups of patients receiving different treatments?

# Example of survival analysis

Retrospective cohort study:
From December 2003 BMJ:
Aspirin, ibuprofen, and mortality after myocardial infarction

# Example of survival analysis

Retrospective cohort study:
From December 2003 BMJ:
Aspirin, ibuprofen, and mortality after myocardial infarction



## Why survival analysis ?

• Mean time-to-event between your groups using t-test/ linear regression ?

Ignores censoring

• Compare proportion of events in group using odds ratios or logistic regression?

Ignores time

## Outline

1. Terms
2. Kaplan-Meier plots for visualizing survival curves
3. Log-rank test to compare survival curves
4. Cox proportional hazards regression

## Outline

1. **Terms**
2. Kaplan-Meier plots for visualizing survival curves
3. Log-rank test to compare survival curves
4. Cox proportional hazards regression

# Fundamental Terms

- Survival time and event
  - Time = time to death, relapse-free survival time
  - Event = death, relapse
  - Important to define when presenting your data
  - Survival time commonly referred to as time from response to treatment to occurrence of event

# Fundamental Terms

- Censoring
  - Survival analysis focuses on the duration of time until the occurrence of an event (relapse or death)
  - The event may not be observed for some people within the time period = *censored* observations (right censoring in this case)
  - **Assumption: censoring must be independent of the event we are looking at**

# Fundamental Terms

- Survival and Hazard Functions
  - S*urvival function* = survival probability, S(t)
    - Probability that an one **survives** from time t (diagnosis), to a future time t
  - H*azard function*= h(t)
    - Probability that one **experiences** event at that time (baseline covariates)

# Censoring (Right)

- A subject does not experience the event before the study ends
- Lost to follow-up during the study period
- Withdrawal from the study

# Review question 1

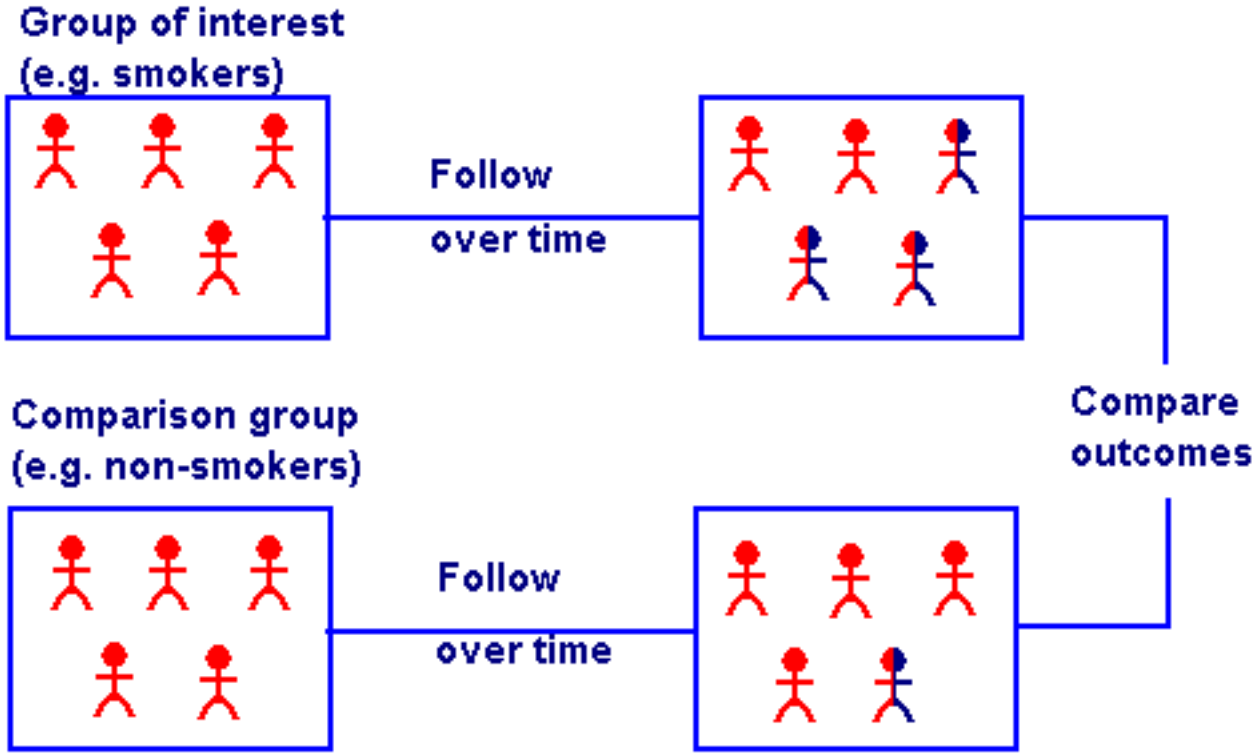
Which of the following data sets is likely to lend itself to survival analysis?

1. A case-control study of caffeine intake and breast cancer
2. A randomized controlled trial where the outcome was whether or not women developed breast cancer in the study period
3. A cohort study where the outcome was the time it took women to develop breast cancer
4. A cross-sectional study which identified both whether or not women have ever had breast cancer and their date of diagnosis

# Review question 1

Which of the following data sets is likely to lend itself to survival analysis?

1. A case-control study of caffeine intake and breast cancer
2. A randomized controlled trial where the outcome was whether or not women developed breast cancer in the study period
3. **A cohort study where the outcome was the time it took women to develop breast cancer**
4. A cross-sectional study which identified both whether or not women have ever had breast cancer and their date of diagnosis




http://www.statisticshowto.com/cohort-study/

## Outline

1. Terms
2. Kaplan-Meier plots for visualizing survival curves
3. Log-rank test to compare survival curves
4. Cox proportional hazards regression

## Outline

1.Terms

**2.Kaplan-Meier plots for visualizing survival curves**

3.Log-rank test to compare survival curves

4.Cox proportional hazards regression

# Kaplan-Meier survival estimate

Non-parametric method that estimates survival probability from observed survival times

The survival probability at time $t_i$, $S(t_i)$, is calculate as:

$$S(t_i) = S(t_{i-1})(1-d_i/n_i)$$

- S(ti-1) = the probability of being alive at ti-1
- ni = the number of patients alive just before ti
- di = the number of events at ti
- t0 = 0, S(0) = 1

