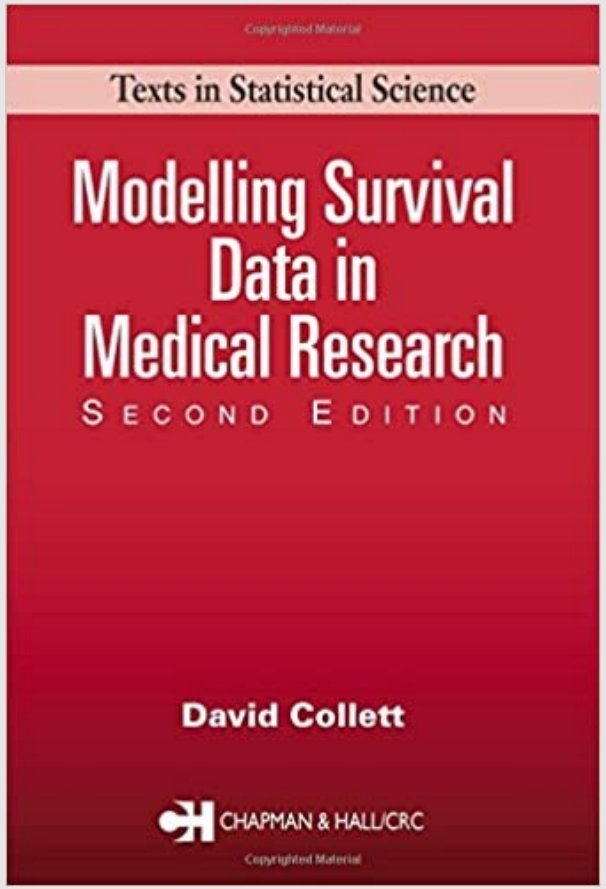
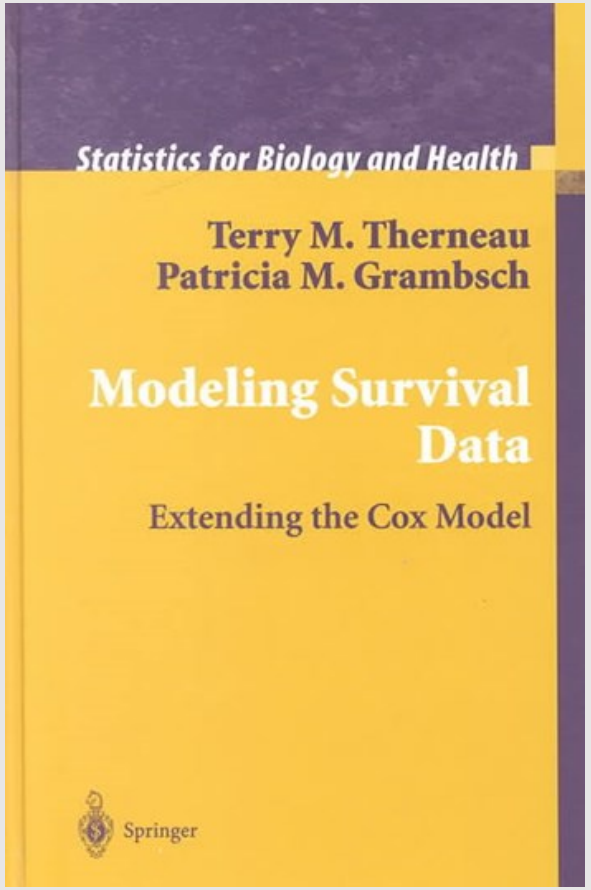
Survival analysis in epidemiological & clinical research, an introduction

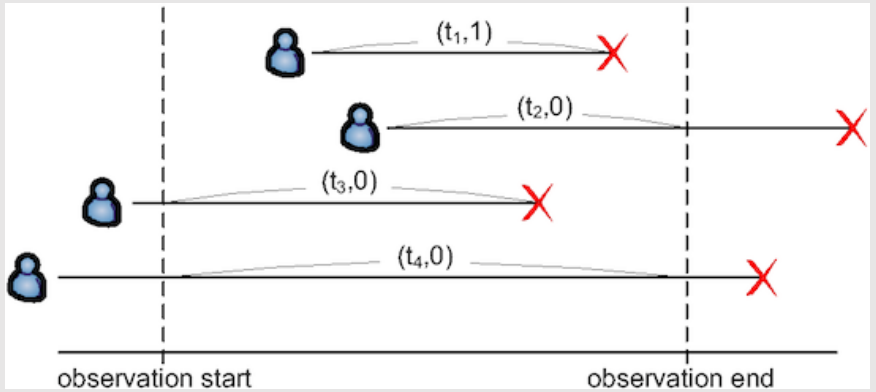


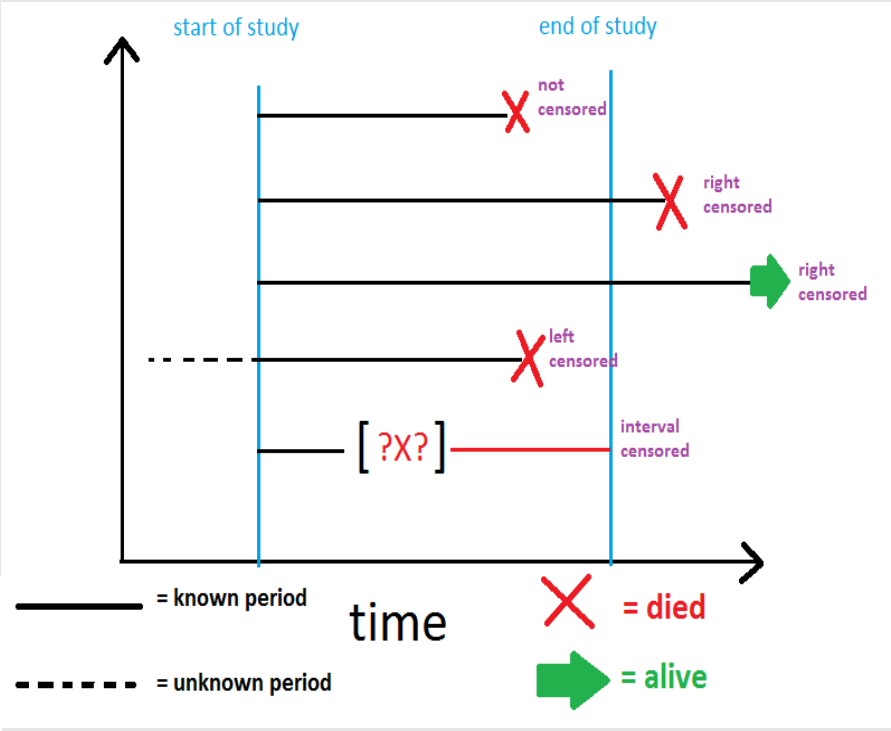


UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Summary:

- ***Survival*** *analysis definition*
- *Censoring*
- ***Kaplan-Meier*** *curve*
- *Cumulative* ***hazard*** *& hazard rate*





On a long enough time line, the survival rate for everyone will drop to zero.

-- Chuck Palahniuk, *Fight Club*

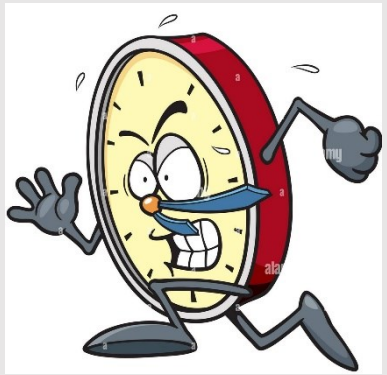
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

- Survival methods analyse data in the form of **times** from a well-defined ***time origin*** until the occurrence of some particular event or ***end-point*** **[OUTCOME]**

- Time is measured from the beginning of the observation until the event or the end of the study period

The object of the study is «time» (duration)

Initial moment
(study entry)

