

# **Survival Analysis (time to event data)**

---

**Dr Azmi Mohd Tamil**

# What is Survival Analysis?

- A collection of statistical tests that analyses the duration of time till the event that we are interested in occurs.
- The time measured can be in years, month, week or days which is also known as survival time.

# Survival Tests

- Life Table
- Kaplan Meier
- Cox Regression
- Cox Regression with time dependent variable

# Event or endpoint

Can be whatever that we are interested in;

- Death
- Disease
- Convalescence
- Recovery/Cured
- Relapse

Although more than 1 event may occur, only one event is taken into account

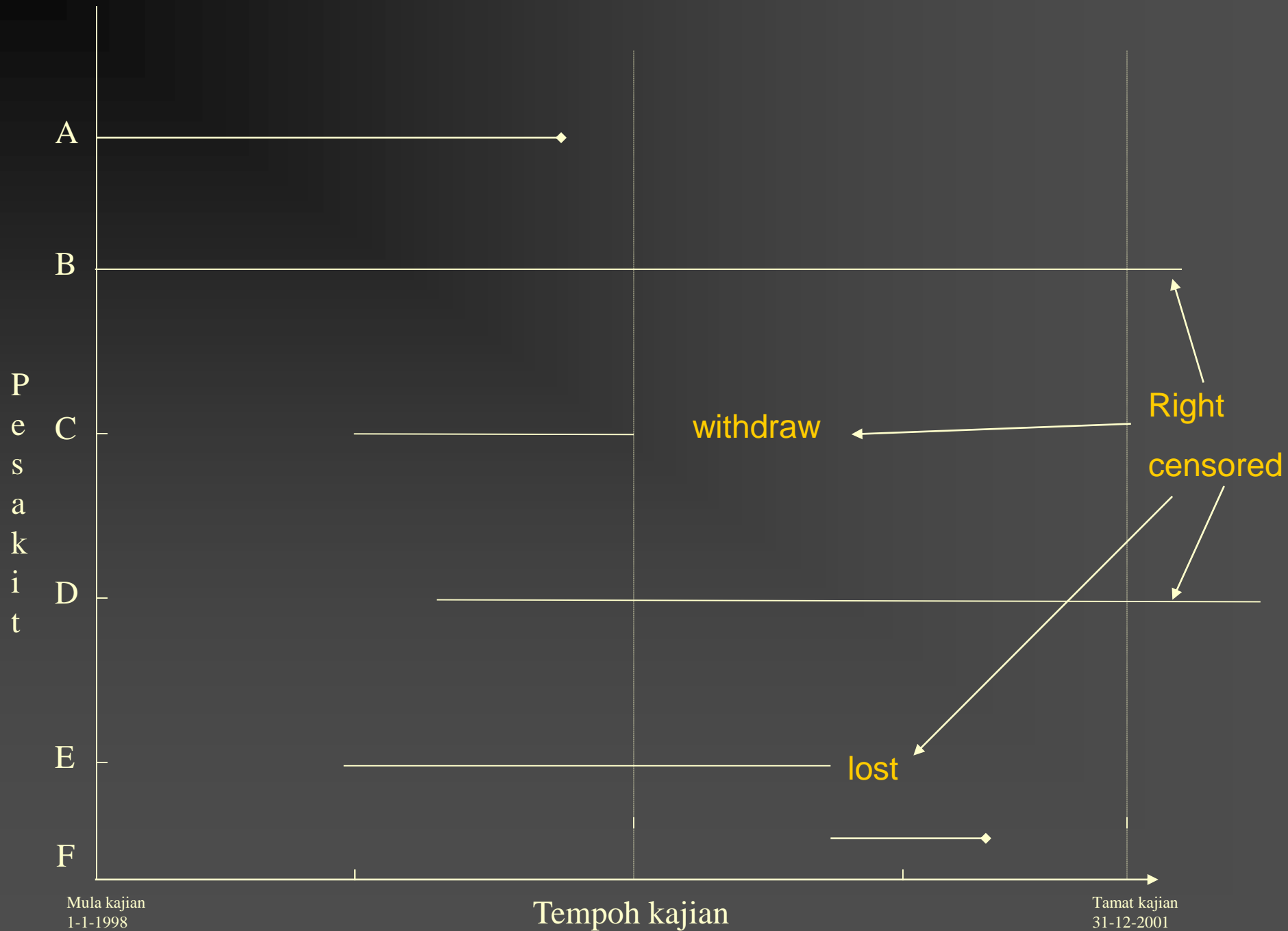


# Survival Analysis

- When doing a time to event study, e.g.,
  - time to cure or
  - time to death or
  - time to pregnancy
- More information is available than alive/dead
- Some individuals may not have a terminating event (they are **censored alive** :o)