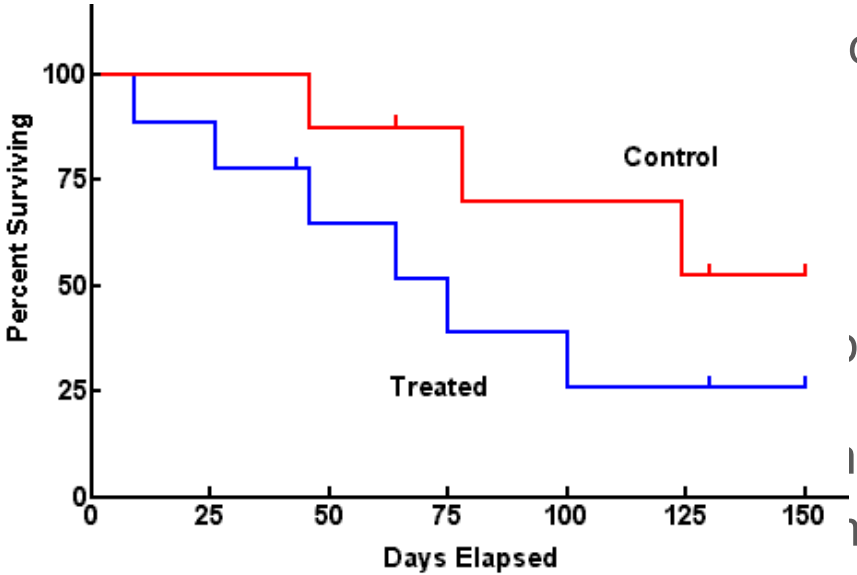
## Kaplan-Meier survival estimate $S(t_i) = S(t_{i-1})(1-d_i/n_i)$

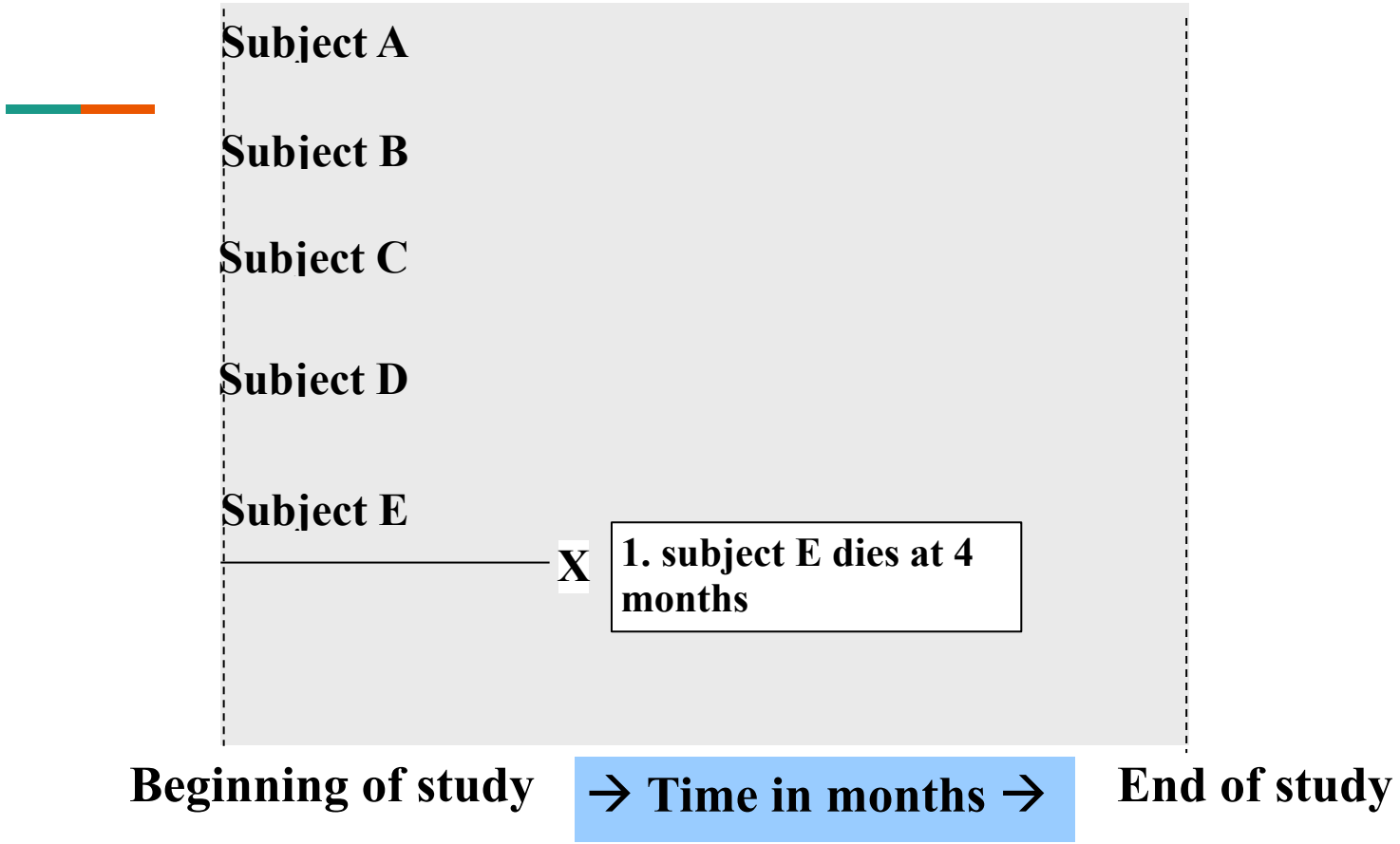
- Non-parametric method that estimates survival probability from observed survival times
- With censoring in mind
- S(t) changes only at the time of each event = step function
- Can generate KM survival curve that nicely summarizes the data and allows further estimation of median survival time for example

