BIOSTATISTICS COURSE #6

SAMPLE SIZE, META-ANALYSES & LONGITUNAL DATA ANALYSIS

NOVEMBER 2025

SUMMARY OF THE COURSE #6

INTRODUCTION

01

INTRODUCTION

Here at St Wadlings we like to treat ALL our patients as INDIVIDUALS...this for example is individual No 76/09bt-c12.

SAMPLE SIZE CALCULATION

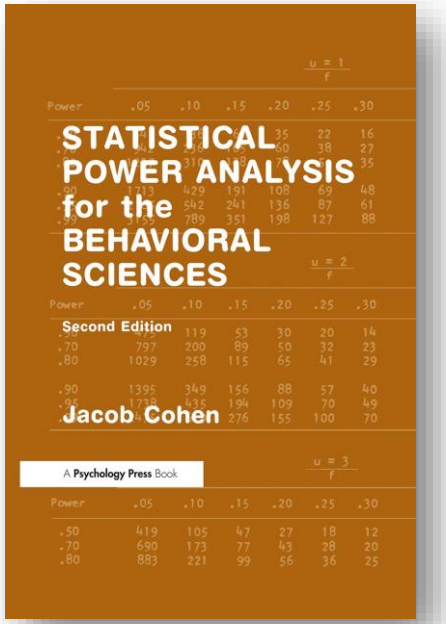
02

SAMPLE SIZE CALCULATION
ARTICLE REVIEW

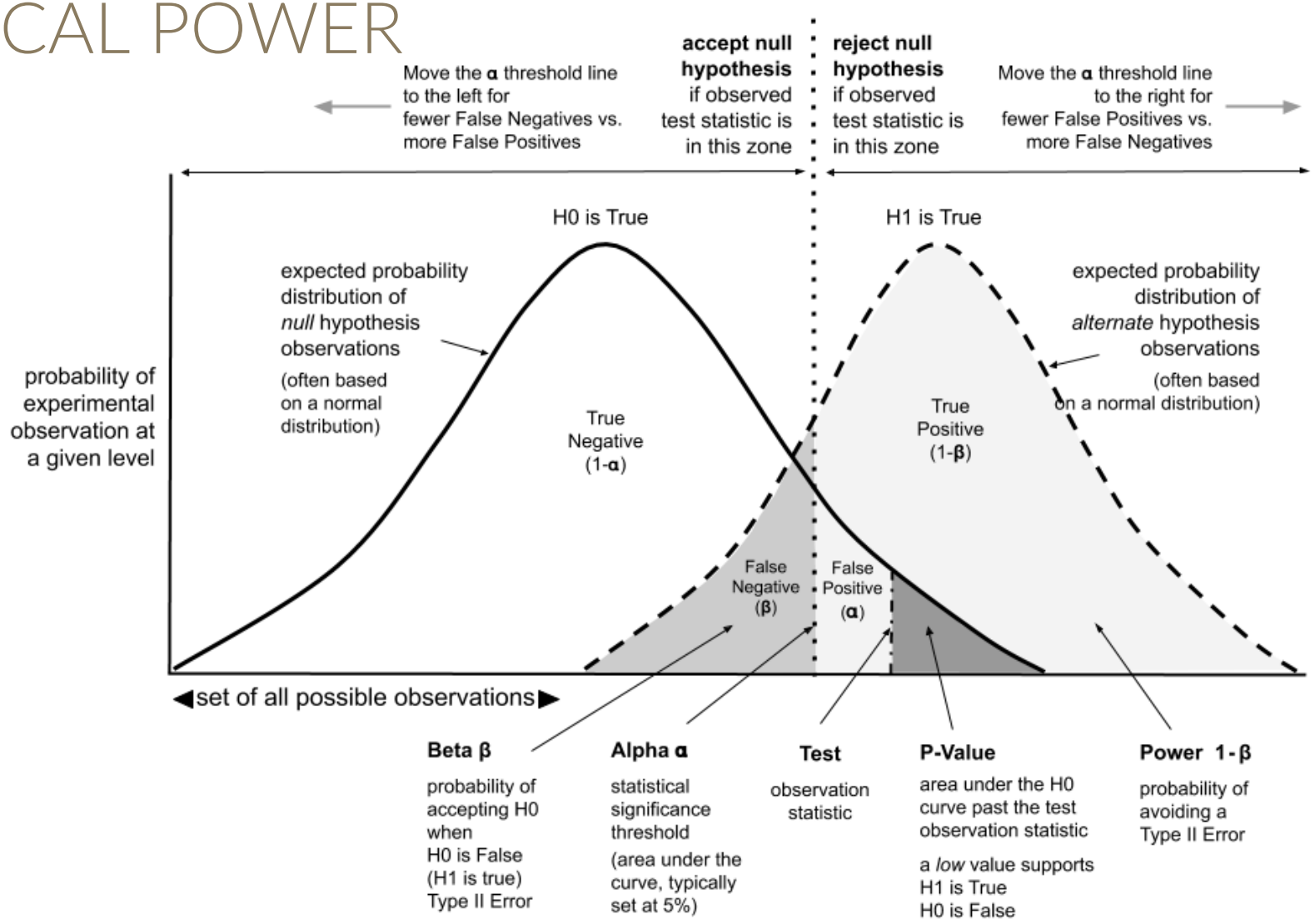
10 minutes

SAMPLE SIZE CALCULATION
HISTORY

- Jacob COHEN (1923 – 1998), American Statistician
- Famous for his works on **statistical power** and **sample size** calculation
- Invented **effect size measures** : Cohen's Kappa, Cohen's d and Cohen's h
- Published in 1988 the book "*Statistical Power Analysis for Behavioral Sciences*"
- Pioneered the principles of **meta-analyzes**