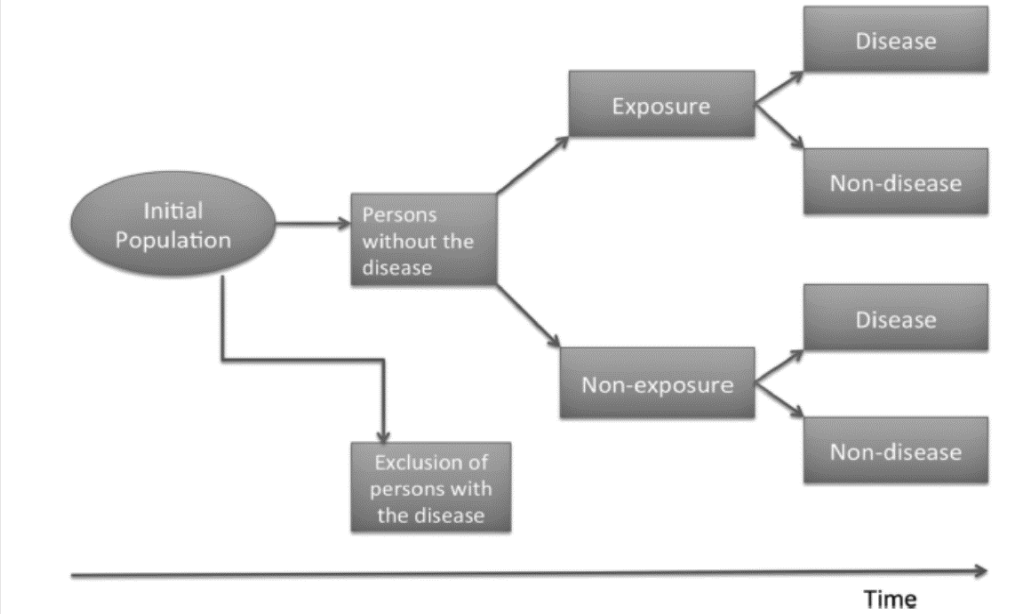


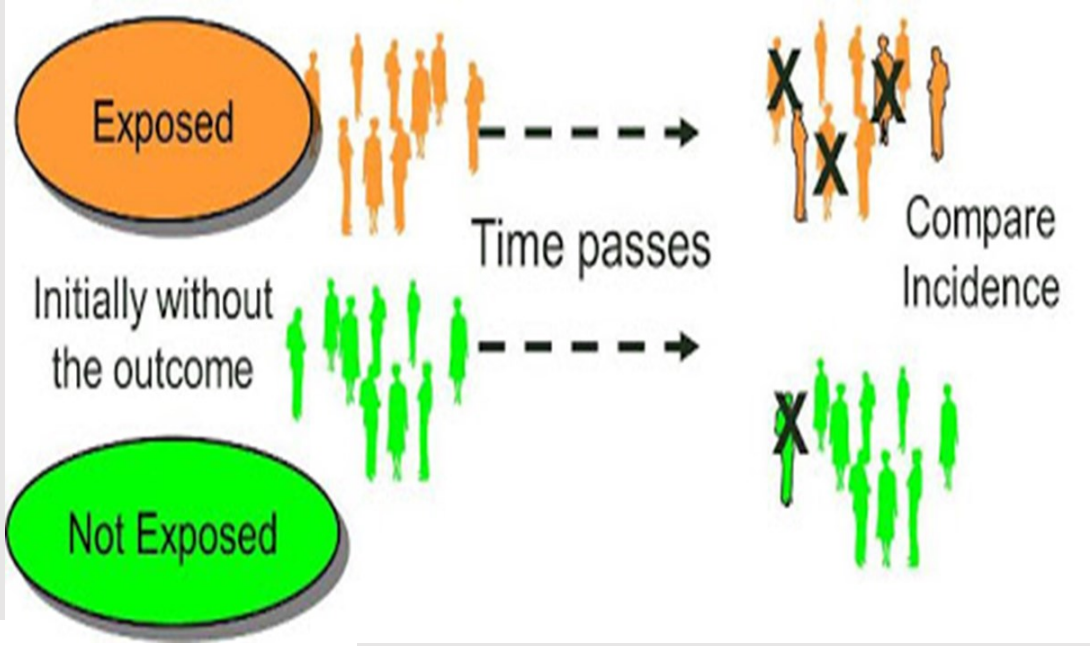
Time

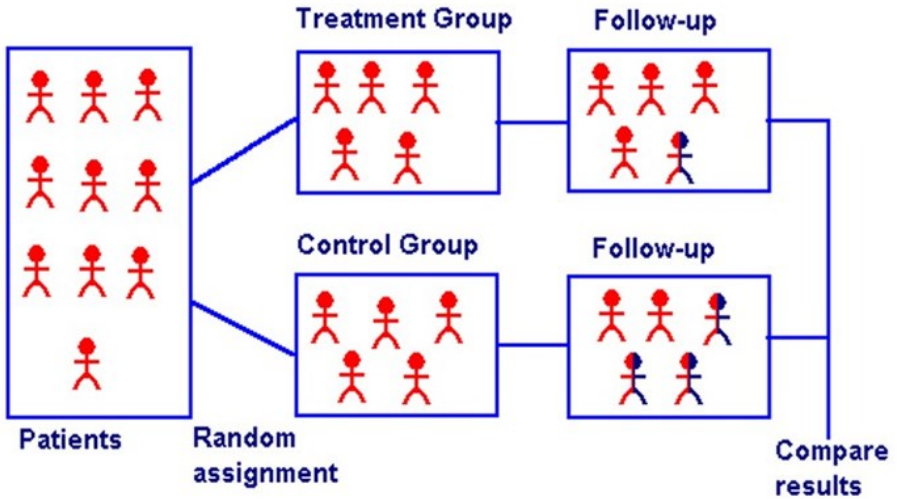
Final moment
(event/end of the study)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Study Design : population-based/cohort/RCTs







UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

A cohort of patients is observed for 24 months after an episode of myocardial infarction (AMI). The outcome is their survival.

- At the end of the observation **(follow-up)** each patient is identified by two values: ***(c,t)***

- **"c"** is the **patient status** (0=survived, 1=dead)

- ***"t"*** is the **duration** of the observation

The «**Time**» Factor:
Observing a group of subjects for a certain period means that the sample size **could change along time** for various reasons

Survival time

Survival time is a general term, could be used also in case of non-fatal events.

Survival analysis is used in all the studies in which **it is of interest** to evaluate the ***time until the occurrence*** of some events of interest (**incidence**).

Incidence : *probability of occurrence of a given condition in a population within a specified period of time*

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Examples

- time from **diagnosis** of a cardiovascular disease to **death** or a **cardiovascular event** (AMI/Stroke)
- time from **commencement** of a treatment to the occurrence of an **adverse effect**
- time from **enrollment in RCT** (drug vs placebo) to a **change** in a biomarker
- time from a **transplantion** procedure to the **Graft-versus-host disease** (GvHD)
- …

Survival time for subjects **without the event of interest** corresponds to the time between the **study entry** and the ***end*** of their observation

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Entry & "Exit"

They have to be clearly stated, without ambiguity

Initial Moment	Final Event/Exit
Randomization Date (RCT)	Date of Death
Date of Enrolment in a clinical registry	Date of First hospitalization
Diagnosis (date of a visit)	Date of recurrence of symptoms
Date of commencement of a therapy	Date of change in a biomarker
Date of a surgical procedure	Date of complication
…	…

Initial Moment, Final Event

To the initial moment should correspond a **date** for each subject :

- Randomization date

- Date of enrollment in a clinical registry….

To the *final* moment should correspond a **date** for each subject :

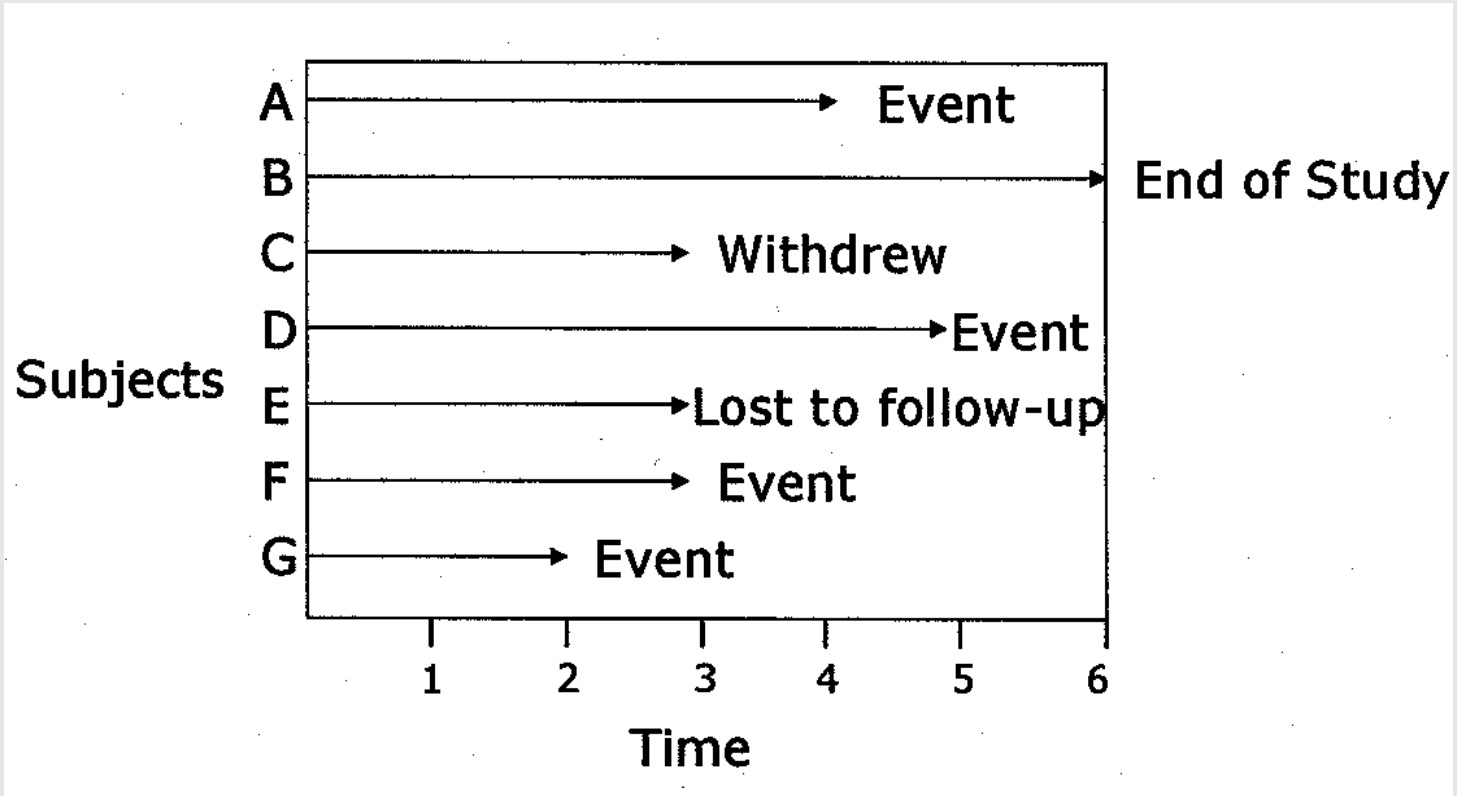
- Date of the event of interest

- **Censoring date :** last contact "free of event" or *administrative study closure*

Data Structure

ID	Initial date	Final date	follow_up	status	age	sex	LVEF	….
BTXVNX440331	3/30/1995	7/15/2000	63	1	50	1	32	…
CCRFRN400628	8/24/2001	11/1/2011	122	0	61	1	34	....
FFTMRN421006	3/31/1988	7/15/1996	99	1	45	0	24	….
GNLMRX381011	4/5/1993	11/1/2011	222	0	54	0	33	….

status: 0 = censoring, 1 = event



FOLLOW UP= Final date-Initial date

Independently from the initial date, all subjects start from **time 0** and are observed until they experience the event or *come out* from the observation.

Aims of Survival Analysis

- **Estimate *time-to-event* for a group of individuals**, such as time until hospitalization or death 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 CV patients?

Special *features* of survival data

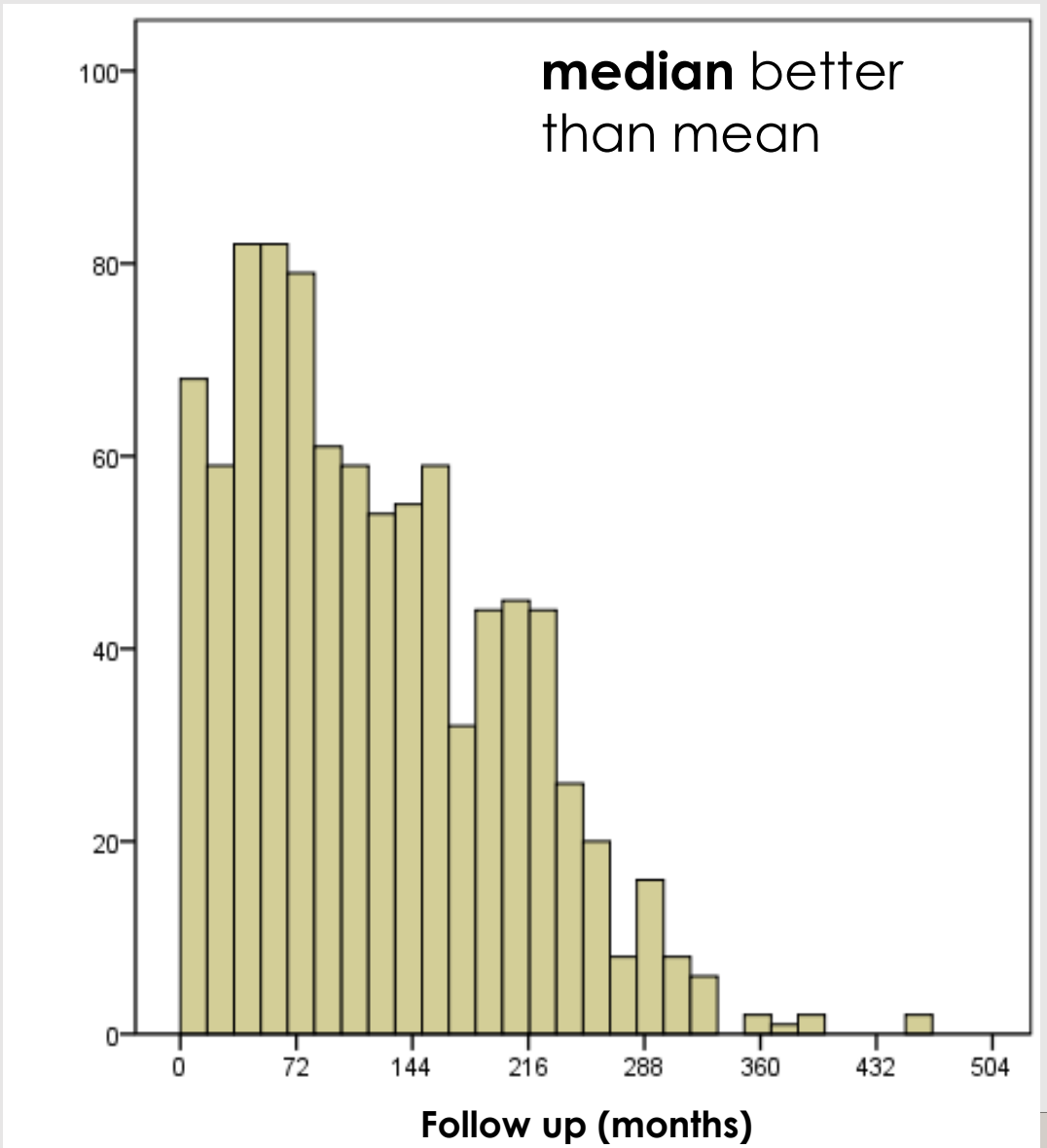
Survival times are in general not **symmetrically** distributed

Survival times are **censored**:

Survival time of an individual is said to be ***censored*** when the end-point of interest **has not been observed** for that individual.
The study **ends before** they have outcome or they are lost to follow-up (dropout)

Independent/non informative censoring*:
survival time *t* of an individual **does not depend** on any mechanism that causes that individual's survival time to be censored at time *c*(< *t*) :

$$T \perp C$$



*If they are not independent, then specialized techniques must be invoked…

Independent censoring:

This means that individuals censored **at any given time** *t* should not be a **biased** sample of those who are still at risk at time *t*.