# Censored Data

- An analytical problem
- Occurs when we do not know the survival time accurately
- Occurs when;
  - Sample do not reach the endpoint within the period of study
  - Sample lost to follow up during the study period
  - Sample withdraw due to death (when death is not the endpoint being studied) or due to adverse events



# Example

Patient	Surv. Time	Status
A	5	1
B	12	0
C	3.5	0
D	8	0
E	6	0
F	3.5	1



# Survival Time (T)

- Any positive value of time
- Equal or larger than zero

# Status

- 1 if endpoint is reached
- 0 if;
  - Patient survive till end of study
  - Patient lost to follow-up
  - Patient withdraw from the study for whatever reasons.

# Right Censored

- Lost to follow up
- Withdraw
- Study ended without endpoint being reached
- Most censored data are censored to the right

# Left Censored

- When we do not know when patient have the inclusion event;
- i.e. Survival of HIV patient, whereby we do not know how long he was HIV positive prior to being diagnosed



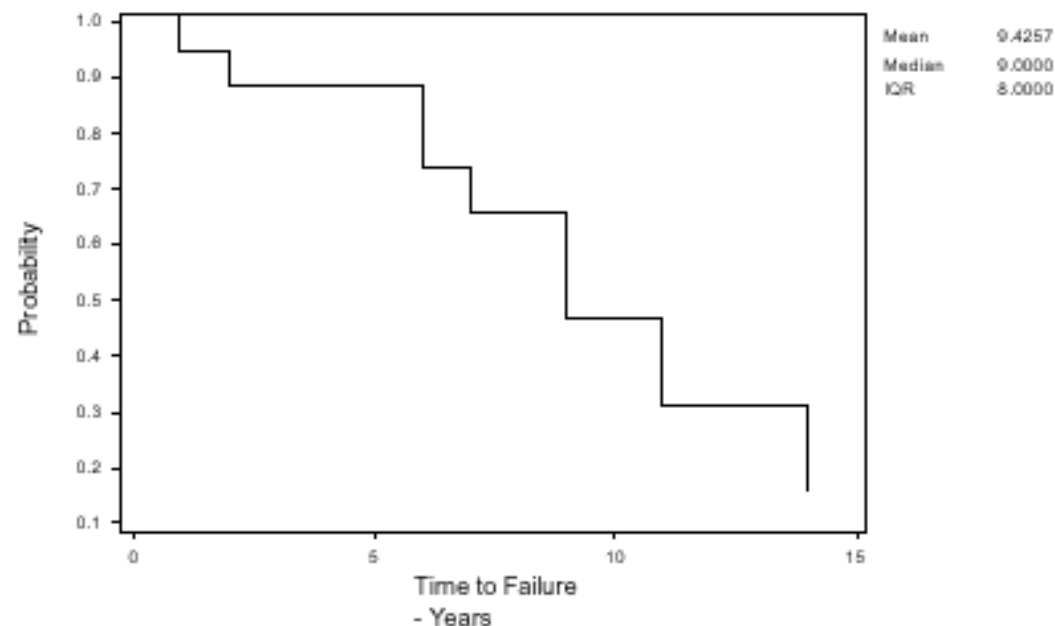
# Survival Analysis

- The aim of survival analysis is to
  - estimate the probability of surviving to a certain time,
  - compare survival in different patient groups and
  - determine what affects this process (much like regression analysis)
- The **Kaplan-Meier** approach assumes deaths only occur at the times they are observed

# Kaplan - Meier Survival Analysis

We estimate the Survival Curve  $S(t)$

Nonparametric Survival Plot for Parathyroid Cancer  
Kaplan-Meier Method



# Kaplan - Meier Survival Analysis

The survival curve is estimated by

$$\left( \begin{array}{c} \text{Proportion} \\ \text{Surviving} \\ \text{at } t \end{array} \right) = \left( \begin{array}{c} \text{Proportion} \\ \text{Surviving} \\ \text{Previously} \end{array} \right) \cdot \left( \frac{\# \text{alive at time } t}{\# \text{alive} + \# \text{dead at } t} \right)$$