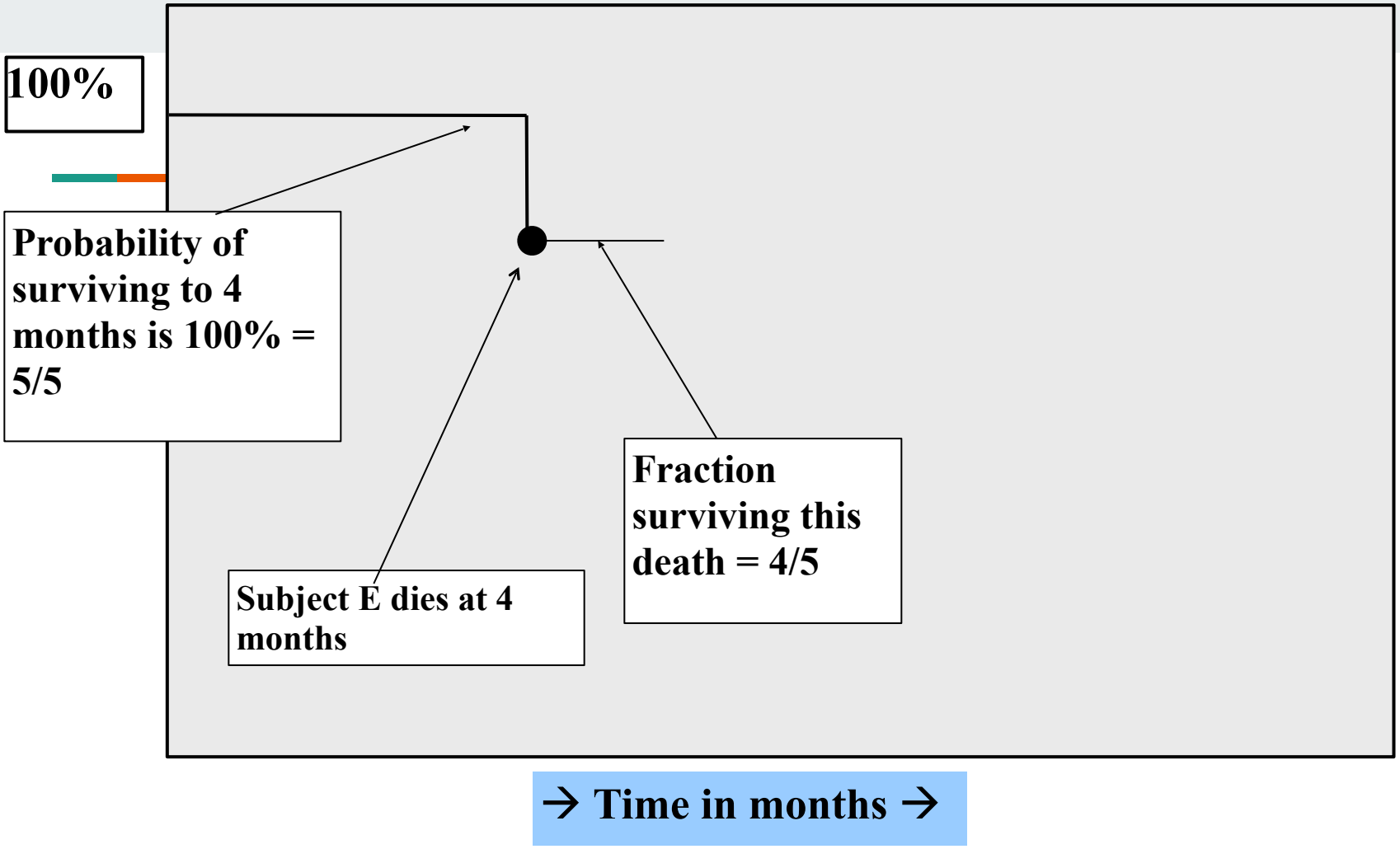
# Kaplan-Meier survival estimate

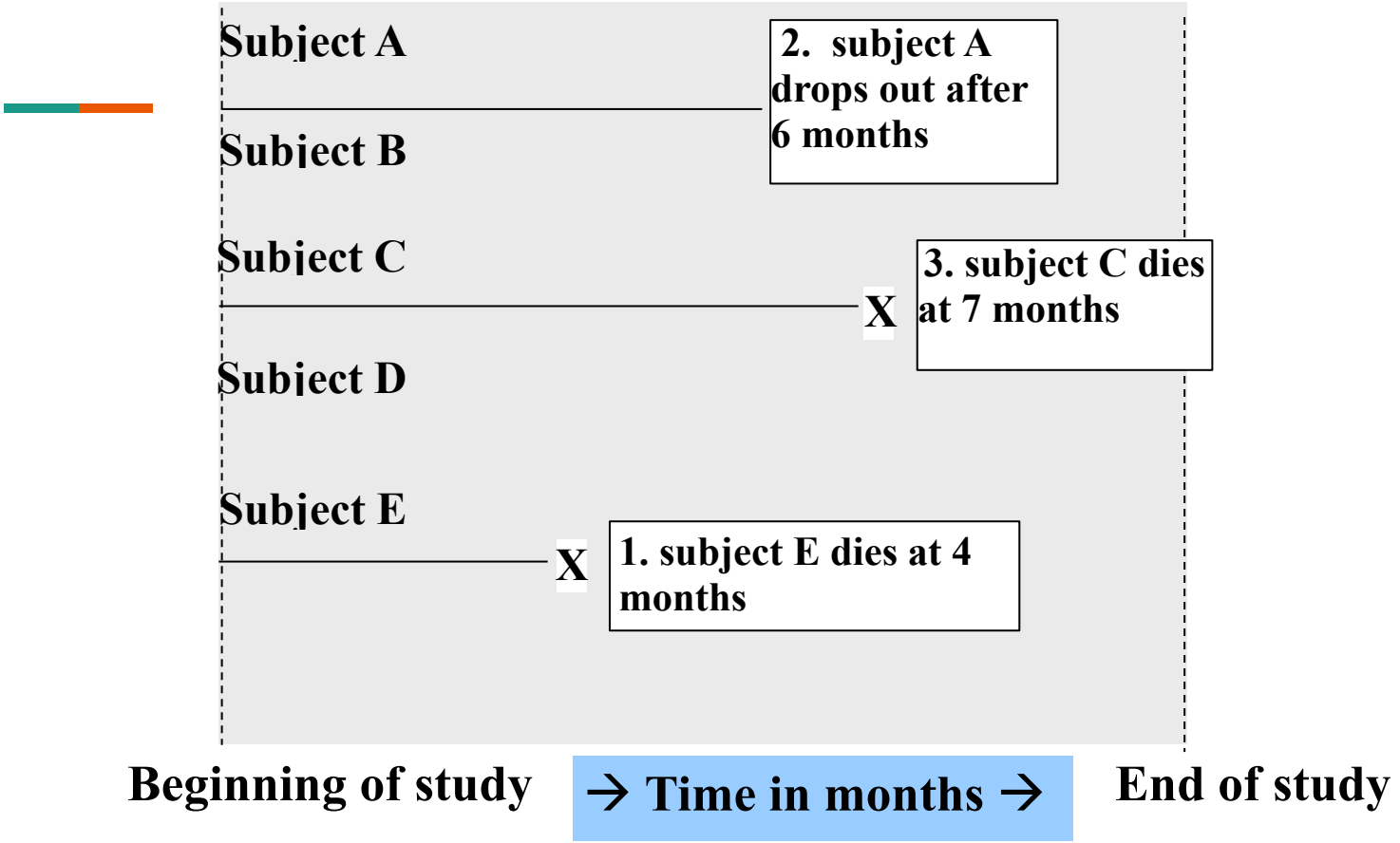
$$S(t_i) = S(t_{i-1})(1-d_i/n_i)$$