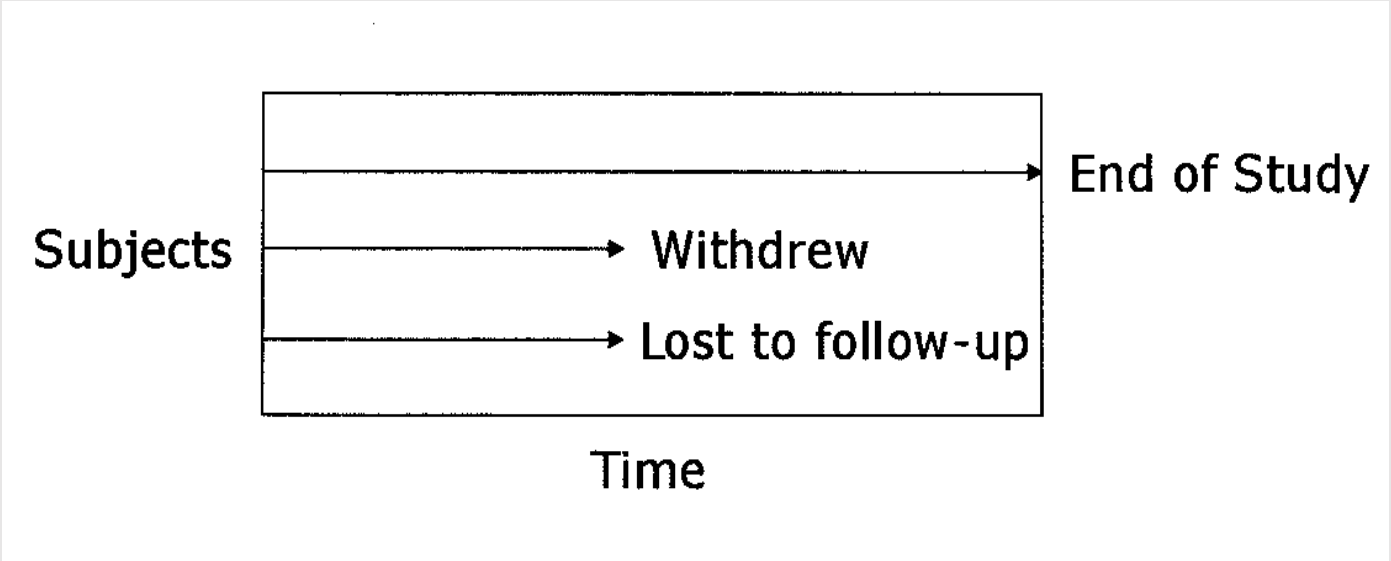
Typically, independent censoring ***cannot be tested*** from the available data - it is a matter of discussion.

Censoring caused by being alive at the administrative closure of the study can usually safely be taken to be *independent*.

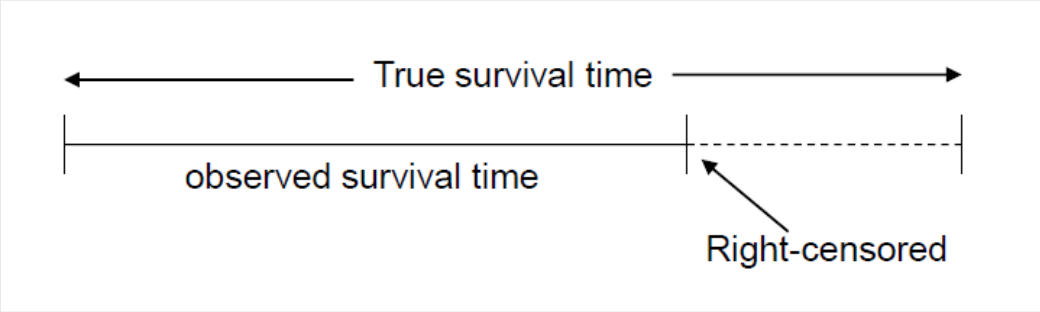
However, one should be more suspicious to other kinds of loss to follow-up **before the end of the study**.

It is strongly advisable (if prospective data collection) to **keep track** of subjects who are lost to follow-up and to note the **reasons** (e.g., worsening of the disease ? emigration ? …).

Right-censoring



The right-censored survival time is less than the «actual» (not observed)



If right censoring occurs when the observation period ends is defined as ***administrative*** censoring.

Aims of Survival Analysis

- **Estimate time-to-event for a group of individuals**, such as time until hospitalization or death 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 CV patients?

Some key quantities

T positive random variable representing *time to event of interest*

$F(t) = P(T \leq t)$ Cumulative Distribution function $f(t) = F'(t)$ *probability density* function

$$S(t) = P(T > t) = 1 - F(t)$$ Survival function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$ Hazard function $H(t) = \int_0^t h(u)du$ Cumulative hazard function

$$h(t) = -\frac{d}{dt}(logS) \qquad S(t) = e^{-H(t)} \qquad f(t) = h(t)S(t)$$

$$S(0) = 1$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

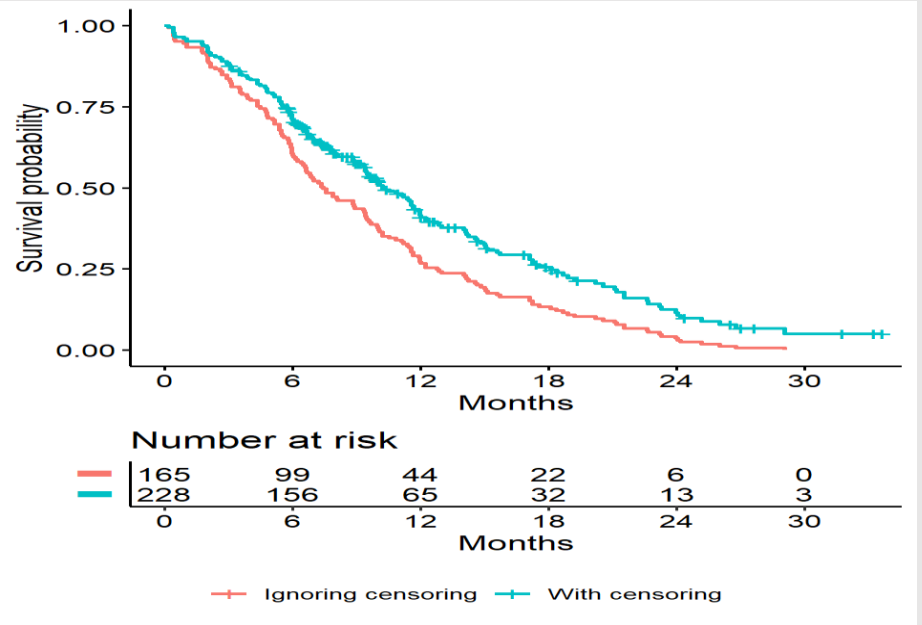
How do we estimate the survival function?

If we observe subjects **all for the same period of time** (and there is **no censoring before the end**):

$$\hat{S}(t) = \frac{\#\ subjects\ survived > t}{\#\ subjects}$$

Or: $$\hat{S}(t) = 1 - \hat{F}(t) = 1 - P(T \le t) = 1 - \frac{\#\,\text{subjects died} \le t}{\#\,\text{subjects}}$$

What happens instead if we have ***censored*** observations and simply we exclude them?



Intuitive explanation: **removing** pts that are censored, creates an **artificially lowered survival curve** because the follow-up time that censored patients contribute is excluded (orange line).

(The **correct estimation of survival** curve for these data is shown in blue for comparison)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

To take correctly into account censoring and variable observation times, there are commonly **3** methods for estimating a survival function [without resorting to parametric models*]:

(1) Kaplan-Meier : most used in clinical/epidemiological setting

(2) Nelson-Aalen : generalizable to *competing risks* setting

(3) **Life-table** (Actuarial Estimator): used in actuarial applications

We will consider the first two.

*parametric survival models are a more advanced topic

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Example:

Follow up (years)	At risk	Censored	Deaths	Survivors
1	100	3	5	?
2	?	3	10	?
3	?	3	15	?
4	?	3	20	?
5	?	3	25	?

* (The censoring mechanism is considered a random variable)

Follow up (years)	At risk	Censored	Deaths	Survivors
1	100	3	5	95
2	92	3	10	82
3	79	3	15	64
4	61	3	20	41
5	38	3	25	13

Censored pts in the *j*-interval are counted in that interval and then **removed** from the risk set.

"These conventions may be paraphrased by saying that deaths recorded as of [time] *t* are treated as if they occurred *slightly before t,* and losses recorded as of [time] *t* are treated as occurring *slightly after t.* In this way the fudging is kept conceptual, systematic, and automatic." (Kaplan & Meier, 1958)

***Kaplan, EL, Meier, P: Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.* 1958.**

Kaplan-Meier estimate* for each time *t* is a conditional probability:

Follow up (years)	At risk	Censored	Deaths	Survivors	KM S(t)
1	100	3	5	95	(95/100)=0.95
2	92	3	10	82	(95/100)x(82/92)= 0.8467
3	79	3	15	64	(95/100)x(82/92)x (64/79)=0.70
4	61	3	20	41	(95/100)x(82/92)x (64/79)x(41/61)= 0.4611
5	38	3	25	13	(95/100)x(82/92)x (64/79)x(41/61)x (13/38)=0.1577

*Non-parametric : no need to assume a specific probability distribution for survival times

Conditional Probability & Independence

The ***conditional probability*** of an event *B given A* is the probability that the event will occur given the knowledge that *A* has already occurred.