$$S_i(t) = e^{-\int_0^t [h_0(t)] e^{b_0 + b_1 x_{i1} + \dots + b_p x_{ip}} dt}$$

where

$S_i(t)$  is the probability the  $i^{\text{th}}$  case survives past time  $t$

# Kaplan - Meier Survival Analysis

- The survival curve is estimated by

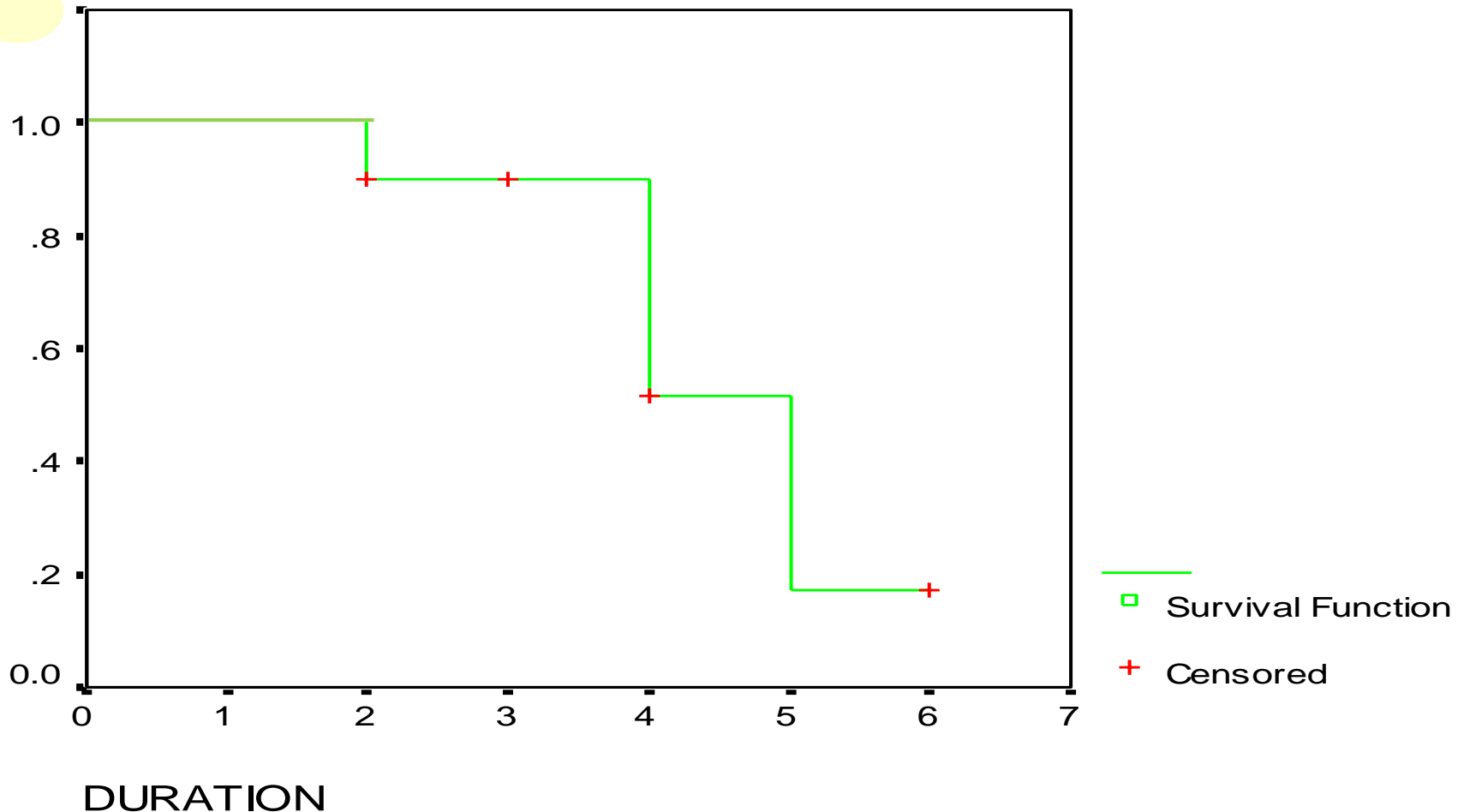
$$\left( \begin{array}{c} \text{Proportion} \\ \text{Surviving} \\ \text{at } t \end{array} \right) = \left( \begin{array}{c} \text{Proportion} \\ \text{Surviving} \\ \text{Previously} \end{array} \right) \cdot \left( \frac{\# \text{alive at time } t}{\# \text{alive} + \# \text{dead at } t} \right)$$