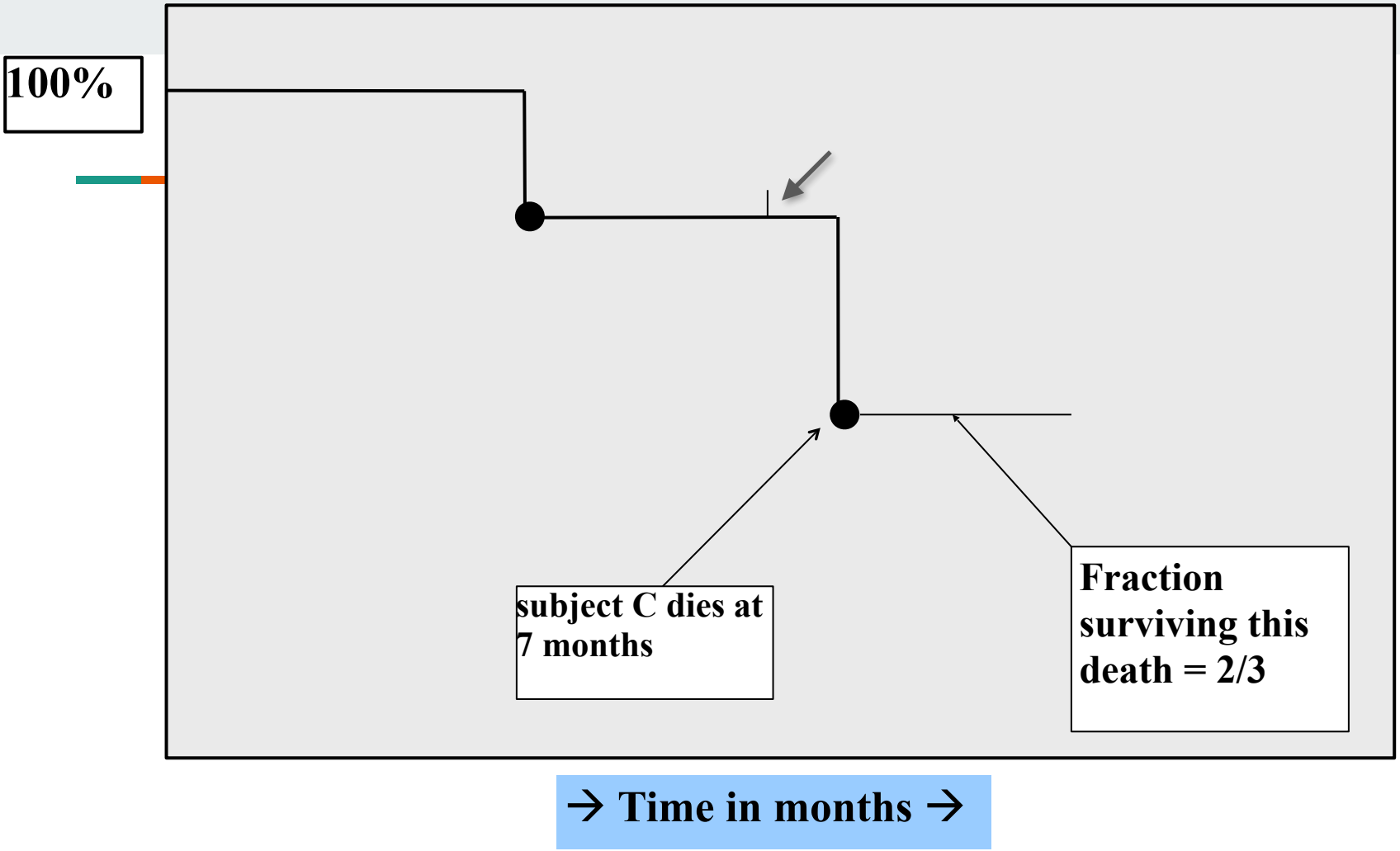
- Non-param obability from observed su
- With censo
- S(t) change ) function
- Can genera arizes the data and allows f ne for example