The probability that a patient survives after 3 days from the study entry is conditional at having survived the first two days.

p1 = probability to survive the first day
p2 = probability to survive the second day
p3 = probability to survive the third day

Cumulative survival probability will be:

$$P=p1*p2*p3$$

(assumption of independence)

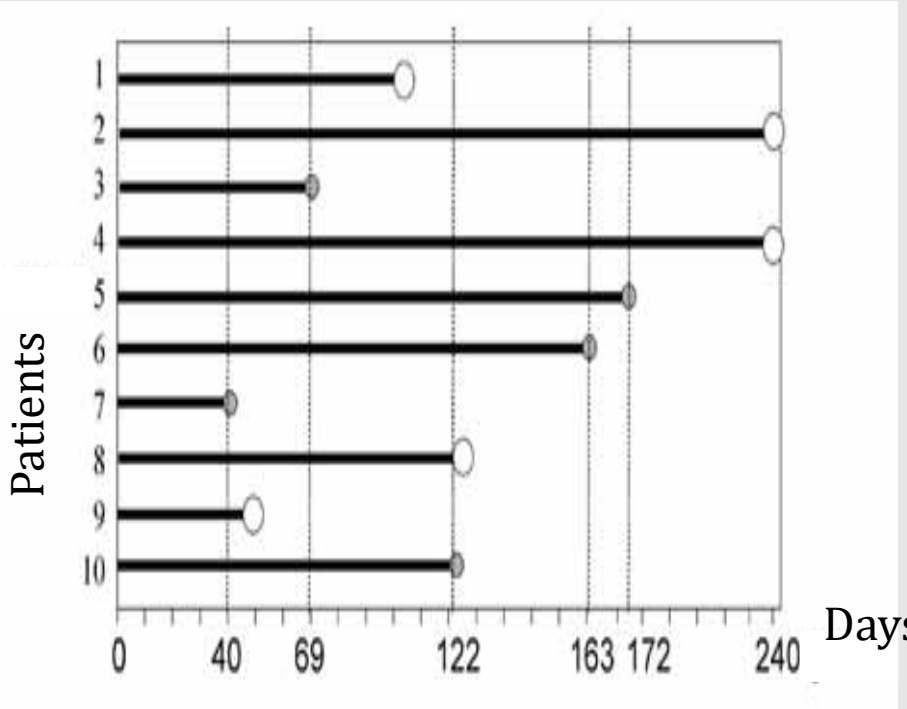
p1=0.80 ; p2=0.68; p3=0.55 → P=0.80 * 0.68 * 0.55=0.30

Kaplan-Meier curve

1. Follow-up is divided into intervals **based on the *observed event times***

2. **Censored** pts in the j-interval are counted in that interval and then removed from the risk set

3. Deaths of the individuals in the sample are assumed to occur *independently* of one another



censored

event

Time intervals	# Risk	# Event	# Censored
0-40	10	1	0
41-69	9	1	1
70-122	7	1	1
123-163	5	1	1
164-172	3	1	0
173-240	2	0	2

Kaplan-Meier *curve*

Events are assumed to occur ***independently*** from one another.

Product-limit estimates of survival

We can multiply probabilities

Time intervals	#Risk	#Event
0-40	10	1
41-69	9	1
70-122	7	1
123-163	5	1
164-172	3	1
173-240	2	0

Survival in interval
9/10: 0.900
8/9: 0.890
6/7: 0.857
4/5: 0.800
2/3: 0.666
2/2: 1.000

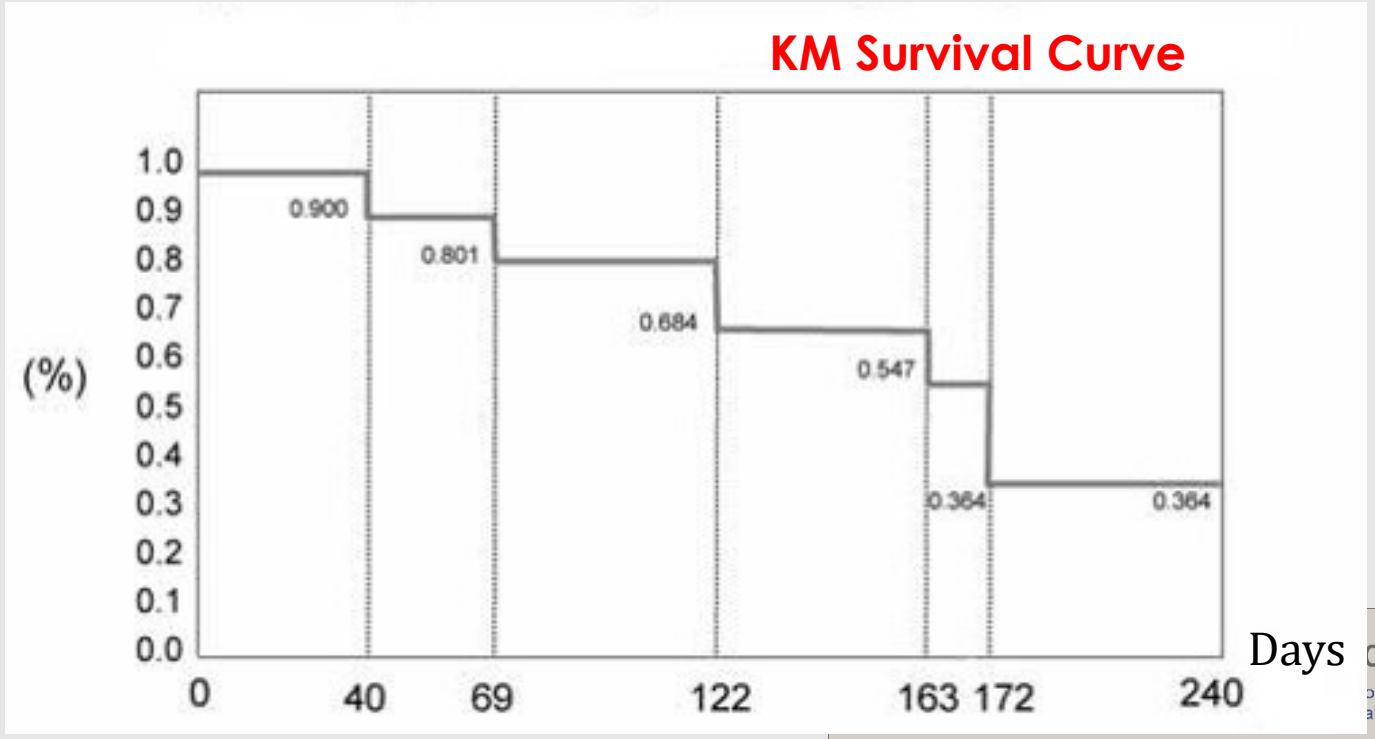
Cumulative Survival	
0.900	S(t1)=9/10=0.90
0.801	S(t2)= 9/10 *8/9=0.90*0.89=0.80
0.684	S(t3)=9/10*8/9*6/7=0.90*0.89*0.85=0.68
0.547	….
0.364	….
0.364	….

In each interval:

$$S_{KM}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

n_j #pts alive at the beginning of j-interval

d_j #events in the j-interval

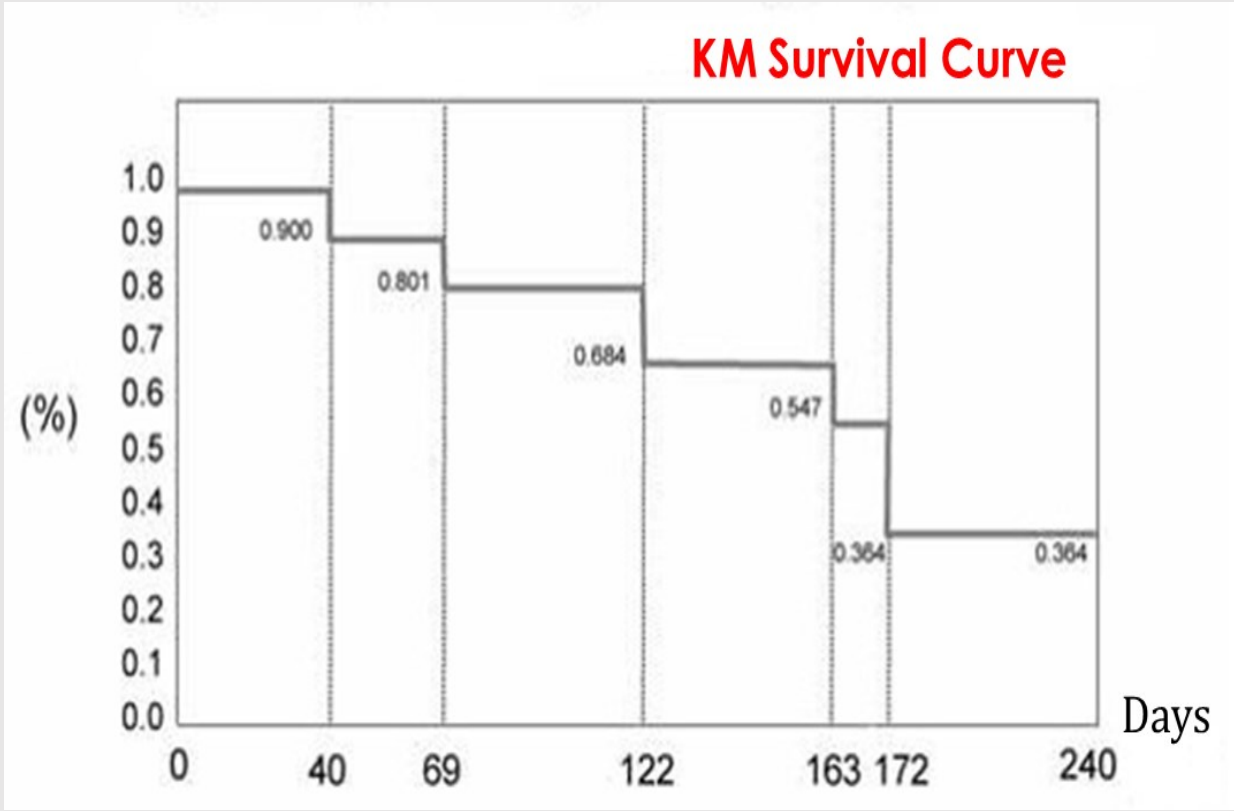


Product-limit estimated survival probabilities are widely used for analyzing survival data because they are estimated **without assumptions** about the probability mechanism that produced the sample of survival times (called **nonparametric** or **model-free** estimates).

Graphical presentation is a series of rectangles (height = S_{i-1} and width = $t_i - t_{i-1}$) placed side by side to display the decreasing pattern of the estimated survival probabilities over time.

It is commonly called a **step** function.

The survival function becomes a "**curve**" in a large sample of distinct survival times where the steps get very small.



