Statistical learning in epidemiology

Lecture IV

Giulia Zamagni

Clinical Epidemiology and Public Health Research Unit, IRCCS «Burlo Garofolo»

University of Trieste

Survival analysis

- **Key idea**: in many studies, we are interested not only in ***whether*** an event occurs, but also ***when*** it happens
- There are many examples in epidemiology:

➢ Death after a diagnosis

➢ Disease relapse after a treatment

➢ Time to hospital discharge

In these situations, the outcome of interest is not simply the occurrence of the event. Instead, the main focus is the **time elapsed until the event happens**.

→ For this reason, these outcomes are referred to as **time-to-event data**.

Survival analysis

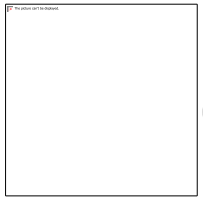
Each individual has:

1. **An event indicator**
2. **A follow-up time**

However, things are always more *complex* in practice.

A major complication arising when working with survival data is that **not all subjects may experience the event during the study period**, and in particular:

- Patients may **drop out**: a patient enrolled in a cancer treatment study decides to stop participating because of side effects from the therapy.
- Some patients may be **lost to follow-up**: a participant changes address and phone number, and researchers are no longer able to contact them for future visits.
- The **study may end earlier**: a 5-year cardiovascular study finishes after 3 years because funding ends, so some participants have not yet experienced the event of interest.



Censoring

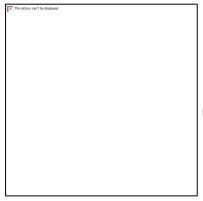
Censoring is a key concept in survival analysis

It occurs when the exact time of the event of interest is not fully observed for some individuals.

Therefore, we know that the event has **not occurred up to a certain time**, but we do not know what happens **afterward**.

However, censored individuals still provide useful informations, as they contribute data for the period during which they were observed.

→ Survival analysis methods are specifically designed to incorporate this partial information correctly.



Censoring: main types

1. **Right censoring**: the most common type of censoring.

   It occurs when ***we know that the event has not happened up to the last observation time, but we do not know if or when it happens afterward***.

   Example: a patient is followed for 3 years after a cancer diagnosis. At the end of the study, the patient is still alive. We do not know the actual time of death, but we know that the patient survived at least 3 years.

2. **Left censoring**: it occurs when ***the event has already happened before observation begins, but the exact time of the event is unknown.***

   Example: in an HIV study, a participant tests positive at the first medical visit. We do not know when the infection occurred; we only know that it happened before the first test.

3. **Interval censoring:** it occurs when ***the event is known to have happened within a specific time interval, but the exact timing is unknown.***

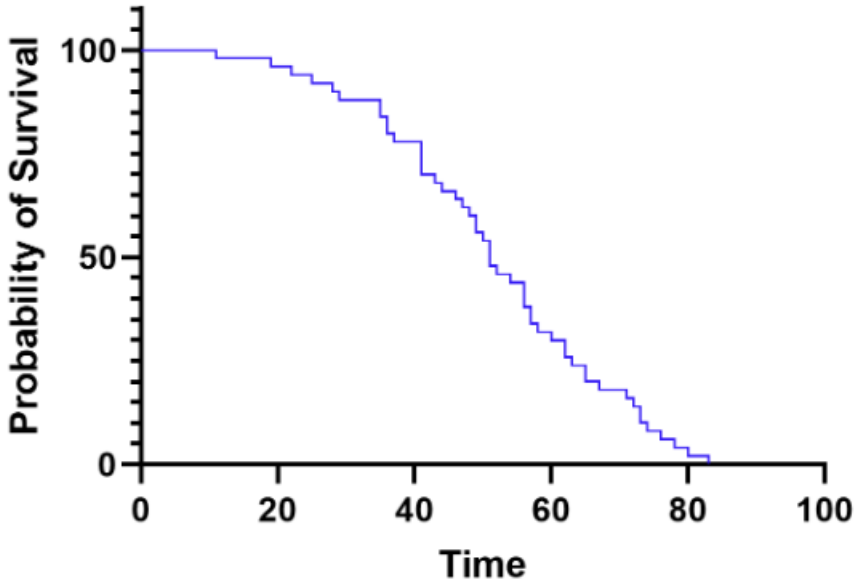
   Example: a patient is examined every 6 months to detect disease relapse. At the January visit there is no relapse, while at the July visit relapse is detected. Therefore, the relapse occurred sometime between January and July, but the exact date is unknown.