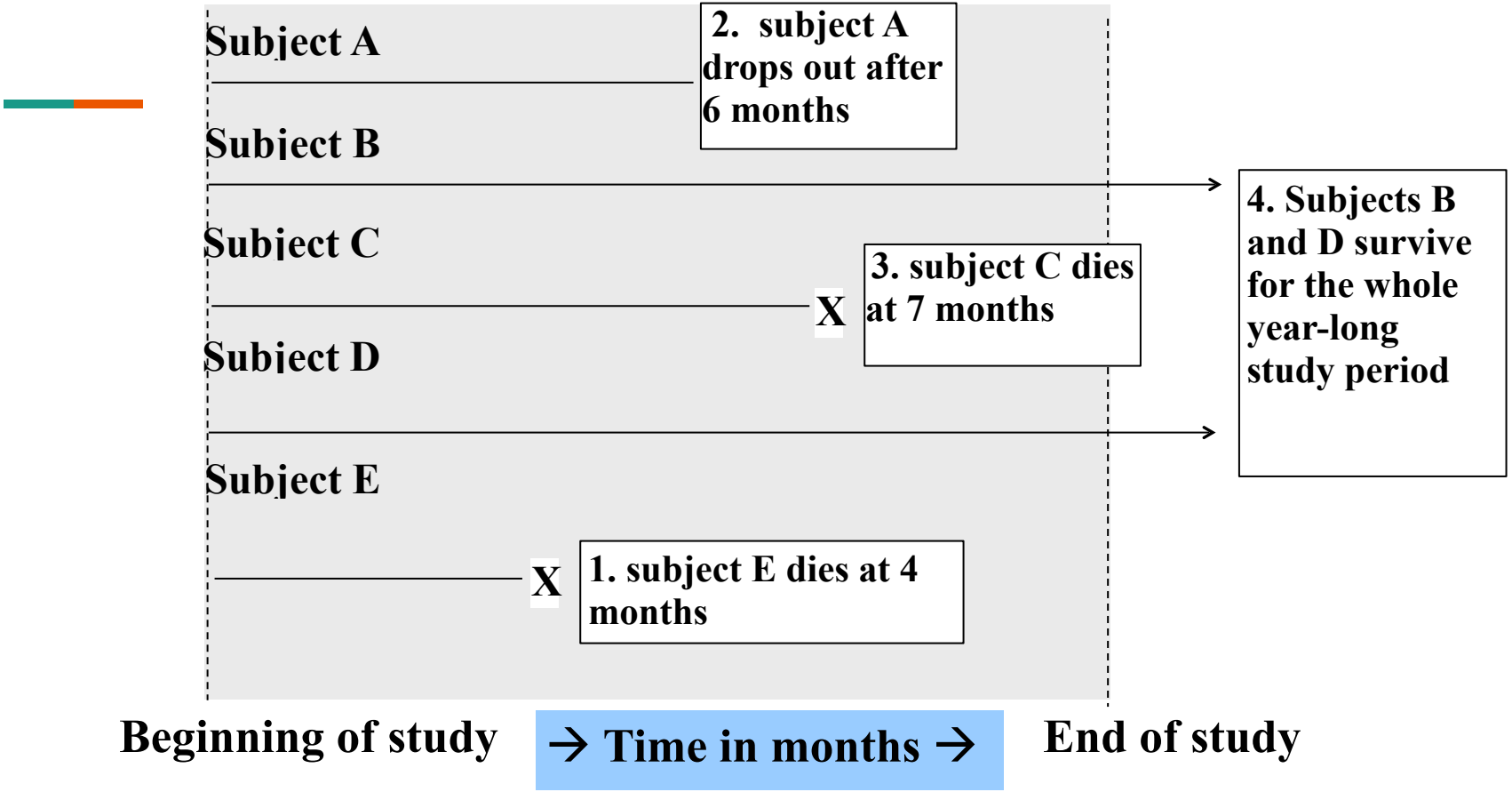



Subject A
Subject B
Subject C
Subject D
Subject E
X
1. subject E dies at 4 months
Beginning of study
→ Time in months →
End of study


100%
Probability of surviving to 4 months is 100% = 5/5
Subject E dies at 4 months
Fraction surviving this death = 4/5
→ Time in months →


Subject A
2. subject A drops out after 6 months
Subject B
Subject C
X
3. subject C dies at 7 months
Subject D
Subject E
X
1. subject E dies at 4 months
Beginning of study
→ Time in months →
End of study


100%
subject C dies at 7 months
Fraction surviving this death = 2/3
→ Time in months →


Subject A
2. subject A drops out after 6 months
Subject B
4. Subjects B and D survive for the whole year-long study period
Subject C
X
3. subject C dies at 7 months
Subject D
Subject E
X
1. subject E dies at 4 months
Beginning of study
→ Time in months →
End of study

# What is the probability of surviving an entire year?

Rule from probability theory:

P(A and B) = P(A) * P(B) if A and B independent

In survival analysis: intervals are defined by failures (2 intervals leading to failures here)

P(surviving intervals 1 and 2) = P(surviving interval 1) * P(surviving interval 2)

∴ Product limit estimate of survival =
P(surviving interval 1 up to failure 1) *
P(surviving interval 2 up to failure 2)
= 4/5 * 2/3 = 0.53

## What is the probability of surviving an entire year?

- = (4/5) * (2/3) = 53%
- > 40% (2/5) because drop-out survived at least until that point
- < 60% (3/5) because unsure if drop-out would have survived until the end of the year

## Outline

1. Terms
2. Kaplan-Meier plots for visualizing survival curves
3. Log-rank test to compare survival curves
4. Cox proportional hazards regression

## Outline

1.Terms

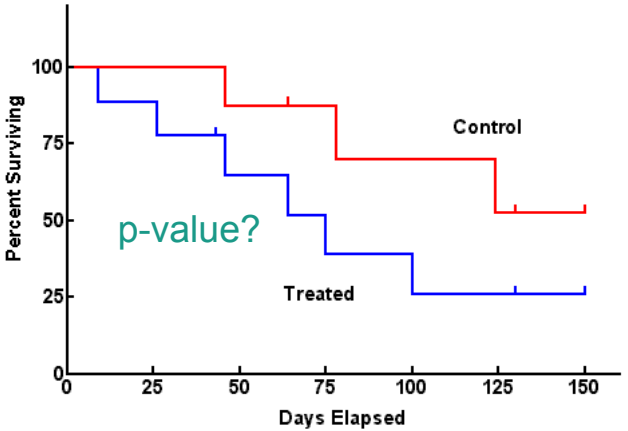
2.Kaplan-Meier plots for visualizing survival curves

**3.Log-rank test to compare survival curves**

4.Cox proportional hazards regression

# Log-Rank test

- Most widely used to compare two or more survival curves
- Null hypothesis is that there is no difference in survival between the two curves or groups
- Makes no assumption about the survival distributions (non-parametric)
- Compares the observed number of events in each group to what would be expected if the null hypothesis was true



# Log-Rank test

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31
- First death in week 6, risk of death during this week = 1/51

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31
- First death in week 6, risk of death during this week = 1/51
- If null hypothesis is true, the expected number of deaths is:

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31
- First death in week 6, risk of death during this week = 1/51
- If null hypothesis is true, the expected number of deaths is:
    - Group1 =  20 * 1/51 = 0.39

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31
- First death in week 6, risk of death during this week = 1/51
- If null hypothesis is true, the expected number of deaths is:
    - Group1 = 20 * 1/51 = 0.39
    - Group2 = 31 * 1/51 = 0.61

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31
- First death in week 6, risk of death during this week = 1/51
- If null hypothesis is true, the expected number of deaths is:
    - Group1 = 20 * 1/51 = 0.39
    - Group2 = 31 * 1/51 = 0.61
- Second death in week 10 with two deaths, at this point 19 and 31 at risk (alive), the probability of event is now 2/50

# Log-Rank test

- 51 subjects, group1 = 20, group2 = 31
- First death in week 6, risk of death during this week = 1/51
- If null hypothesis is true, the expected number of deaths is:
    - Group1 = 20 * 1/51 = 0.39
    - Group2 = 31 * 1/51 = 0.61
- Second death in week 10 with two deaths, at this point 19 and 31 at risk (alive), the probability of event is now 2/50
- Expected, group1 = 19*2/50 = 0.76, group2= 31*2/50 = 1.24