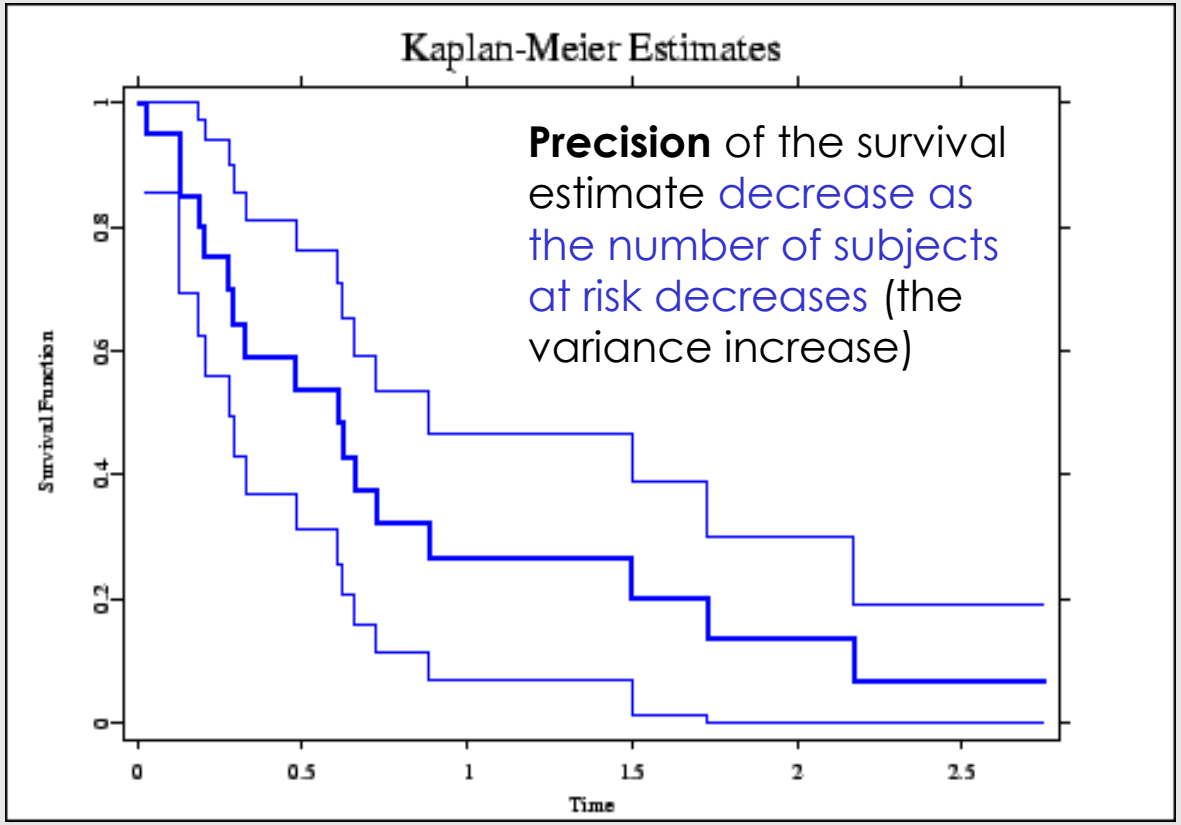
Confidence Intervals

Kaplan-Meier curve

Variance (Greenwood's formula):

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

$$\sqrt{\hat{V}\left(\hat{S}(t)\right)}$$



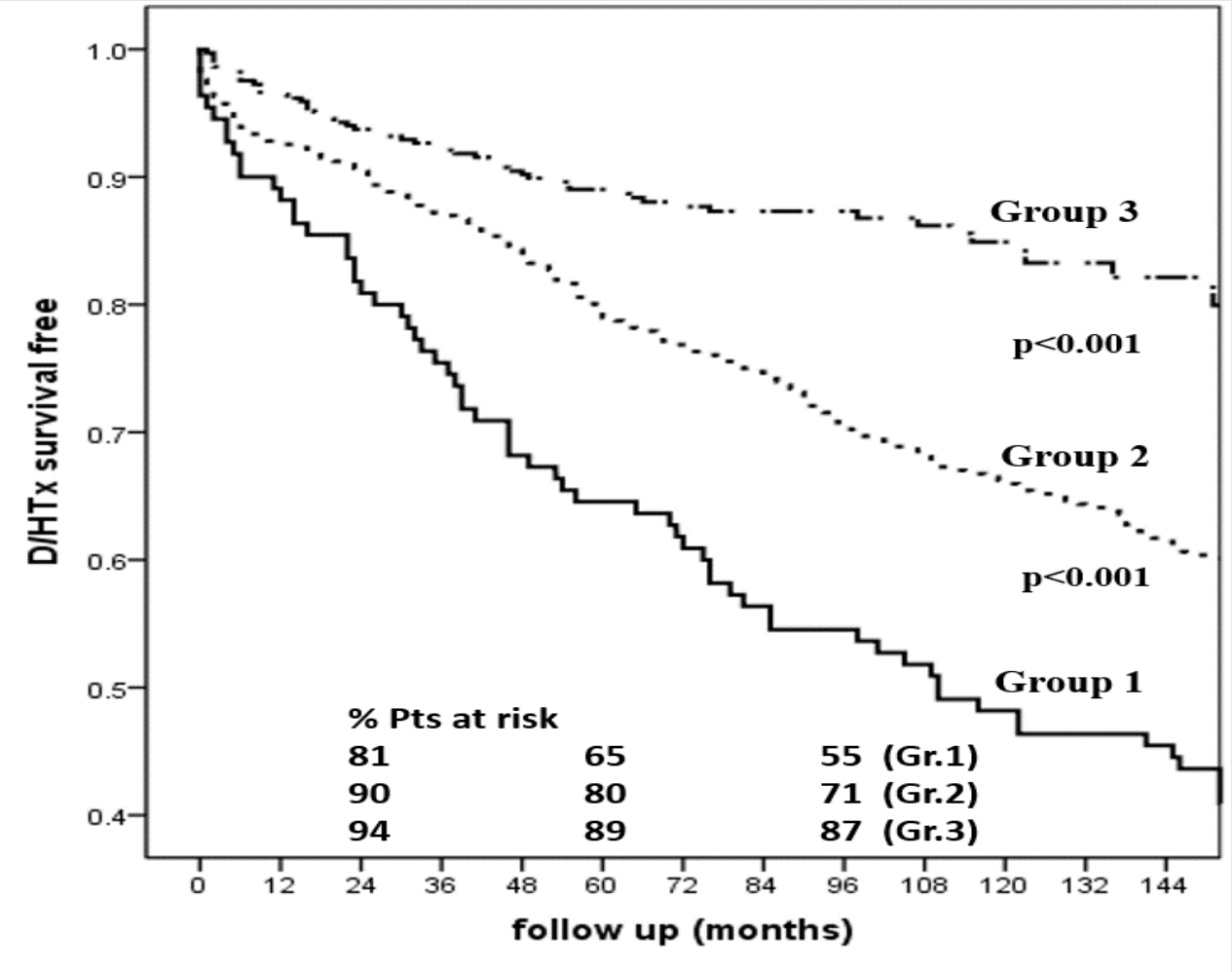
Confidence intervals* can be computed on the scale of the survival curve using the normal approximation:

$$\hat{S}(t) \sim N\left(S(t), V(S(t))\right) \quad \Rightarrow \quad \hat{S}(t) \pm 1.96\sqrt{\hat{V}\left(\hat{S}(t)\right)}$$

*At each time point *t*

Kaplan-Meier curve

From 1978 to 2007, 853 IDCM patients (45 ± 15 years, 72% males) were enrolled and classified as follows: Group 1, 110 patients (12.8%) enrolled during 1978–1987; Group 2, 376 patients (44.1%) enrolled during 1988–1997; Group 3, 367 patients (43.1%) enrolled during 1998–2007.



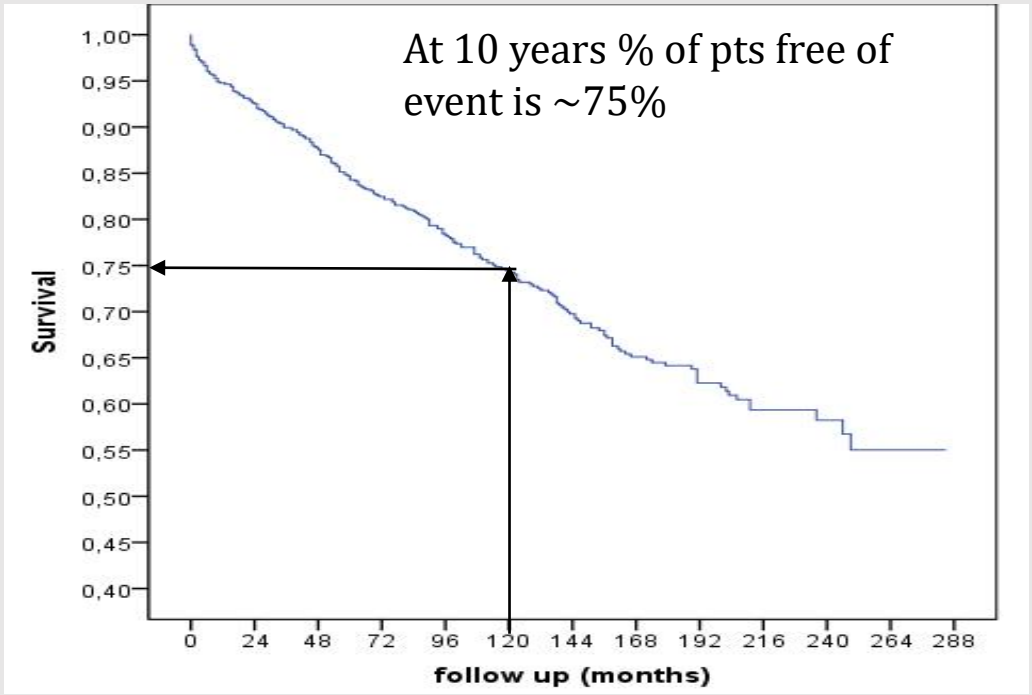
The aim of the study was to describe the impact of therapeutic approaches in the last 30 years on the long-term natural history of a large cohort of IDCM patients enrolled in the clinical registry in Trieste.

KM curves with decade of enrollment as a *proxy* of changing therapies.

Long-term prognostic impact of therapeutic strategies in patients with idiopathic dilated cardiomyopathy: changing mortality over the last 30 years. Merlo et al., Eur J Heart Fail. 2014.