Survival curves

- To properly account for both the timing of events and censored individuals, we use survival analysis methods, particularly **survival curves**.
- A survival curve describes the probability of remaining event-free over time. More specifically, it shows **the probability that an individual survives (or does not experience the event) beyond a given time point.**

The curve typically:

1. starts at 1 (or 100%) at time zero,
2. decreases over time as events occur,
3. changes only when an event happens.



The hazard function

- While the survival curve focuses on the probability of surviving over time, another important quantity in survival analysis is the **hazard function**.
- It describes the **instantaneous risk** of experiencing the event at a specific time, among individuals who have not yet experienced the event.

In simpler terms, it answers the question:

"At this exact moment, how likely is the event to occur in individuals who are still event-free?"

The hazard is therefore a measure of **risk over time**.

Aspect	Survival Curve	Hazard Function
Definition	Describes the probability of remaining event-free over time	Describes the instantaneous risk of experiencing the event at a given time
Main focus	Survival probability	Risk intensity
Behavior over time	Starts at 1 and decreases over time	Can increase, decrease, or remain constant
Clinical interpretation	Easier to interpret clinically	Useful for understanding how risk changes during follow-up
Key question answered	"What proportion of patients are still event-free over time?"	"Among patients still at risk, how rapidly are events occurring right now?"
Interpretation of high values	High survival means many individuals remain event-free	High hazard means a high instantaneous risk of the event
Relationship between the two	Declines faster when hazard is high	Determines how quickly the survival curve decreases
Overall role	Summarizes the cumulative effect of risk over time	Describes the underlying event dynamics over time

Suppose we are studying mortality after a cancer diagnosis.

- The **survival curve** tells us the probability that a **patient** is still **alive after 1 year, 3 years, or 5 years.**
- The **hazard function** tells us the **instantaneous risk of death** at each moment **among patients** who are **still alive**.
- Immediately after diagnosis, the **hazard** may be **high**, then **decrease during recovery**, or **increase again later** if the disease progresses.