# Log-Rank test

- Group1
    - Total # of expected deaths = 22.48
    - Observed = 14
- Group2
    - Total # of expected deaths = 19.52
    - Observed = 28

# Log-Rank test

- The test statistic is the sum of (O - E)^2/E for each group
- O and E are the totals of the observed and expected events
- Here (14 - 22.48)^2 / 22.48 + (28 - 19.52)^2 / 19.52 = 6.88
- Chi Square Test
- $p < 0.01$ with 1 DF

## Log-Rank test

- Limitation = only assesses the effect of one variable at a time
- Test of significance, does not provide an estimate of the size of the difference between the groups
- Gives all calculations the same weight regardless of the time at which event occurs
    - Peto log-rank test statistic gives more weight to earlier events when there are a large number of observations

# Review question 2

Investigators studied a cohort of individuals who joined a weight-loss program by tracking their weight loss over 1 year. Which of the following statistical test is likely the most appropriate test for evaluating the effectiveness of the weight loss program?

1. A two-sample t-test.
2. ANOVA
3. Repeated-measures ANOVA
4. Chi-square
5. Kaplan-Meier methods

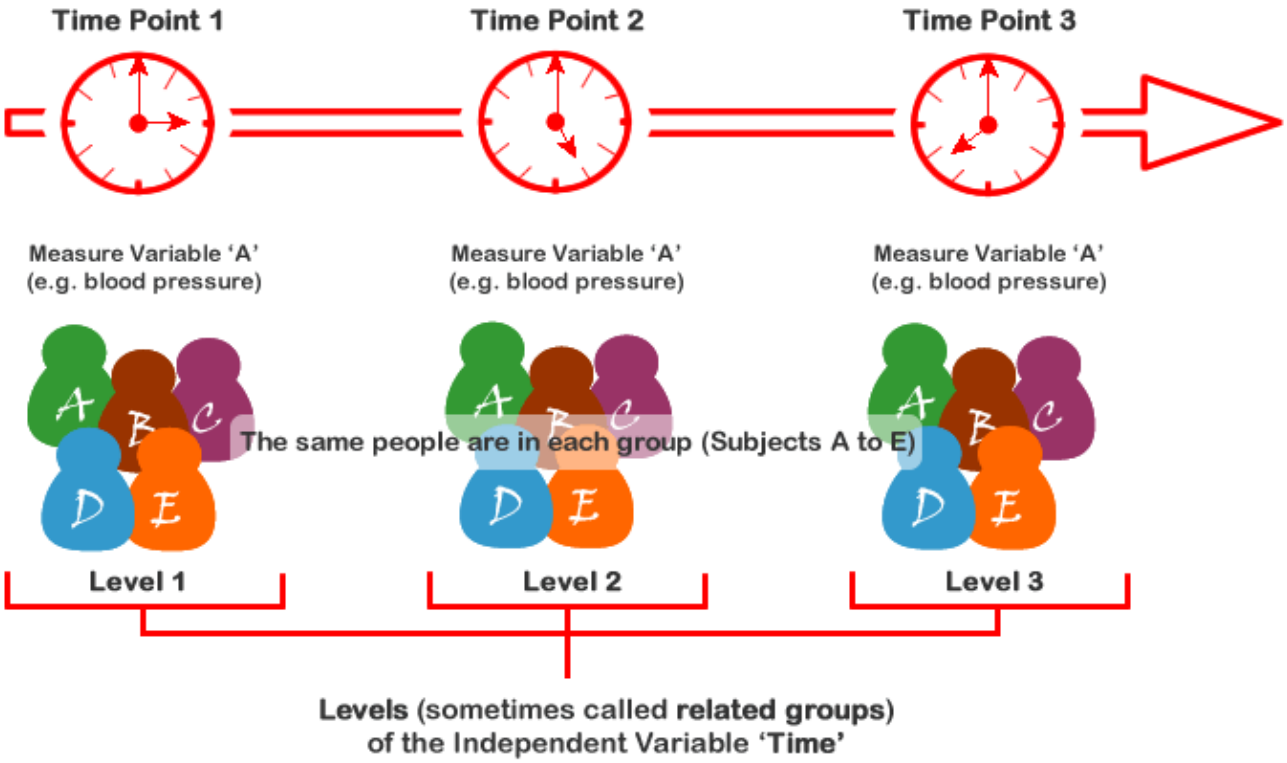
# Review question 2

Investigators studied a cohort of individuals who joined a weight-loss program by tracking their weight loss over 1 year. Which of the following statistical test is likely the most appropriate test for evaluating the effectiveness of the weight loss program?

1. A two-sample t-test
2. ANOVA
3. **Repeated-measures ANOVA**
4. Chi-square
5. Kaplan-Meier methods

Time is the independent variable

In survival analysis, **time to event** is the dependent variable



# Review question 3

Investigators compared mean cholesterol level between cases with heart disease and controls without heart disease. Which of the following is likely the most appropriate statistical test for this comparison?

1. A two-sample t-test
2. ANOVA
3. Repeated-measures ANOVA
4. Chi-square
5. Kaplan-Meier methods

# Review question 3

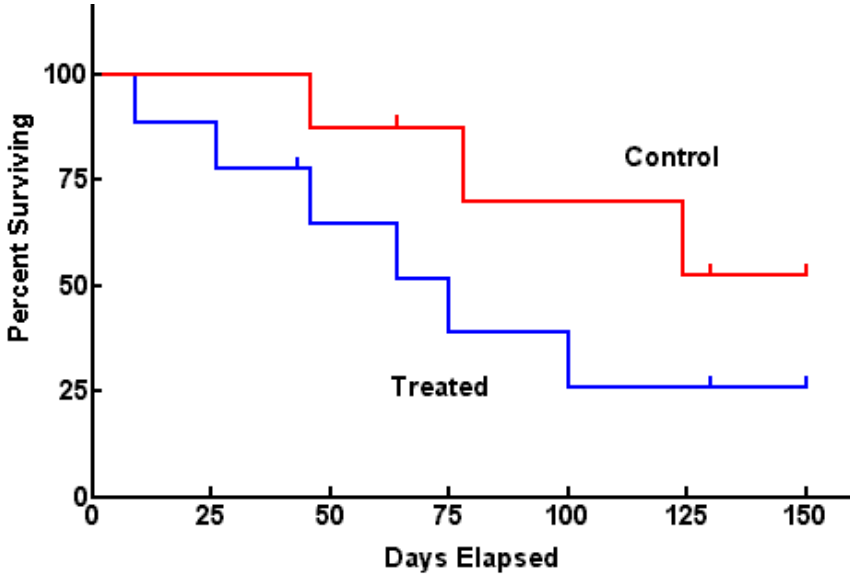
Investigators compared mean cholesterol level between cases with heart disease and controls without heart disease. Which of the following is likely the most appropriate statistical test for this comparison?