Survival function

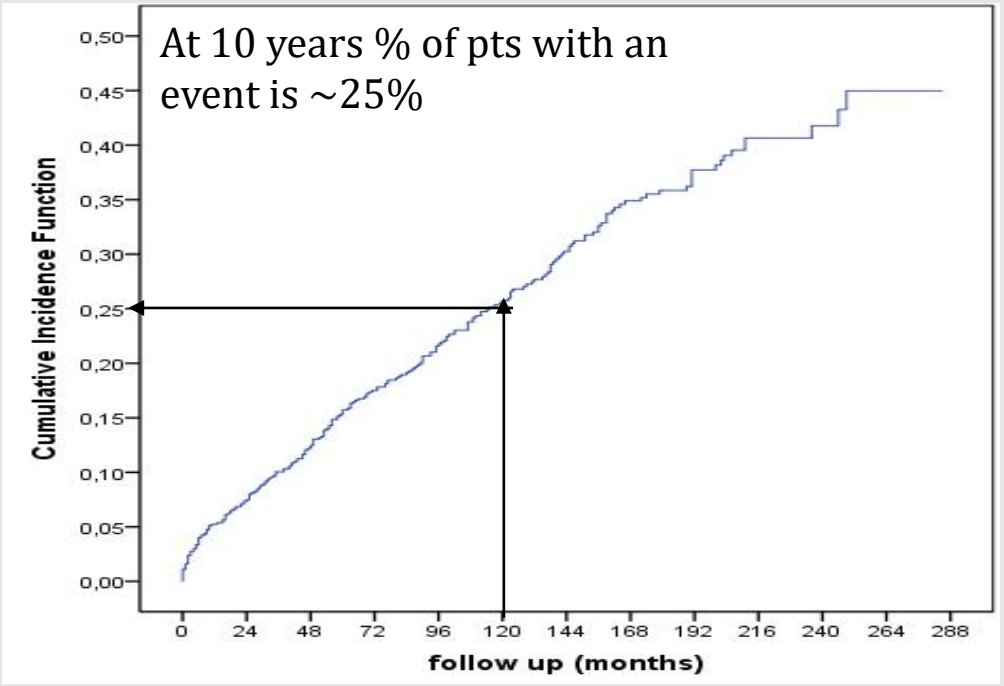
$$S(t) = P(T > t) = 1 - F(t)$$



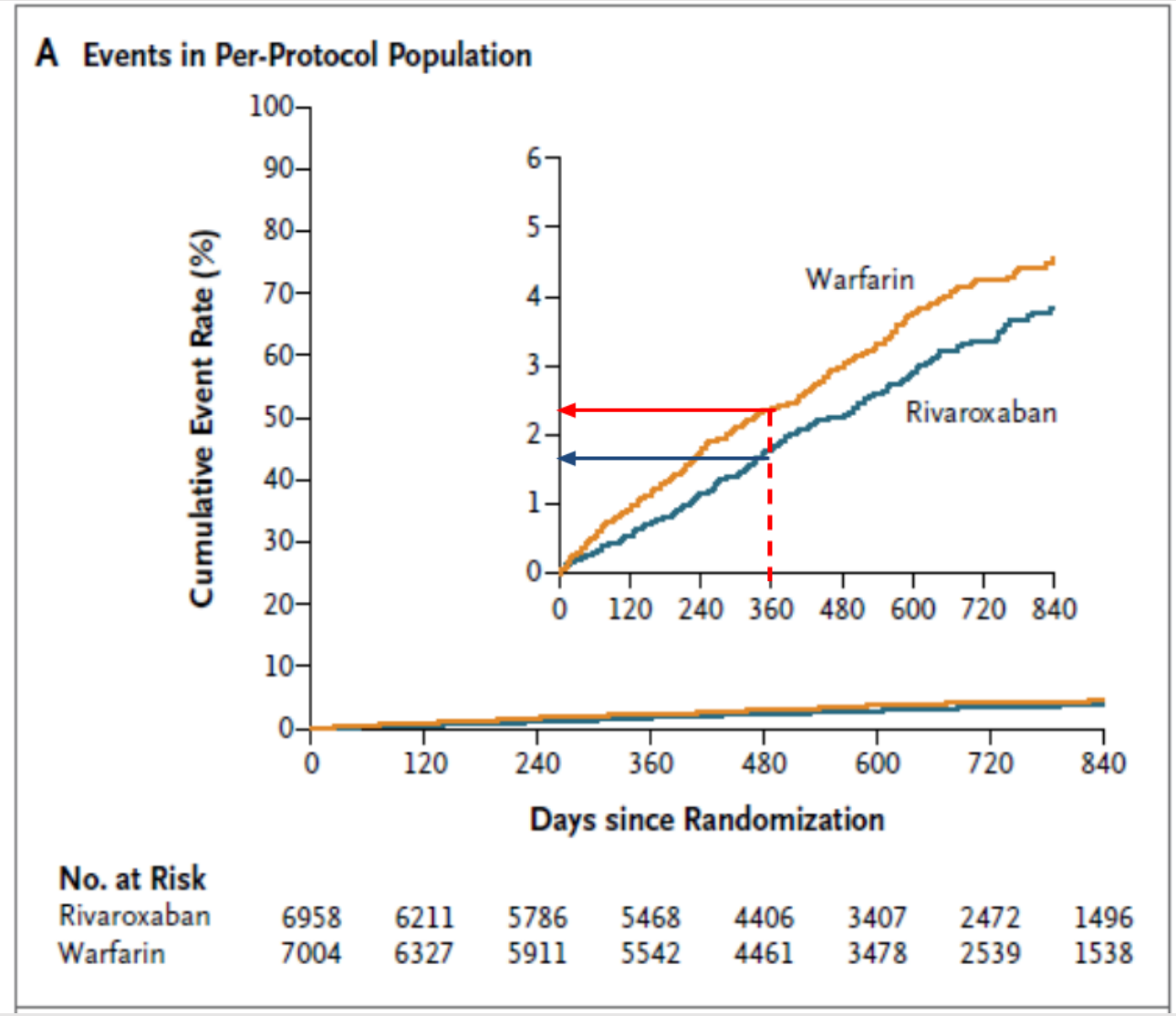
Probability that survival is «beyond» *t*

Cumulative Incidence Function (Cumulative Event Rate):

$$F(t) = P(T \leq t)$$



Probability that survival is «less than» *t*



When the event rate is **low** could be more informative plot the cumulative event rate instead of the survival function.

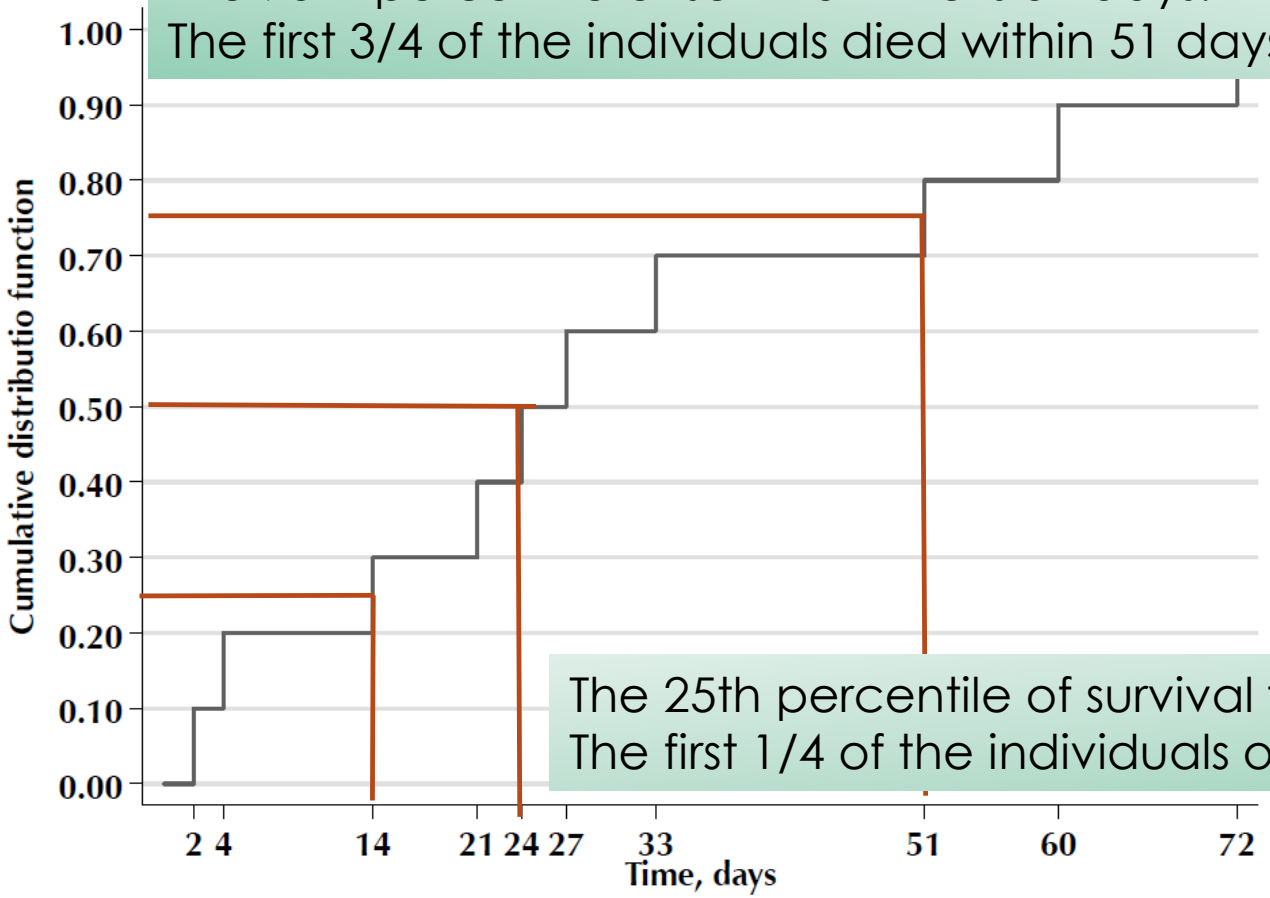
Median survival time

The generic pth quantile of the survival time is the *smallest* time so that : $S(t_p) \leq 1 - p$

If $F(t_p) \geq p$ then the 100 pth percentile of survival time is t_p

What is the time by which a certain probability of dying is achieved?

The 75th percentile of survival time is 51 days.
The first 3/4 of the individuals died within 51 days.



S(T > 24 days) = 0.5

F(T ≤ 24 days) = 0.5

50th survival percentile = 24 days

Half of the individuals died within 24 days.

Half of the individuals lived longer than 24 days.

The 25th percentile of survival time is 14 days.
The first 1/4 of the individuals of the sample died within 14 days.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Rates

What is the ***change*** of the survival probability over time?

Rate = change in the survival function S(t) for a change in time

It is related to the **derivative** of the survival function with respect to time.

For example, what is the change in survival probability during the first 2 days of observation ?

S(t=0 days) = 1
S(t=2 days) = 0.9

→ (0.9-1) / (2-0) = -0.1/2 = -0.05

Between 0 and 2 days, the survival function decreases by 0.05% (of the initial survival) per day.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

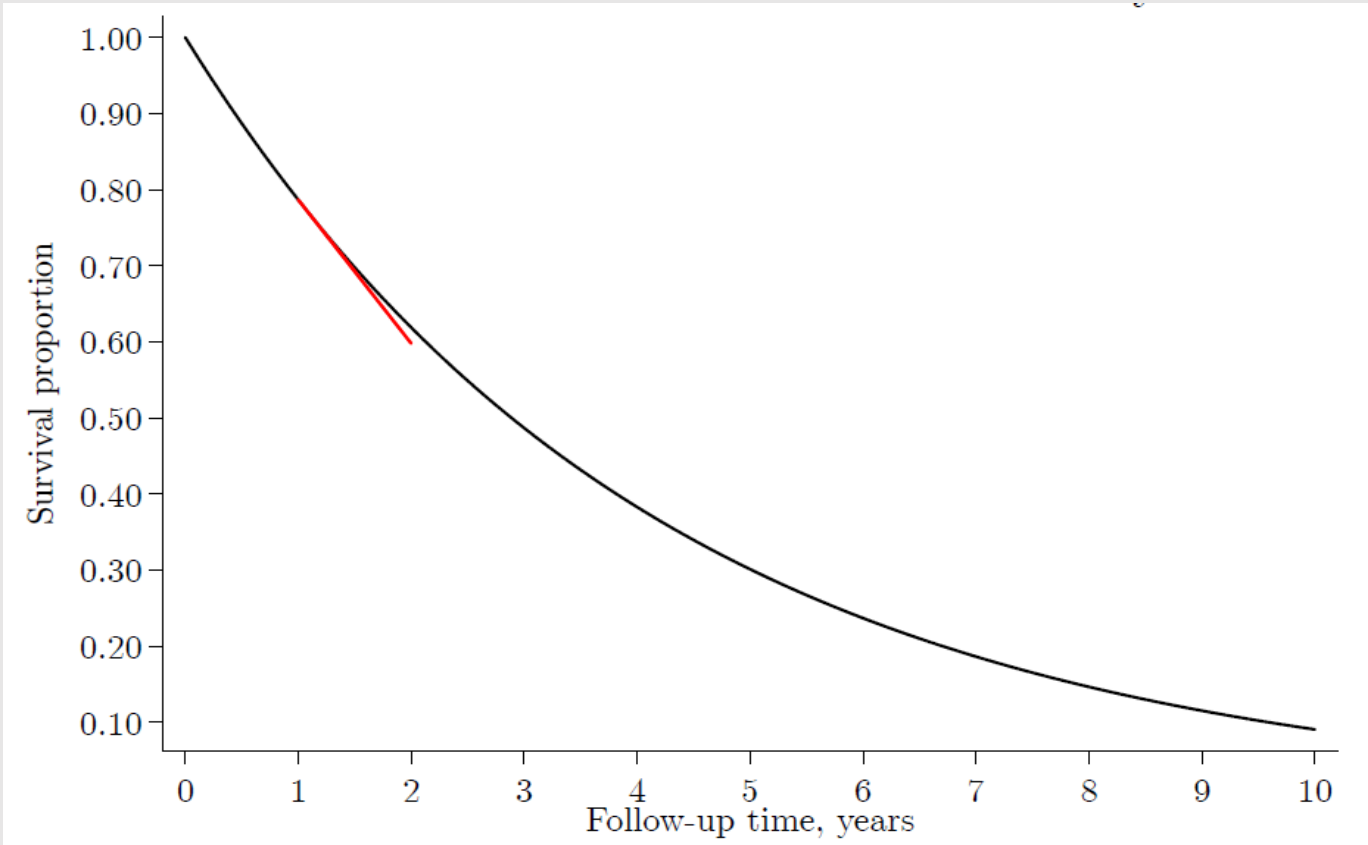
Hazard rate

What is the **instantaneous** rate of death at some point in time *t* relative to the survival probability at *t* ?