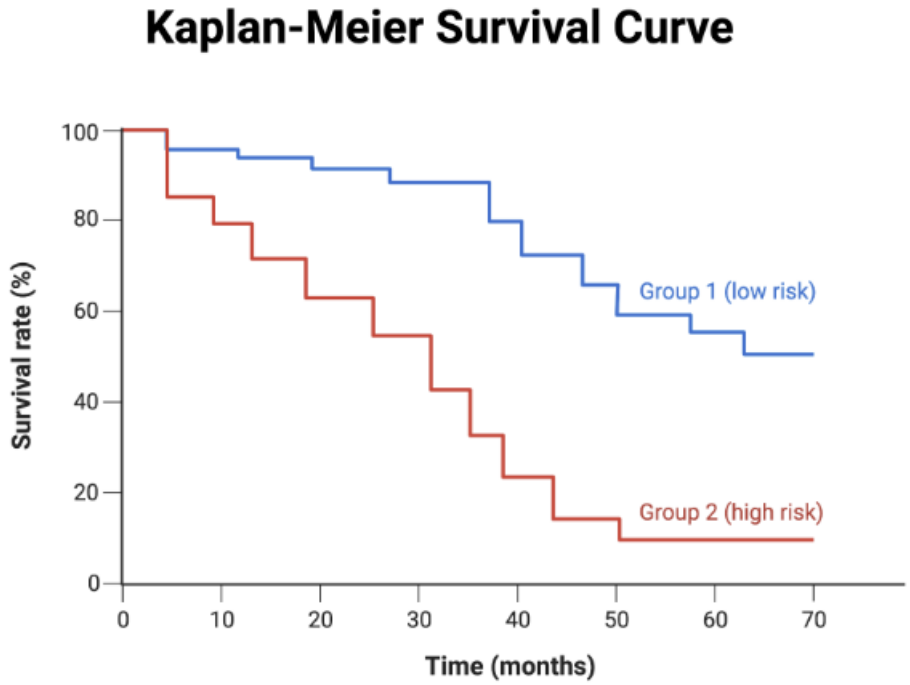
The Kaplan-Meier estimator

- The most commonly used method to estimate a survival curve is the **Kaplan–Meier estimator**
- It is a non-parametric method used to estimate the probability of surviving beyond different time points while properly accounting for censored observations.

The Kaplan–Meier curve is built by updating the survival probability every time an event occurs.

- The curve starts at 1 (100% survival at time zero).
- Each observed event causes a downward step in the curve.
- Censored observations do not cause the curve to drop: they simply reduce the number of individuals still under observation afterward.

For this reason, the Kaplan–Meier curve has a characteristic **stepwise shape**.



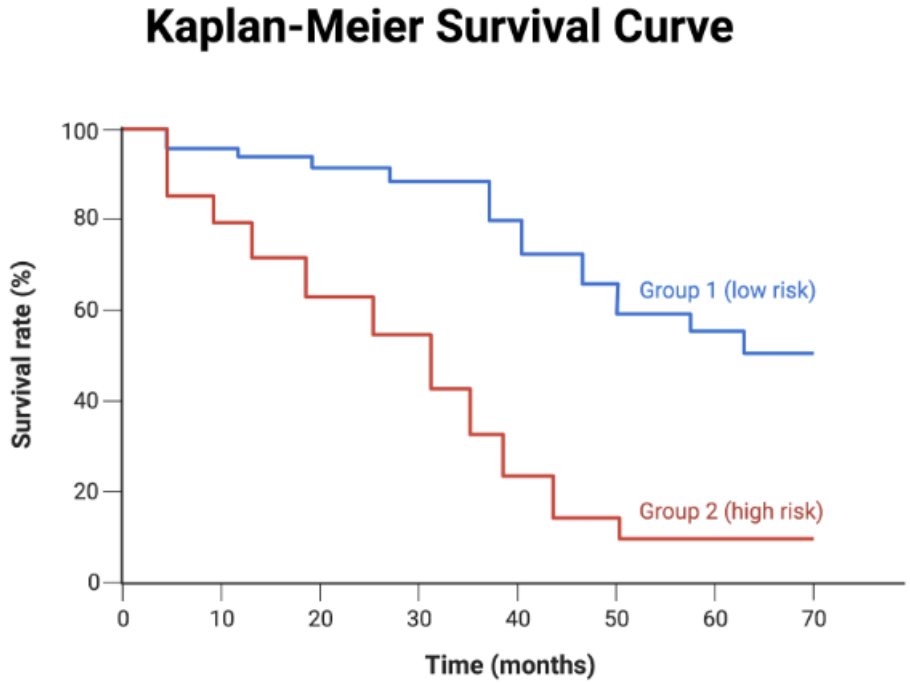
How to compare survival between groups?

Very often, researchers want to compare survival between two or more groups.

Examples:

- *treated vs untreated patients,*
- *smokers vs non-smokers,*
- *different therapeutic strategies.*

In these cases, we can visually compare Kaplan-Meier curves. However, visual differences alone are not sufficient!



- **We need a statistical test to determine whether the observed differences are likely due to chance.**
- The most commonly used test is the **log-rank test**.