- for example for 10 subjects with survival time (in years) of  
2, 2\*, 3\*, 4, 4, 4, 4\*, 5, 5, 6\*
- $S(0) = 1$  is the starting point
- $S(2) = 9/10 * 1 = 0.90$
- $S(4) = 4/7 * (0.9) = 0.56 * 0.9 = 0.50$
- $S(5) = 1/3 * (0.5) = 0.33 * 0.5 = 0.17$



# Kaplan - Meier Survival Analysis

Survival Function



# Survivor Function $S(t)$

- Gives the probability of the patient surviving from the time given
- $S(t) = P(T > t)$
- e.g.  $S(5) = P(T > 5)$
- $t$  can be any value between 0 to infinity

# Survivor Function $S(t)$

- The value will decrease over time
- When  $t=0$ ,  $S(t)=S(0)=1$ . Since no one has yet reached endpoint at time 0.
- When  $t=\text{infinity}$ ,  $S(t)=S(\text{infinity})=0$ . Because by right nobody survives, therefore the curve turns to 0.



# Kaplan - Meier Survival Analysis

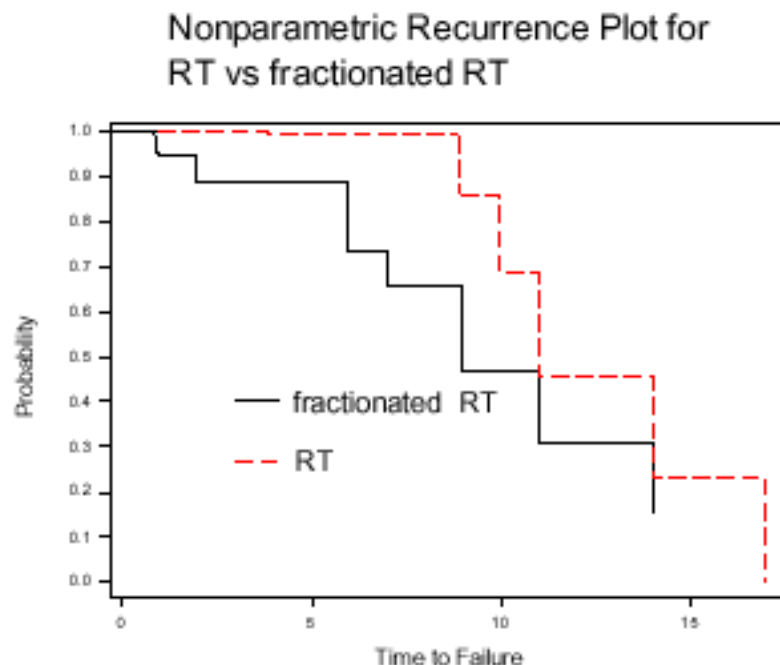
- Usually the aim is compare survival curves between groups,

You want to test whether a new fractionated radiotherapy treatment improves the proportion surviving compared to the standard two dose treatment

# Survival Analysis

The null hypothesis to compare the two groups is that the Curves are the same:

$$H_0: S_{RT}(t) = S_{\text{frac.RT}}(t)$$





# Survival Analysis Comparisons I

Comparing two Groups

$$H_0: S_{RT}(t) = S_{\text{frac.RT}}(t)$$

One method for comparing survival curves is the **Wilcoxon-Gehan** which is an extension of the Mann-Whitney nonparametric test comparing the **medians** of the survival times in each group by **ranking** them all together and adding up the ranks for the smaller group.

# Comparing Survival Curve

statistics to test the equality of the survival distributions

- log rank – all time points are weighted equally in this test
- Breslow – time points are weighted by the number of cases at risk at each time point
- Tarone-Ware – time points are weighted by the square root of the number of cases at risk at each time point.