The hazard at time t is obtained by: $h(t) = -\frac{\frac{d}{dt}S(t)}{S(t)}$ $h(t) = -\frac{d}{dt}(logS)$

The *instantaneous* relative rate h(t) is usually called a **hazard rate** in human populations and a **failure rate** in other contexts.

The same rate is sometimes called the **force of mortality** or an instantaneous rate of death.

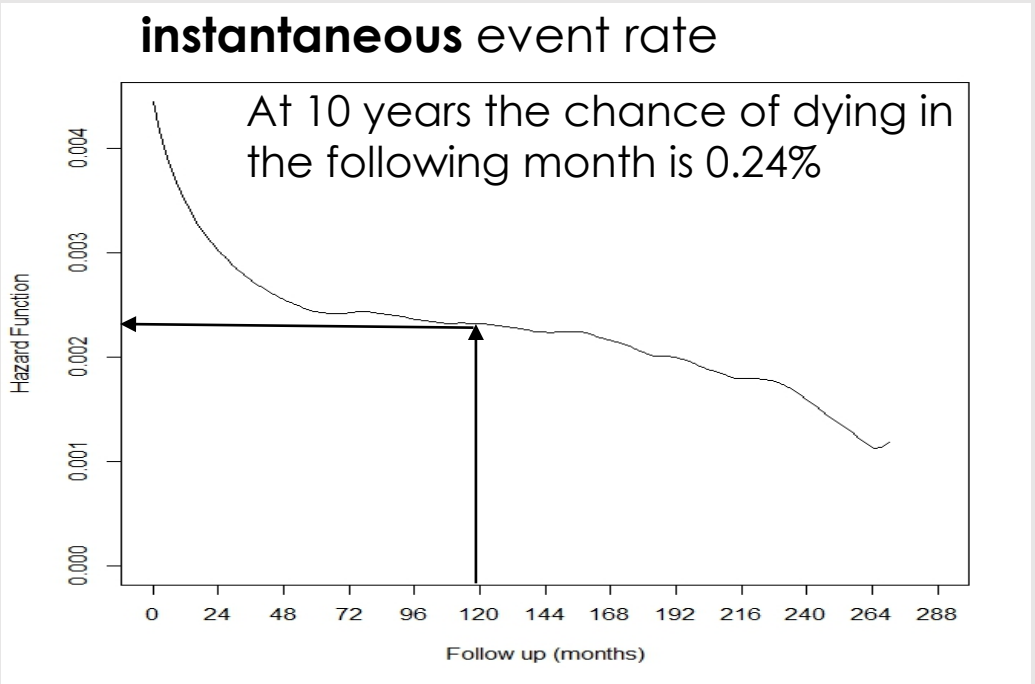


Hazard & Cumulative Hazard Rate

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

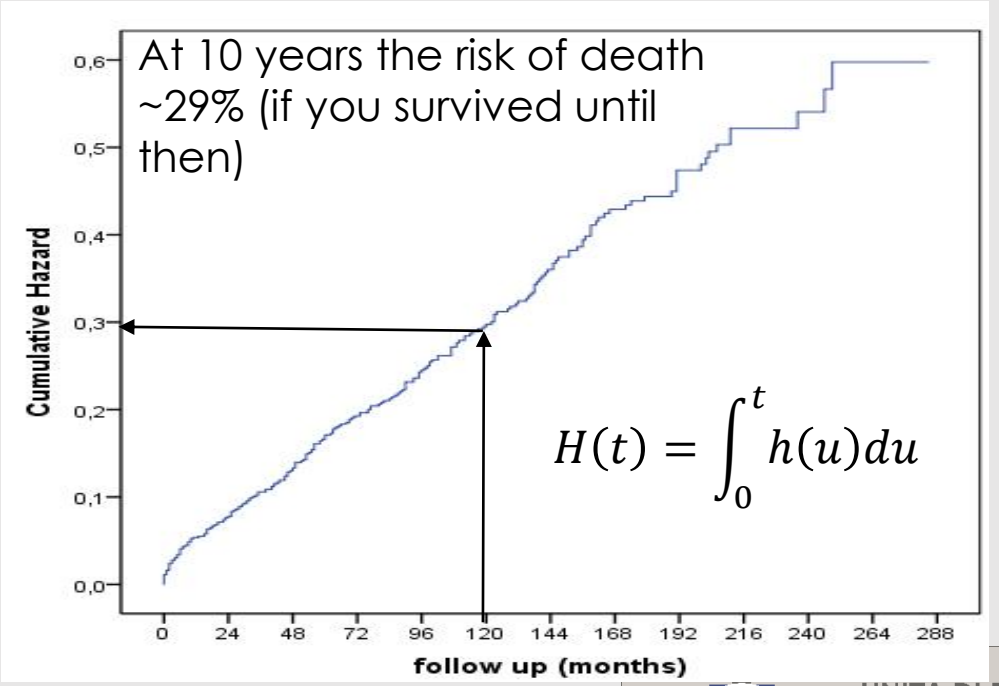
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

instantaneous event rate



Risk of event **up to time** *t* given that the event has not occurred before *t*

Cumulative Hazard Function
(Cumulative Hazard Rate)



UNITA DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

How do we estimate the cumulative hazard function?

$$H(t) = \int_0^t h(u)du \quad \Rightarrow \quad S(t) = e^{-H(t)}$$

The most common estimate of the cumulative hazard is the **Nelson-Aalen** estimate

Just as we did for the KM, we divide follow-up in ***intervals*** corresponding to event times, in this way $H(t)$ can be approximated by a sum:

$$H(t) \approx \sum_{t_j \leq t} \tilde{h}_j * \Delta_j$$

$\tilde{h}_j$: hazard in the j-th interval and Δ_j is the interval width

$\tilde{h}_j * \Delta_j$ is the **conditional probability of the event in the interval**, estimated by $\frac{d_j}{n_j}$: number of events in the interval over the set of subjects at risk at the beginning of the interval

Cumulative Hazard *function*

The Nelson-Aalen estimate of the cumulative hazard is:

$$H_{NA}(t) = \sum_{t_j \le t} \frac{d_j}{n_j}$$

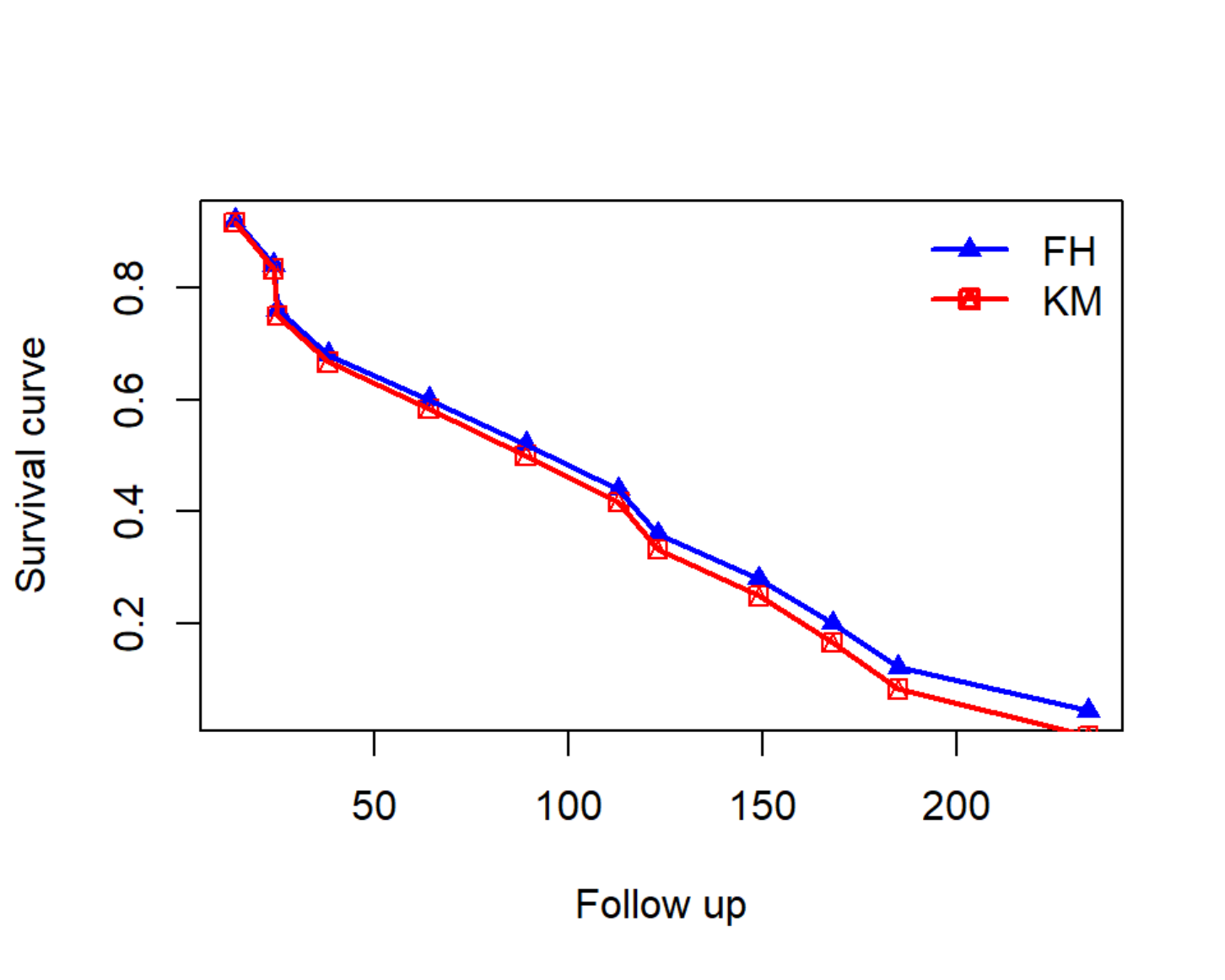
$H_{NA}(t)$, like the KM, changes only at the observed event times.

Once we have $H_{NA}(t)$, we can obtain the Fleming-Harrington estimator of S(t): $\hat{S}_{FH}(t) = e^{-H_{NA}(t)}$

In general, the FH estimator of the survival function **is very close** to the Kaplan-Meier estimator.

The two estimators are similar when the increments $\frac{d_j}{n_j}$ are small, that is, when there are many subjects still at risk.