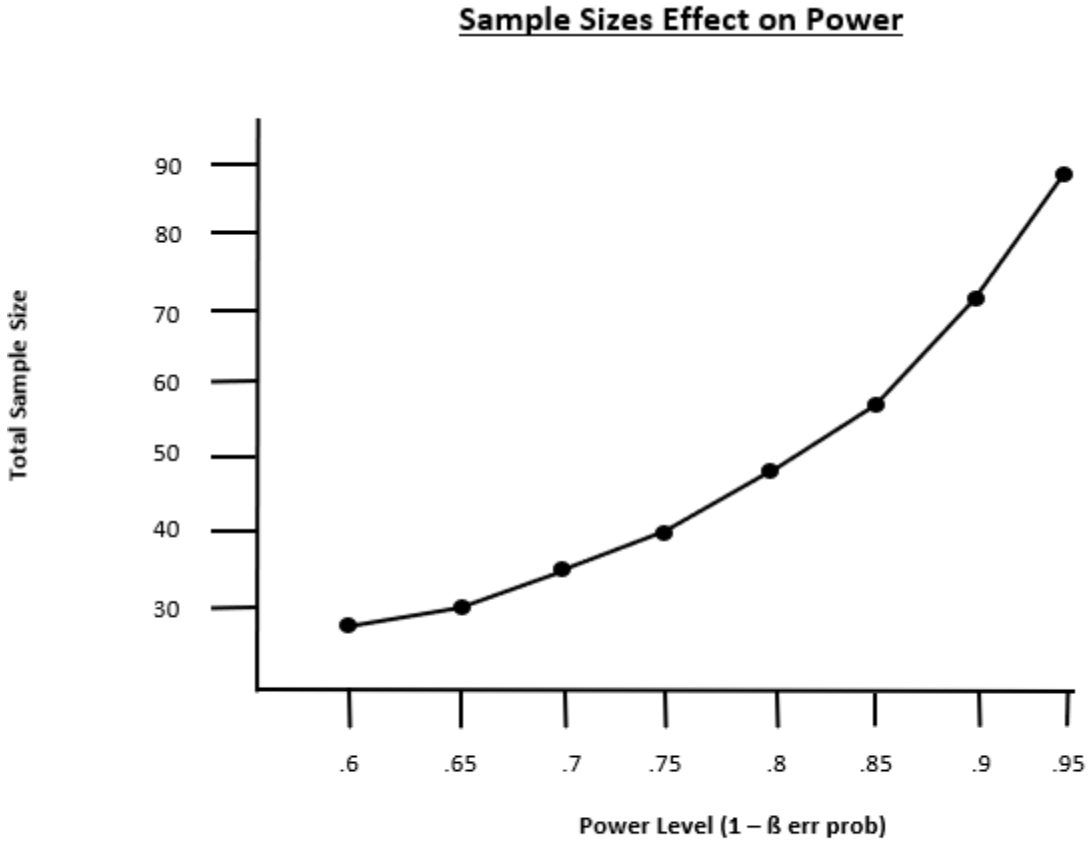


SAMPLE SIZE CALCULATION
STATISTICAL POWER



SAMPLE SIZE CALCULATION
STATISTICAL POWER

- Statistical power = 1 - β (beta risk)
- Represents the **ability of a test to detect a significant effect size**
- Strongly linked with:
  - Sample size
  - Effect size
  - Significance level



Note: As the sample size increases in the model, so does power.

SAMPLE SIZE CALCULATION
APPLICATION WITH R

Package **pwr** contains **many functions for power / sample size calculation :**

- **pwr.2p.test :** test of **two proportions** (same N in groups)
- **pwr.2p2n.test :** test of **two proportions** (unequal N in groups)
- **pwr.p.test** : Proportion (one sample)
- **pwr.r.test** : Correlation
- **pwr.chisq.test** : Chi-Square test
- **pwr.f2.test** : General Linear Model
- **pwr.anova.test** : balanced one-way ANOVA
- **pwr.norm.test** : Test of a mean to a reference (known variance)
- **pwr.t.test** : T-tests (one sample, two samples, paired)
- **pwr.t2n.test** : T-tests (two samples with unequal N)

Same way to use: all parameter values provided **except one** (result of the function)

SAMPLE SIZE CALCULATION
APPLICATION WITH R

pwr.2p.test function (*pwr* package)

Parameters : *h* = effect size : $2 \times \arcsin\left(\sqrt{p_1}\right) - 2 \times \arcsin\left(\sqrt{p_2}\right)$

n = common sample size

sig.level = significance level (default : 5%)

power = power (in %)

pwr.2p2n.test function (*pwr* package)

Parameters : *h* = effect size : $2 \times \arcsin\left(\sqrt{p_1}\right) - 2 \times \arcsin\left(\sqrt{p_2}\right)$

n1 = sample size group 1

n2 = sample size group 2

sig.level = significance level (default : 5%)

power = power (in %)

SAMPLE SIZE CALCULATION
APPLICATION WITH R

pwr.p.test function (*pwr* package)