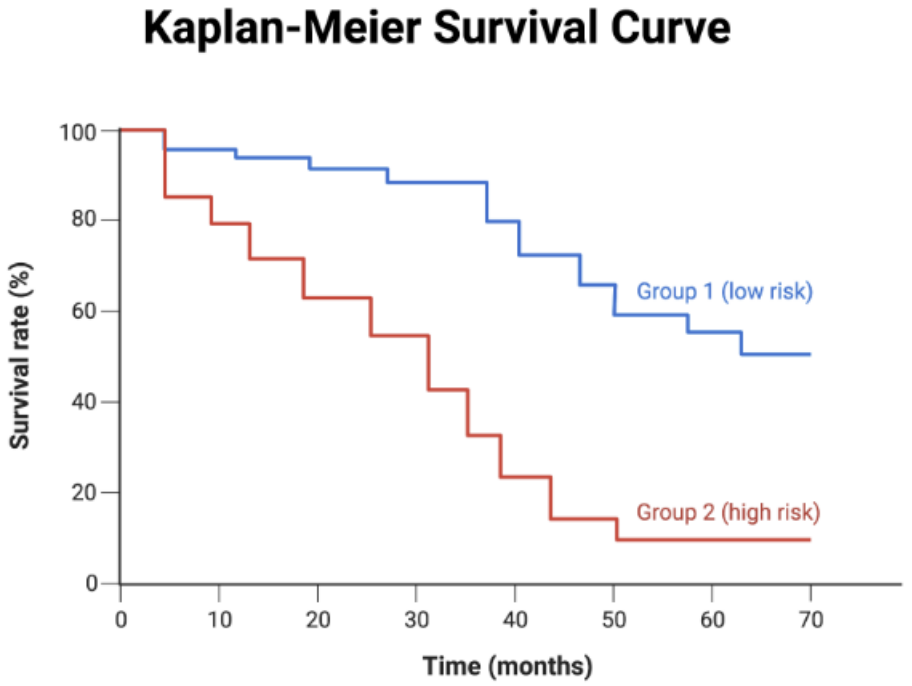
The log rank test

- A non-parametric statistical test used to compare survival curves between groups.
- It evaluates whether there is a significant difference in survival experience over the entire follow-up period.

At each event time, the test compares:

- the **observed number of events** in each group,
- with the **expected number of events** if the groups had the same survival pattern.

If the observed and expected numbers differ substantially over time, the test suggests that the survival curves are significantly different.



Null hypothesis (H_0):
survival is the same in the two groups.
Alternative hypothesis (H_1):
survival differs between groups.

Limitation: the log-rank test **does not quantify the magnitude of the difference or adjust for other variables**.
→ For this, we need more advanced methods, such as the **Cox PH model**

The Cox Proportional Hazards model

- More complex questions, more complex methods to provide answers
- For example:

1. ***Does treatment still affect survival after adjusting for age and sex?***
2. ***Which variables are associated with a higher risk of the event?***
3. ***How large is the effect of a risk factor?***

- The **Cox PH model** is a regression model used in survival analysis to evaluate the association between covariates and the hazard of an event.
- It relates the **hazard function** to one or more explanatory variables (e.g., age, sex, clinical characteristics, biomarkers…)

- Rather than modeling survival probabilities directly, the model focuses on the **hazard**.
- The key output is the **hazard ratio**, which compares the **hazard between two groups** and tells us **how much the event risk changes according to a predictor.**

Hazard Ratio	Interpretation
HR = 1	No difference in hazard
HR > 1	Higher hazard (increased risk)
HR < 1	Lower hazard (protective effect)

The Proportional Hazard assumption

- Assuming PH means that the hazard ratio between groups remains constant over time.
- In other words, **the relative difference in hazard does not change during follow-up.**

Suppose:
treatment group hazard = 0.5 × control hazard.