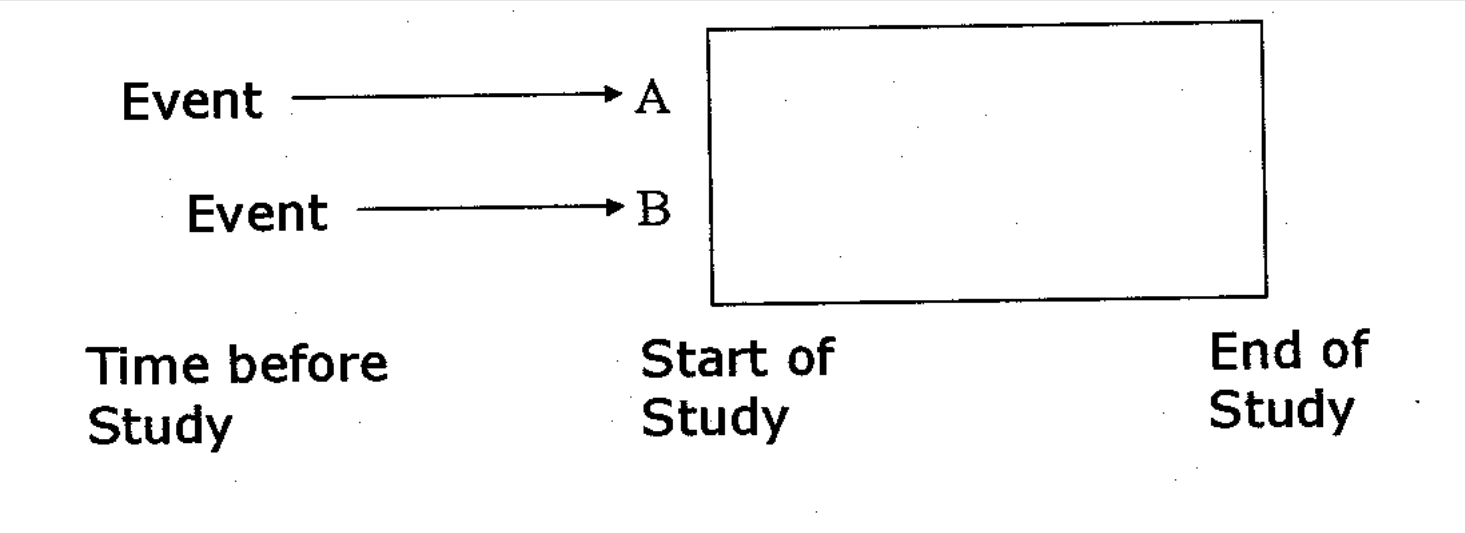
Generally, the Fleming-Harrington estimator is slightly higher than the KM at every time point, but with larger datasets the two will typically be much closer.



The Nelson-Aalen estimator is generalisable to situations with ***competing risks*** and ***multiple states*** (advanced topic!).

Supplementary materials

Left-censoring



The actual survival time [non fatal event] is less* than that observed.

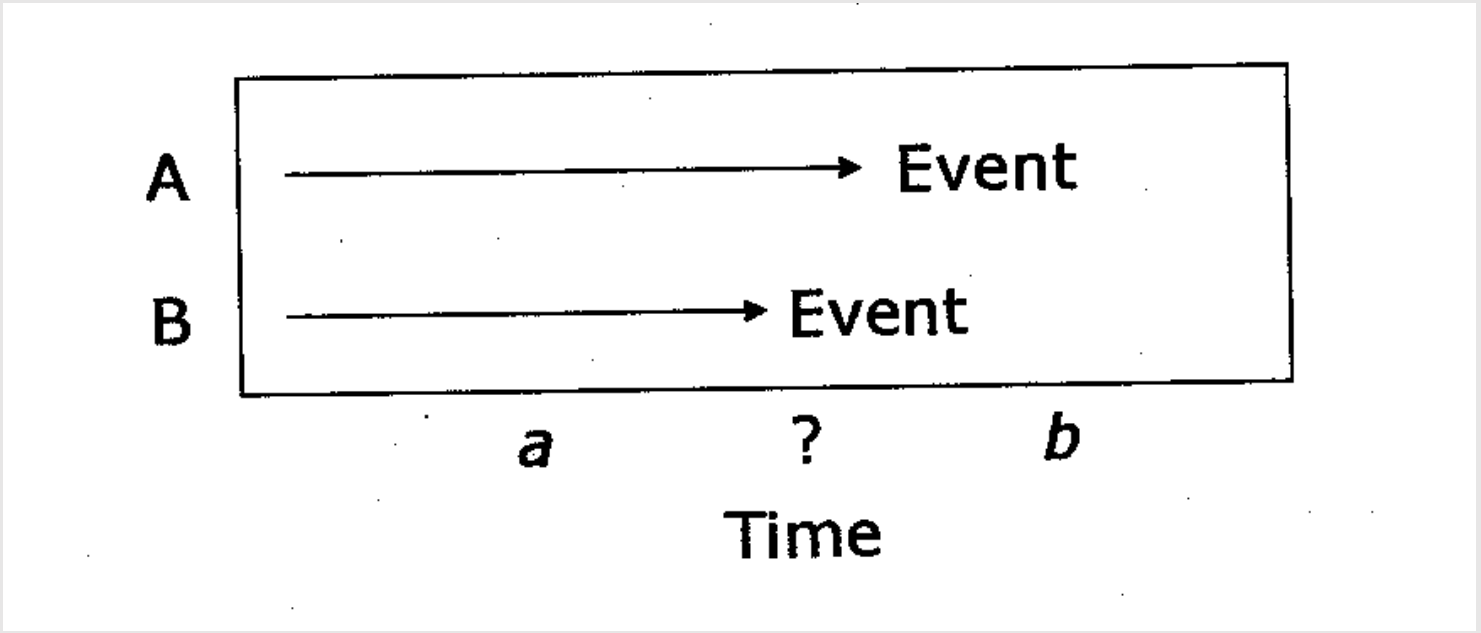
Outcome: time to recurrence of cancer after a surgical intervention.

Start of the study: 3 months after operation patients are examined to determine recurrence.

Some of them may be found to have recurrence.

* If left censoring occurs in a study, right censoring may also occur, and the lifetimes are considered ***doubly censored***

Interval-censoring



Subjects are known to have experienced an event ***within an interval*** of time.

Outcome: time to recurrence of cancer after a surgical intervention.

Start of the study: 3 months after operation patients are examined to determine recurrence.

A patient is **free from recurrence** at 3 months and then is found to have recurrence at 6 months.

Observations from most studies with a nonlethal outcome are *interval censored* since we usually cannot monitor subjects continuously.

The issue is whether we should analyze the data as *interval censored* or *point censored*.

For instance, if the ***median survival time*** is 5 years and the intervals are between 3 and 6 months wide, then we have no reason to complicate the analysis by considering interval censoring.

On the other hand, if the intervals are about 1 year or longer, then we should account for such uncertainty in the analysis…

CENSORING ISSUES IN SURVIVAL ANALYSIS

Leung KM, Elashoff RM, Afifi AA

Annu. Rev. Public Health 1997. 18:83–104

A special case (by design): *truncation*

In some epidemiological studies a specific ***initial event*** prior to the event of interest is defined as the *time origin* (e.g : a specific diagnosis) but subjects are recruited only in a pre-defined *calendar* window.

For some subjects a «***delayed*** entry time» is observed, since we only observe those subjects after the time origin (unknown), until they have the event of interest or are censored.

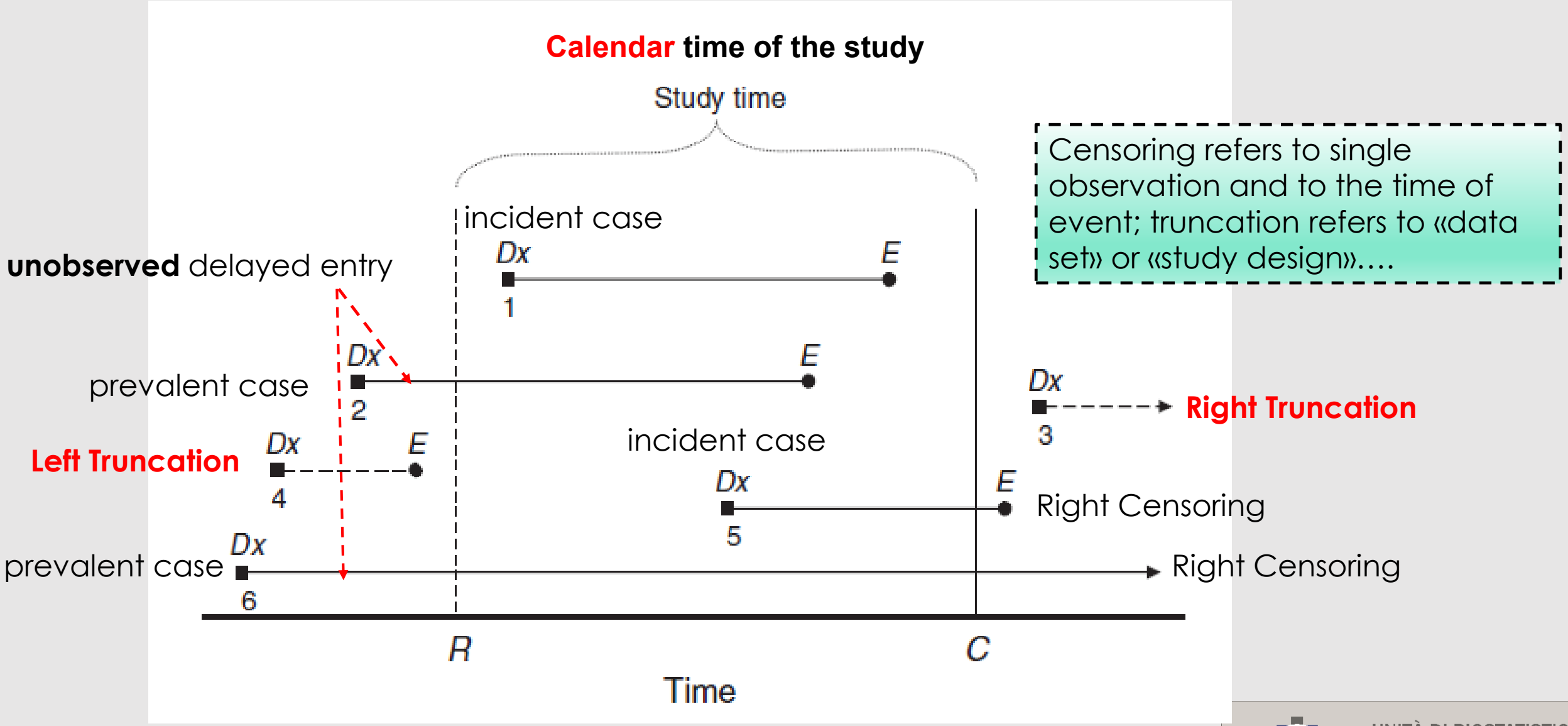
Others subjects are «**truncated**» since they experience both the initial event and the event of interest out of the observation window.

Different from *left censoring*: there the **event of interest** has already occurred for the individual before he/she is observed in the study…it is not a question related to a specific **time origin**.

In our examples, we will focus only on methods that account for **right censoring**, other mechanisms will not be covered, but always pay a great attention to the **study design** !

A special case (by design): truncation

Study recruitment starts at *R* and ends at C. **Date of diagnosis=Dx** and **event=E**



Censoring refers to single observation and to the time of event; truncation refers to «data set» or «study design»….

«Mean» survival time

The *average* length of time from the start of observation that patients are still alive.

Mean survival time is estimated as the area under the survival curve. The estimator is based upon the entire range of data. Rarely used in clinical studies.

For "complete" survival data (no censoring): $\hat{\mu} = \frac{\sum t_i}{n}$

For censored data is the area under the estimated survival function*: $\hat{\mu} = \sum_{i=1}^{n} S_{i-1} * (t_i - t_{i-1})$

The sum of the *n* rectangles' area is the total area enclosed by the estimated product-limit survival function.

Samples of survival times are frequently **highly skewed**, therefore, in survival analysis, the **median** is generally a better measure of central location than the mean.

* A large sample method is used to estimate the variance of the mean survival time and thus to construct a confidence interval