# COMPARING 2 (OR MORE) SURVIVAL CURVES

Log rank test: equal weight to deaths  
throughout survival times

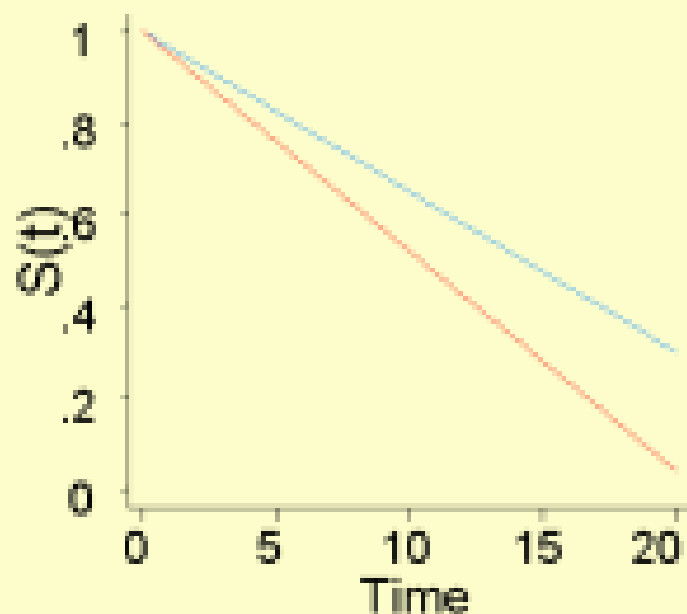
Breslow test: more weight to earlier  
deaths than later deaths

Tarone-Ware test: in between

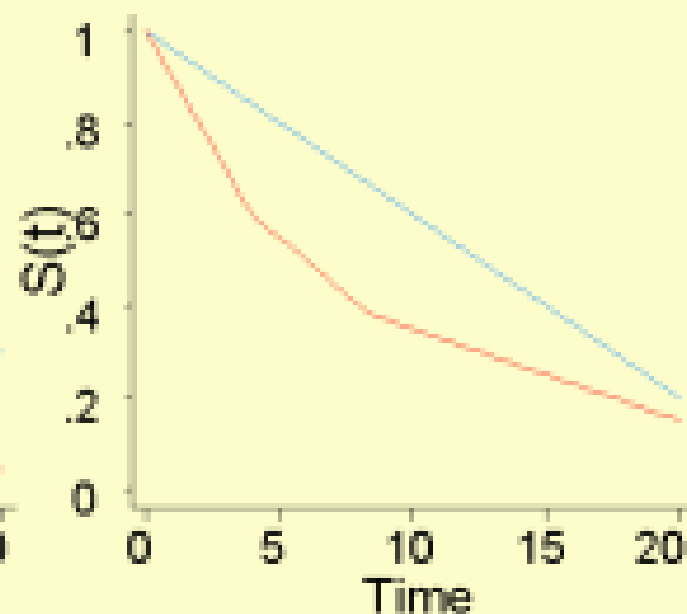


# WHICH TEST TO USE?

## Proportional hazard assumption



**Log rank test**  
**preferred (PH true )**



**Breslow test**  
**preferred (non-PH)**



# Survival Analysis Comparisons II

Comparing groups with covariate information, **e.g. age, gender, etc.** The most common method is the **Cox proportional hazards**, which is a regression-based method which assumes that instantaneous risk is proportional at each time point.

$$\lambda(t) = \lambda_0(t) e^{\alpha + \beta \text{treatment} + \gamma \text{age}}$$

It allows adjustment for differences between the groups.

$$H_0: \beta = 0$$



# Survival Analysis Comparisons II

The interpretation of  $\beta$  is similar to logistic regression.

$e^{\beta}$  is the hazard ratio

# Survival Analysis




Exponential Survival



Constant Hazard

# Hazard Function $h(t)$

- Potential of the event occurring after the patient has survive till time  $t$
- $h(t) \geq 0$ ;
  - No upward limit
  - Not a probability
  - Depends upon time
- Can predict  $h(t)$  if we know  $S(t)$



# An Example of a Survival Analysis from the NEJM

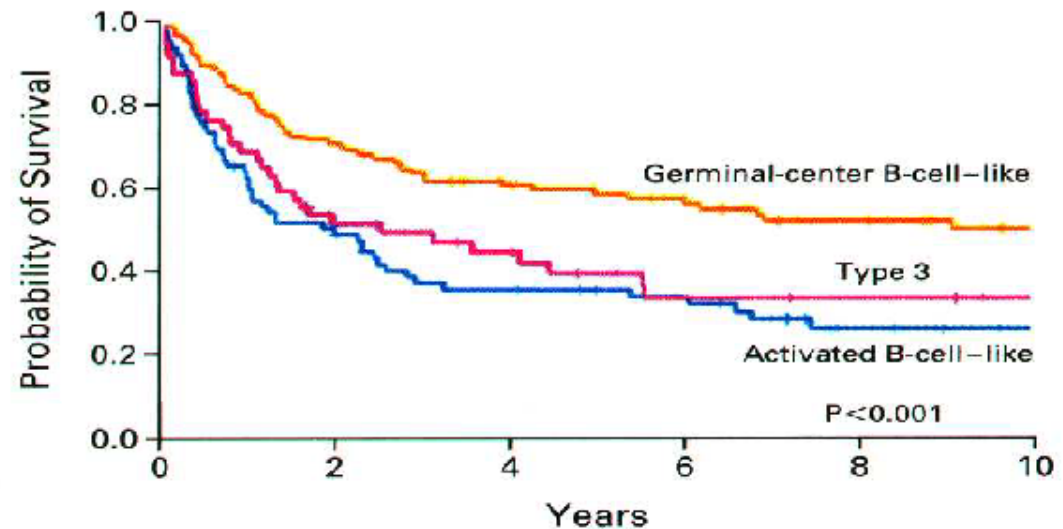
## The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma

Rosenwald, Andreas; Wright, George; Chan, Wing C.; Connors, Joseph M.; Campo, Elias; Fisher, Richard I.; Gascoyne, Randy D.; Muller-Hermelink, H. Konrad; Smeland, Erlend B.; Staudt, Louis M. ([N Engl J Med 2002;346:1937-47.](#))

- **Background:** The survival of patients with diffuse large-B-cell lymphoma after chemotherapy is influenced by molecular features of the tumors. We used the gene-expression profiles of these lymphomas to develop a molecular predictor of survival.
- **Methods:** Biopsy samples of diffuse large-B-cell lymphoma from 240 patients were examined for gene expression with the use of DNA microarrays and analyzed for genomic abnormalities.
- **Subgroups with distinctive gene-expression profiles were defined on the basis of hierarchical clustering.**
- **A molecular predictor of risk was constructed with the use of genes with expression patterns that were associated with survival in a preliminary group of 160 patients and was then tested in a validation group of 80 patients.**

# The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphom

C



No. AT RISK

Germinal-center B-cell-like	115	81	60	46	32	19
Type 3	52	24	18	10	8	5
Activated B-cell-like	73	35	23	19	8	5

(They) used a Cox proportional-hazards model to identify individual genes whose expression correlated with the outcome.

Data from 670 of 7399 microarray features were significantly associated with a good or a bad outcome in the preliminary group ( $P < 0.01$ ).





# The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma

## Conclusions:

**The subgroups differed substantially with respect to two recurrent oncogenic events.** The t(14;18) translocation involving the bcl-2 gene and the amplification of the c-rel locus on chromosome 2p occurred exclusively in germinal-center B-cell-like diffuse large-B-cell lymphomas.

These findings support the view that the various subgroups represent different diseases that arise as a result of distinct mechanisms of malignant transformation.



# Survival Analysis Methods

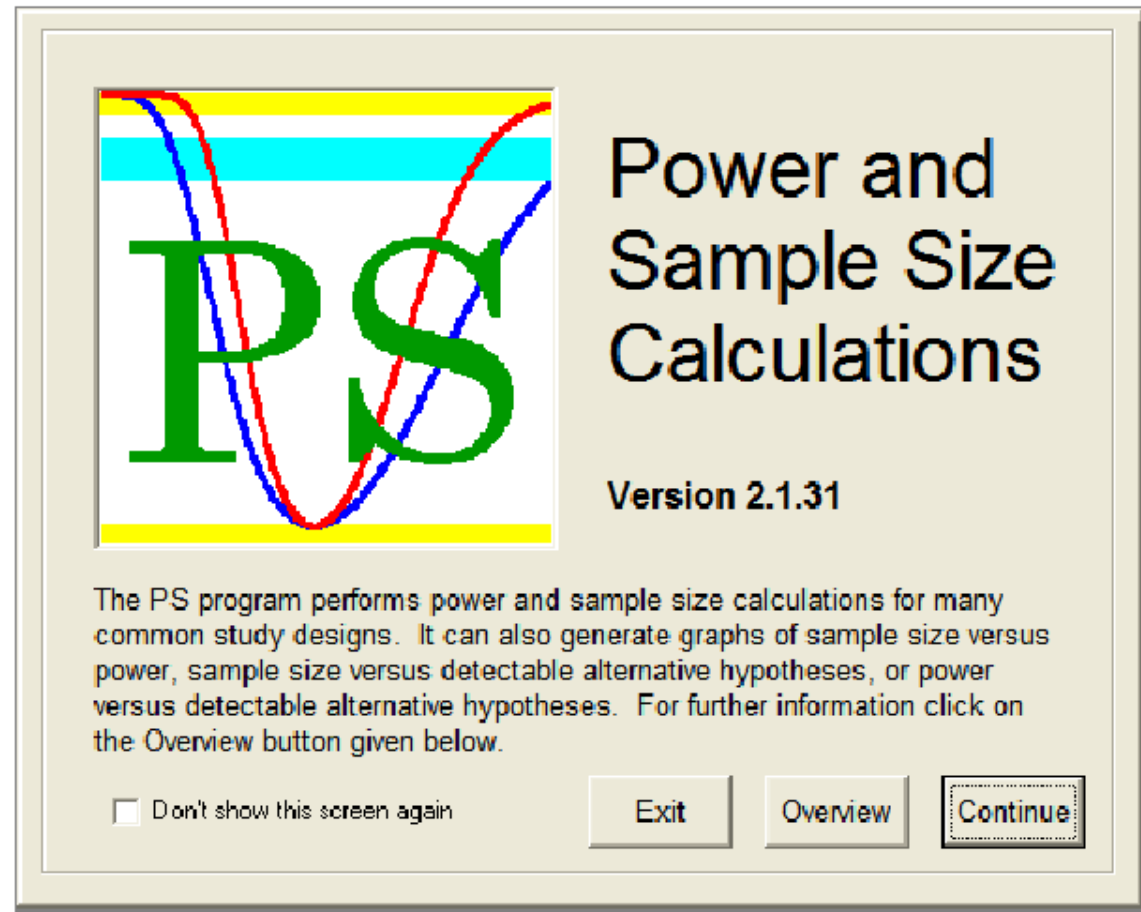
- Cox proportional hazards analysis will be used to compare the groups adjusting for differences in age of patient and stage of disease. The analysis will be stratified on stage if the proportionality assumption does not hold. (Survival Analysis: Techniques for Censored and Truncated Data, J Klein and M Moeschberger, Springer, 1997)