1. **A two-sample t-test**
2. ANOVA
3. Repeated-measures ANOVA
4. Chi-square
5. Kaplan-Meier methods

# Review question 4

- Which is the correct statement describing this K-M curve?

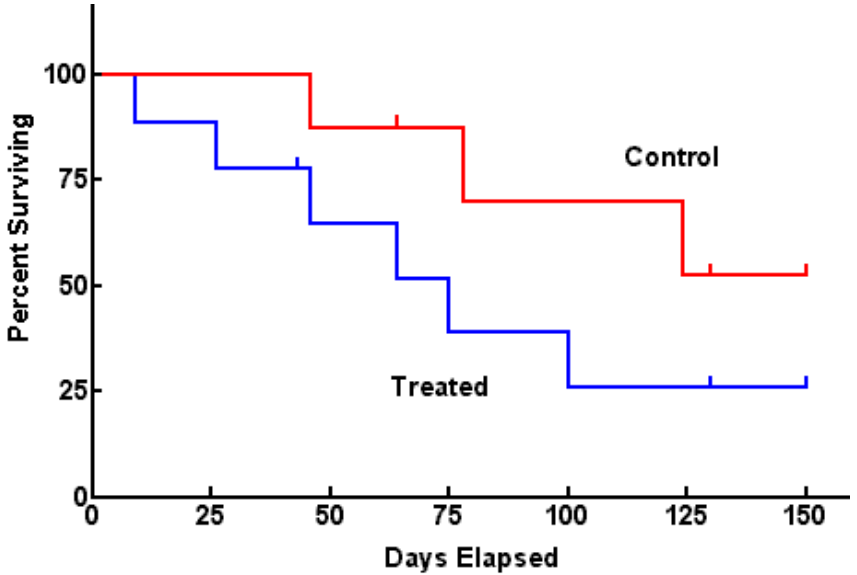
1. The mortality rate was higher in the control group than the treated group.
2. The probability of surviving past 100 days was about 50% in the treated group.
3. The probability of surviving past 100 days was about 70% in the control group.
4. Treatment should be recommended.



# Review question 4

- Which is the correct statement describing this K-M curve?

1. The mortality rate was higher in the control group than the treated group.
2. The probability of surviving past 100 days was about 50% in the treated group.
3. **The probability of surviving past 100 days was about 70% in the control group.**
4. Treatment should be recommended.



## Outline

1. Terms
2. Kaplan-Meier plots for visualizing survival curves
3. Log-rank test to compare survival curves
4. Cox proportional hazards regression

## Outline

1.Terms

2.Kaplan-Meier plots for visualizing survival curves

3.Log-rank test to compare survival curves

**4.Cox proportional hazards regression**

# Cox Proportional Hazards Model

- KM curves and log-rank tests are univariate types of analysis, ignore impact of other factors
- KM and log-rank tests are most often used when variables are categorical, wouldn’t work for quantitate variables such as gene expression…
- We can assess the effects of several variables or risk factors on time-to-event (dependent variable)
- Estimates adjusted hazard ratios (relative rather than absolute risk)

# Cox Proportional Hazards Model

- Cox model is expressed by the hazard function

$$h(t) = h_0(t) \times exp(b_1x_1 + b_2x_2 + \ldots + b_px_p)$$

- $h_0$ is the baseline hazard corresponding to the hazard if all the variable coefficients are set to 0
- $exp(b_i)$ are the hazard ratios (HR)
- A covariate with HR > 1 is called bad prognostic factor
- A covariate with HR < 1 is called good prognostic factor

# Cox Proportional Hazards Model - Assumptions

Key assumption is that the hazard curves for groups of observations should be proportional and cannot cross

- Hazard function for the patient k:

$$h_k(t) = h_0(t)e^{\sum_{i=1}^{n} \beta x}$$

- Hazard function for the patient k':

$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^{n} \beta x'}$$

- The hazard ratio for these two patients $[\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_{i=1}^{n} \beta x}}{h_0(t)e^{\sum_{i=1}^{n} \beta x'}} = \frac{e^{\sum_{i=1}^{n} \beta x}}{e^{\sum_{i=1}^{n} \beta x'}}]$ is independent of time t.

http://www.sthda.com/english/wiki/cox-proportional-hazards-model#the-need-for-multivariate-statistical-modeling

# Cox Proportional Hazards Model

- This is why it is called a proportional-hazards model = hazard curves for groups should be proportional
- This baseline hazard function itself it not estimated within the model (the hazard function obtained when all covariates are set to 0)
- Pro: No risk of misspecifying baseline distribution, doesn't make arbitrary assumption about the shape/form of the baseline hazard function
- Con: Model is incompletely specified for future uses of the model

# Cox model - Proportional Hazards Assumption

- Important to assess whether a fitted Cox regression adequately describes our data
- The proportional hazards assumption can be checked used scaled Schoenfeld **residuals** which are independent of time (plot!)
- Residuals = difference between the observed predictor and the expected given the risk set at that time
- Residuals should have no correlation
    - Calculated for each covariate

# Review question 5

Exponentiating a beta-coefficient from Cox regression gives you what?

1. Odds ratios
2. Risk ratios
3. Hazard ratios
4. None of the above

# Review question 5

Exponentiating a beta-coefficient from Cox regression gives you what?