- If hazards are proportional, the treatment group maintains approximately half the risk throughout the study period.

- The absolute risk may change over time, but the ratio between groups remains stable.

$$h(t|X) = h_0(t)e^{\beta X}$$

Where:

- $h(t|X)$ = hazard at time t for an individual with covariates X,
- $h_0(t)$ = baseline hazard,
- β = regression coefficient.

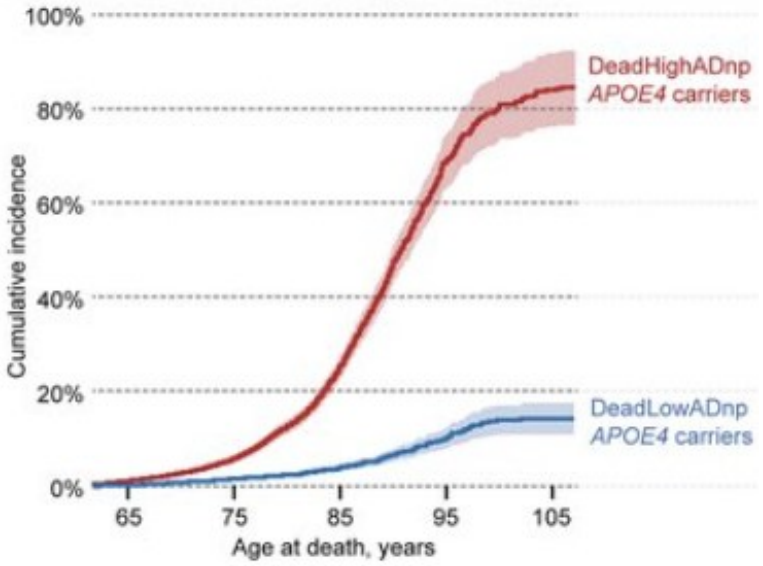
The hazard ratio is obtained from:

$$HR = e^{\beta}$$

Competing risks – another challenge

- In some studies, individuals may experience different types of events, and one event can prevent the occurrence of the event of interest.
- These are called **competing risks**.

Suppose the event of interest is **death from cancer**.
However, some patients *may die from cardiovascular disease before dying from cancer*.
→ Death from cardiovascular disease is a **competing event**, because once it occurs death from cancer can no longer occur.



Why so important?

Standard survival methods assume that censored individuals could still experience the event later.
But in competing risks settings, this assumption is not true.

In these situations, standard survival methods may produce biased estimates, so specialized competing-risk approaches are required, such as **Cumulative incidence functions (CIF)**,

Exercise: Survival Analysis in Oncology Patients

A cohort study was conducted to evaluate survival outcomes in patients diagnosed with a specific type of cancer.
A total of 300 patients were enrolled and followed over time after diagnosis. For each patient, the following information was collected:

- age at diagnosis,
- sex,
- treatment group: **Standard treatment vs New treatment**

The primary endpoint of the study is: **time to death from cancer**.
However, not all patients experience the event during follow-up:
some patients are still alive at the end of the study, and therefore their observations are censored.

In a second part of the analysis, researchers also consider **death from other causes** as a competing event, because patients who die from another cause can no longer experience cancer-related death.

Exercise: Survival Analysis in Oncology Patients

Part 1 — Classical survival analysis

Estimate Kaplan–Meier survival curves:

- by treatment group,
- by age group (<65 vs ≥65 years).

Interpret the survival curves.

Perform log-rank tests to compare groups.

Fit a Cox proportional hazards model including: treatment, age, sex.

Interpret the hazard ratios.

Assess the proportional hazards assumption using Schoenfeld residuals.

Part 2 — Competing risks analysis

In the second analysis:

- event code 1 = death from cancer,
- event code 2 = death from other causes,
- event code 0 = censored.

Using competing risks methods:

Estimate cumulative incidence functions by:

- treatment group,
- age group.

Compare cumulative incidence curves between groups.