# The Sample Size for Survival Curves is Complicated

- <http://www.mc.vanderbilt.edu/prevmed/p/s/index.htm>
- **PS** is an interactive program for performing power and sample size calculations.
- It can be used for studies with dichotomous, continuous, or survival response measures. The alternative hypothesis of interest may be specified either in terms of differing response rates, means, or survival times, or in terms of relative risks or odds ratios.

# Power and Precision



# Data management

## ■ Required data for analysis

- ✓ ID, DOA – date of admission , DOT – date of termination, ST- status on termination
- ✓ Only enter other required data i.e. covariates
- ✓ Keep whatever data entered as it is

ID	DOA	DOT	ST	RS1	RS2	...	....	...
001								
002								
003								
...	....	...	....	...	...	...	...	..

# Data Analysis

- SPSS

- Analyze

- Survival

- Life Table

- Kaplan Meier

- Cox-Regression

- Cox x/time-dep cov

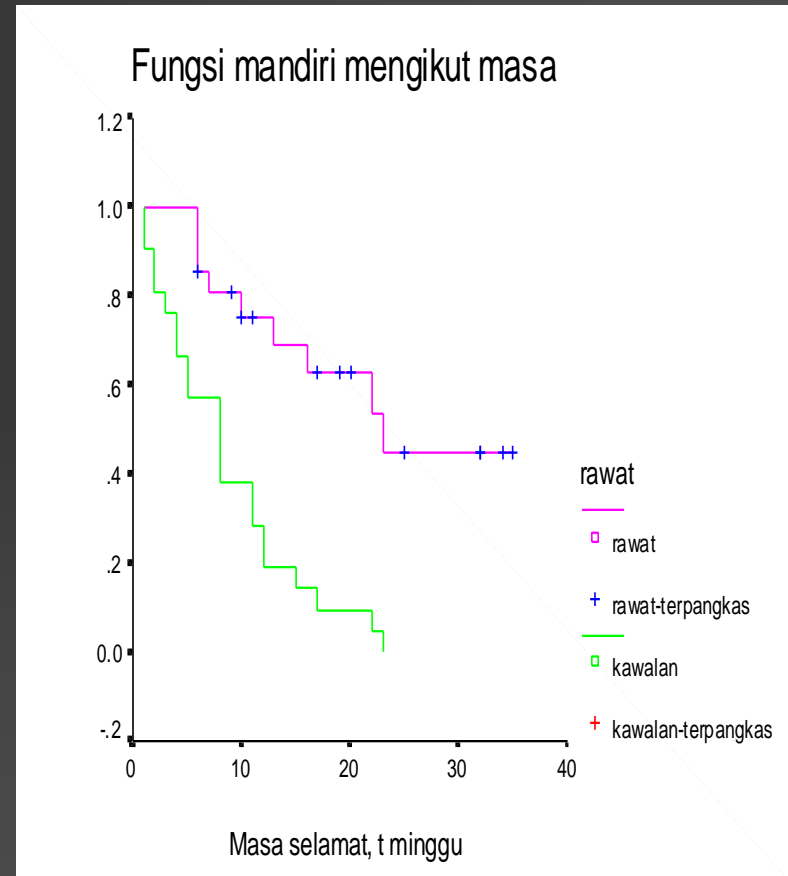
## .... Data Analysis

### ■ Life Table

- Time interval
- Graph plots – Survival, hazard, density function

### ■ Kaplan Meier

- Time by status
- Graph Plots – Survival, hazard, density function
- Log-rank test



## .... Data Analysis

### ■ Cox – Regression

- Identifies the association between hazard function, survival time density distribution and survival time.

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

$$h(t; \beta) = h_o(t) \exp(\beta_1 z_1 + \dots + \beta_p z_p)$$



$$h_i(t) = [h_0(t)] e^{b_0 + b_1 x_{i1} + \dots + b_p x_{ip}}$$

where

$h_i(t)$  is the hazard for the  $i^{\text{th}}$  case at time  $t$

$h_0(t)$  is the baseline hazard at time  $t$

$p$  is the number of covariates

$b_j$  is the value of the  $j^{\text{th}}$  regression coefficient

$x_{ij}$  is the value of the  $i^{\text{th}}$  case of the  $j^{\text{th}}$  covariate

## .... Data Analysis

### ■ Cox-Regression

- Will give an ANOVA table with comparative risk values and p values for significance of a variable. The example given is for 1 then 2 variables.

Model	Beta	Std Err	P-value	Risk exp( $\beta$ )
$h(t; \beta) = h_o(t) \exp(\beta_1 z_1)$	-1.509	0.410	0.000	0.221
$h(t; \beta) = h_o(t) \exp(\beta_1 z_1 +$ $\beta_2 z_2)$	-1.294	0.422	0.002	0.274
	1.604	0.329	0.000	4.95

## .... Data Analysis

- Cox w/ time dep. cov
  - ✓ A specific analysis used if there is a time related covariates such as age, follow-up interval etc.

## Survival Analysis

Survival analysis is concerned with analyzing the time to the occurrence of an event.