Parameters : *h* = effect size : $2 \times \arcsin\left(\sqrt{p_1}\right) - 2 \times \arcsin\left(\sqrt{p_2}\right)$

n = sample size

sig.level = significance level (default : 5%)

power = power (in %)

pwr.r.test function (*pwr* package)

Parameters : *n* = Total sample size

r = correlation

sig.level = significance level (default : 5%)

power = power (in %)

SAMPLE SIZE CALCULATION
APPLICATION WITH R

pwr.chisq.test function (*pwr* package)

Parameters : *w* = effect size : $\sqrt{\sum_{i=1}^{m} \frac{(p0_i - p1_i)^2}{p0_i}}$

with m : number of cells

$p0_i$: cell probability in ith cell under H0

$p1_i$: cell probability in ith cell under H1

N = Total sample size

df= degrees of freedom (number of categories -1)

sig.level = significance level (default : 5%)

power = power (in %)

Package *effectsize* useful for extracting f effect size from an **test** (function *cohens_w*)

SAMPLE SIZE CALCULATION
APPLICATION WITH R

pwr.anova.test function (*pwr* package)

Parameters : *k* = number of groups

f = effect size : $\sqrt{\frac{\sum_{i=1}^{k} p_i \times (\mu_i - \mu)^2}{\sigma^2}}$

with $p_i = \frac{n_i}{N}$ and σ^2 : error variance within groups

n = common sample size in each group

sig.level = significance level (default : 5%)

power = power (in %)

Package *effectsize* useful for extracting f effect size from an ANOVA (function *cohens_f*)

SAMPLE SIZE CALCULATION
APPLICATION WITH R

pwr.t.test function (*pwr* package)

Parameters : ***d*** = effect size : $\frac{|\mu_1 - \mu_2|}{\sigma}$

with μ_1 : mean of group 1

μ_2 : mean of group 2

σ : common error standard deviation

n = Total sample size

type = type of data (two.sample, one.sample, paired)

sig.level = significance level (default : 5%)

power = power (in %)

pwr.norm.test function (*pwr* package)