1. Odds ratios
2. Risk ratios
3. **Hazard ratios**
4. None of the above

# Cox model - Proportional Hazards Assumption Violations

- A violations of proportional hazards assumption can be resolved by:
    - Adding covariate*time interaction
- Stratification
    - Allows the form of the underlying hazard function to vary across levels of stratification variables (treatment or age variables for example)
    - Allows factor to be adjusted without estimating its effect

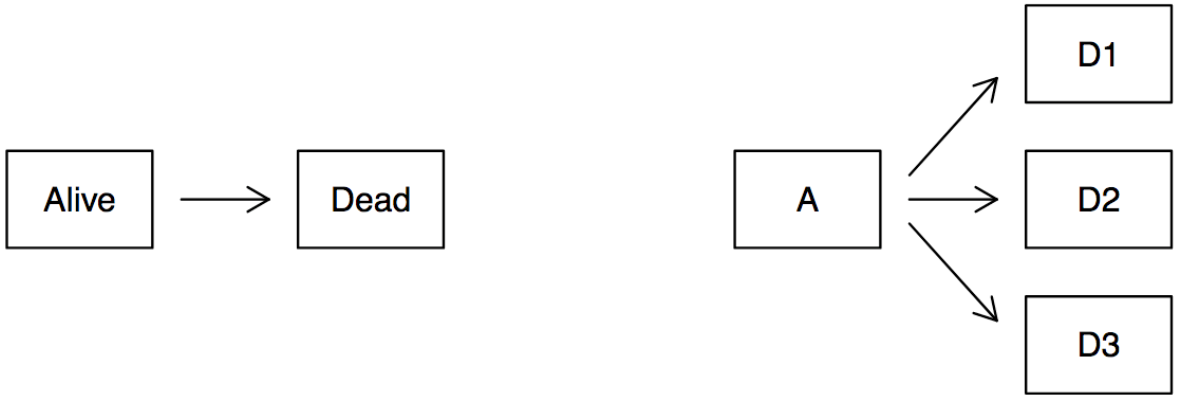
# Cox model - Non independent observations

- Can deal with this using cluster(variable_name) within coxph
- For example patients in a practice in one hospital versus another

# A little on competing risks in survival analysis

- Until now we have assumed there is only one survival endpoint of interest, death for example
- Also assumed that censoring is independent of the event of interest
- In real life, there could be several different types of events (relapse, infection) all which could be of interest to us
- The occurrence of one of these may or may influence the occurrence of other events —> competing risks

# A little on competing risks in survival analysis



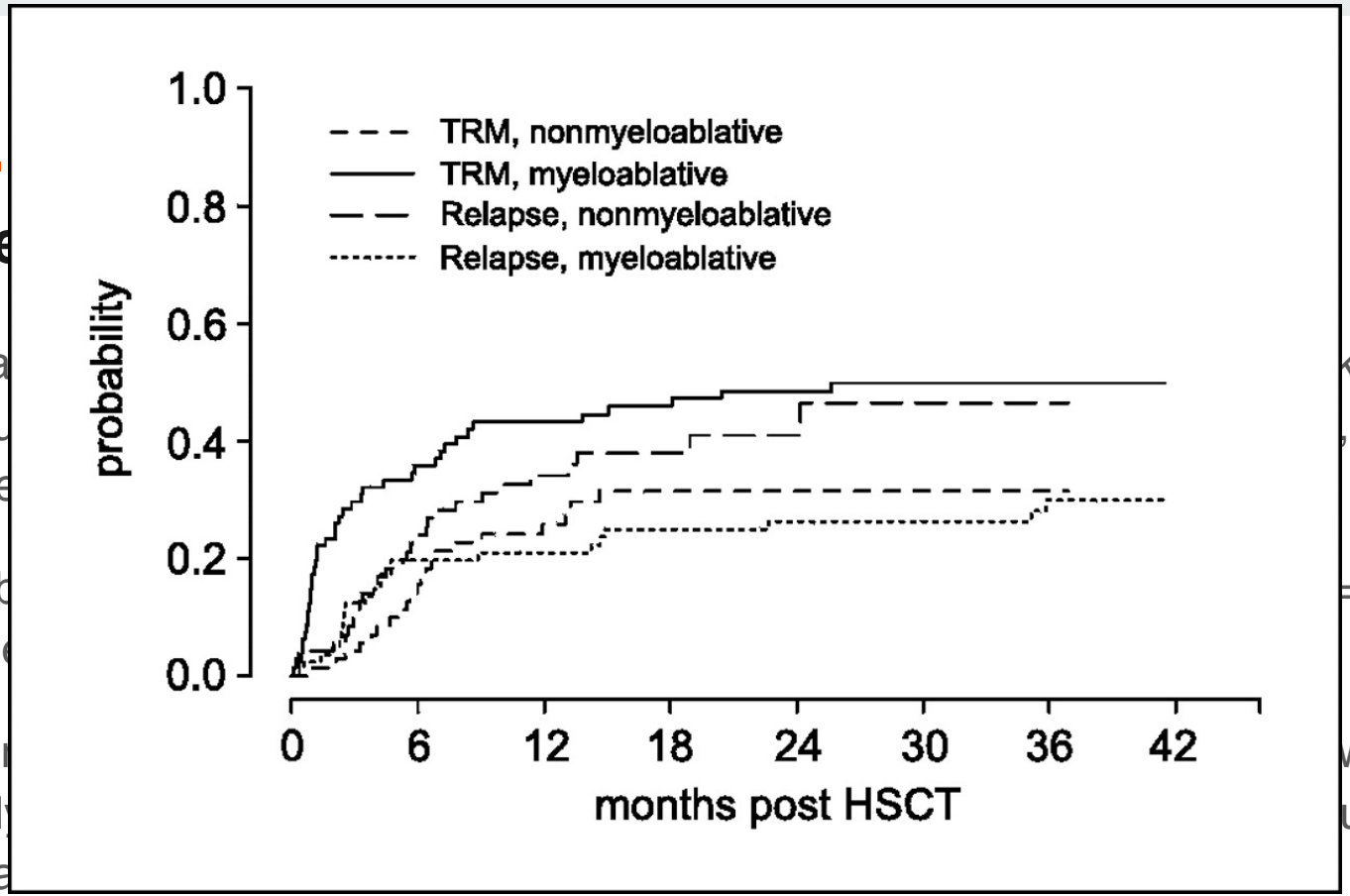
# A little on competing risks in survival analysis

- After a bone marrow transplantation, patients are followed to evaluate leukemia-free survival where the end point is time to relapse of leukaemia or death , whichever occurs first
- The above endpoint is made up of two types of failures (competing risks) = relapse and non-relapse deaths
- If different event types are independent of each other, we can apply what we already learned , the only difference would be that the “event” variable would now be a factored variable and we will have more curves

Kim, 2007
DOI: 10.1158/1078-0432.CCR-06-1210

# A little

- After a
- The ab
- If differ




Kim, 2007
DOI: 10.1158/1078-0432.CCR-06-1210

# Let’s try running some analysis in R!