For instance, we have a dataset in which the times are 1, 5, 9, 20 and 22. These measures could be in seconds, hours, or days.

For example, the event is the time until a generator's bearings seize, the time until a cancer patient dies, or the time until a person finds employment.

The survival time  $T$  may be regarded as a random variable with a probability distribution  $F(t)$  and probability density function  $f(t)$ .

An obvious quantity of interest is the probability of surviving to time  $t$  or beyond, the survivor function or survival curve  $S(t)$ , which is given by

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

## Survival Analysis

A further function which is of interest for survival data is the hazard function.

This represents the instantaneous failure rate, that is, the probability that an individual experiences the event of interest at a time point given that the event has not yet occurred.

It can be shown that the hazard function is given by:

$$h(t) = f(t)/S(t),$$

the instantaneous probability of failure at time  $t$  divided by the probability of surviving up to time  $t$ .

# Survival Analysis

## Kaplan-Meier Estimator

**The Kaplan-Meier estimator is a nonparametric estimator of the survivor function  $S(t)$ .**

**If all the failure times, or times at which the event occurs in the sample, are ordered and labeled  $t(j)$  such that  $t(1) \leq t(2) \leq \dots \leq t(n)$ , the estimator is given by:**

$$\hat{S}(t) = \prod 1 - d_j / n_j$$

where  $d_j$  is the number of individuals who experience the event at time  $t(j)$ , and  $n_j$  is the number of individuals who have not yet experienced the event at that time and are therefore still “at risk” of experiencing it.

The product is over all failure times less than or equal to  $t$ .

We can compare survival in different subgroups by plotting the Kaplan-Meier estimators of the group-specific survivor functions and applying simple significance tests.

# Survival Analysis

## Cox Regression

**When there are several explanatory variables, and in particular when some of these are continuous, it is much more useful to use a regression method such as Cox.**

**Here the hazard function for individual  $i$  is modeled as:**

$$h_i(t) = h_0(t) \exp(\beta^T x_i)$$

where  $h_0(t)$  is the baseline hazard function,  $\beta$  are regression coefficients, and  $x_i$  covariates.

The baseline hazard is the hazard when all covariates are zero.