Parameters : ***same as above*** *but* ***d*** = effect size : $\frac{|\mu - reference|}{\sigma}$

SAMPLE SIZE CALCULATION
APPLICATION WITH R

pwr.t2n.test function (*pwr* package)

Parameters : *d* = effect size : $\frac{|\mu_1 - \mu_2|}{\sigma}$

with μ_1 : mean of group 1

μ_2 : mean of group 2

σ : common error standard deviation

n1 = Sample size of group 1

n2 = Sample size of group 2

type = type of data (two.sample, one.sample, paired)

sig.level = significance level (default : 5%)

power = power (in %)

SAMPLE SIZE CALCULATION
APPLICATION WITH R

Package **pwr** contains **other functions for effect size calculation :**

- **cohen.ES:** give the conventional effect size (small, medium, large)

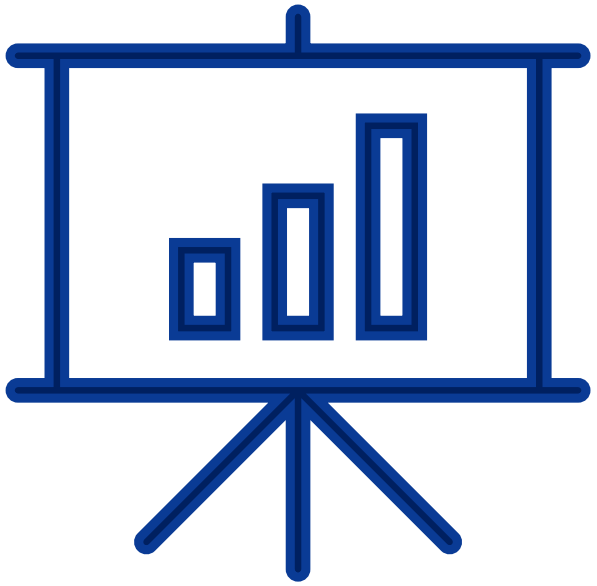
Table 1 Values of Effect Sizes and Their Interpretation

Kind of Effect Size	*Small*	*Medium*	*Large*
r	.10	.30	.50
d	0.20	0.50	0.80
η^2_P	.01	.06	.14
f^2	.02	.15	.35

Source: Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

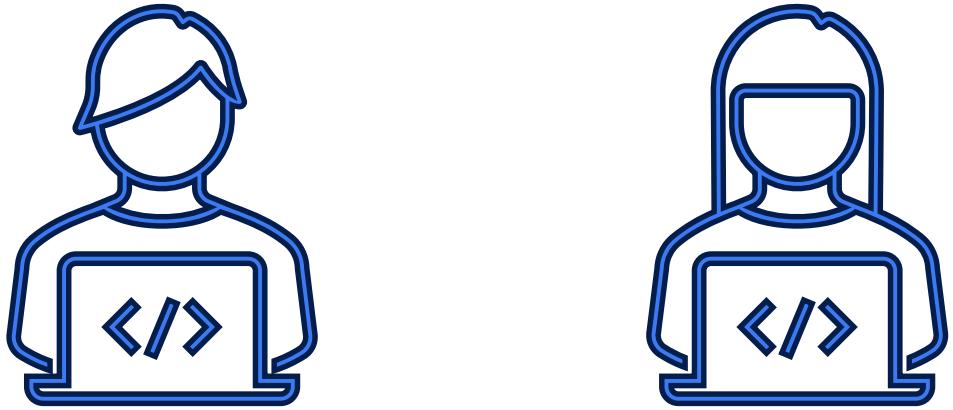
- **ES.h:** calculates effect size for proportion (h = effect size : $2 \times \arcsin(\sqrt{p_1}) - 2 \times \arcsin(\sqrt{p_2})$
- **ES.w1** : calculates effect size in the chi-squared test for goodness of fit
- **ES.w2** : calculates effect size in the chi-squared test for association
- **plot** : plot evolution of power vs sample size

SAMPLE SIZE CALCULATION



Live demo

SAMPLE SIZE CALCULATION



Time to play !
(15 minutes)

META-ANALYSIS

03

META-ANALYSIS
ARTICLE REVIEW

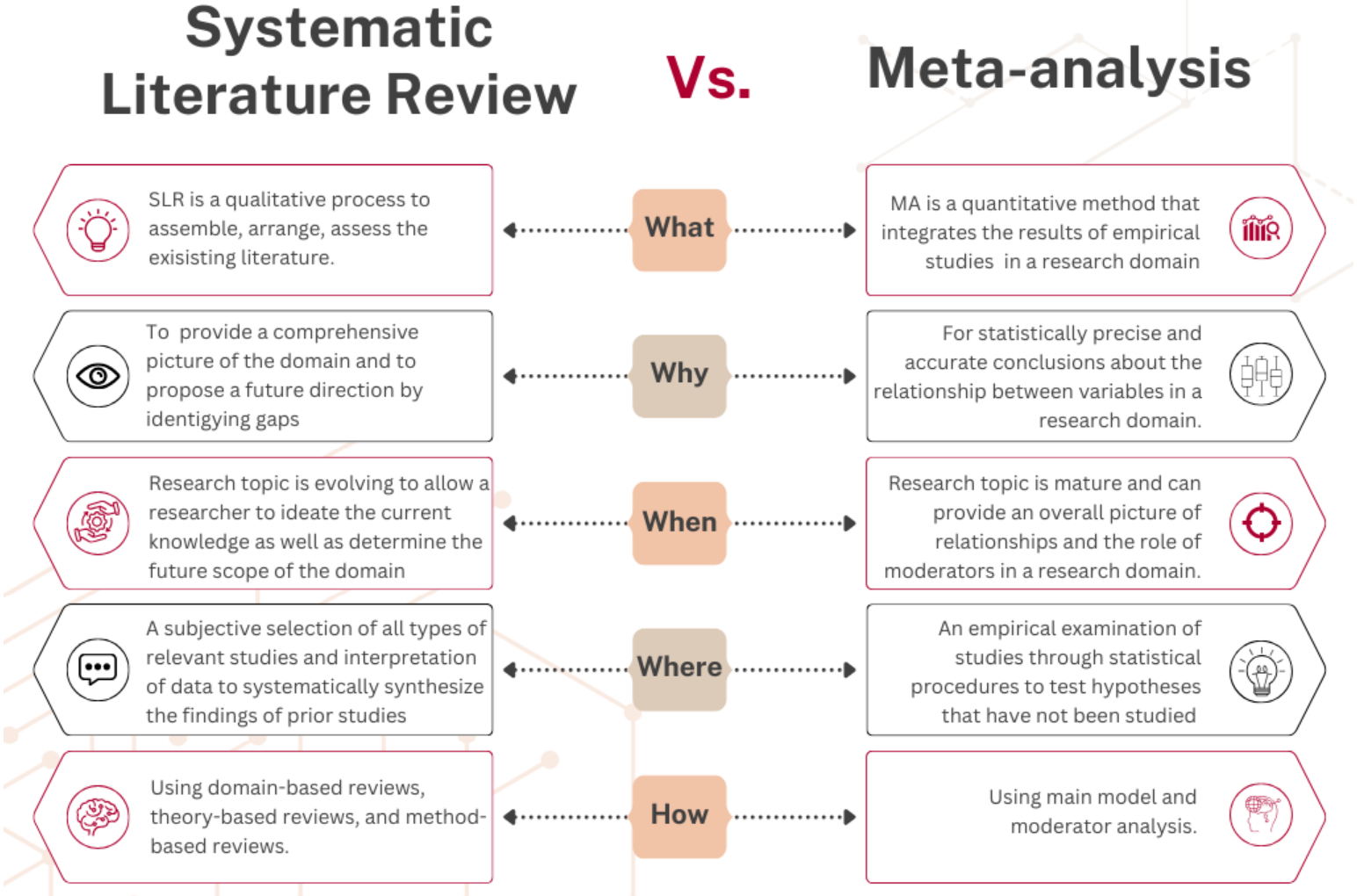
10 minutes

META-ANALYSIS
OVERVIEW

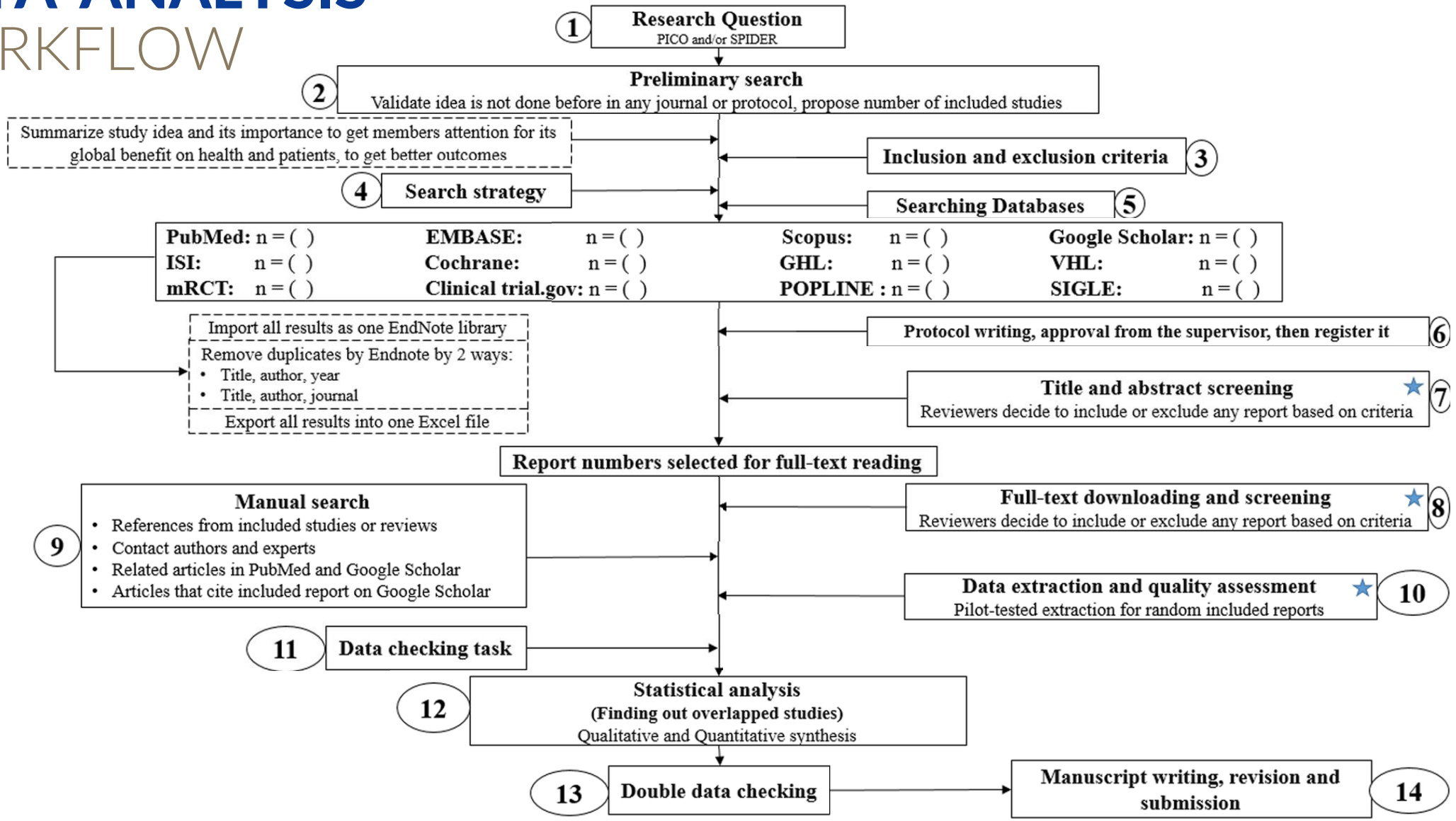
- **Definition** : method of synthesis of quantitative data from multiple independent studies addressing a common **research question**.
- The term "***Meta-Analysis***" emerged with **Gene GLASS** (born in 1940), American statistician specialized in educational psychology and social sciences.
- He published in 1978 his book "*Meta-analysis refers to the analysis of analyses*"
- **Goal** : **combine results** of many studies with statistical models in the aim to get **more statistical** power and get an overall result.

META-ANALYSIS

SYSTEMATIC LITTERATURE REVIEW VS META-ANALYSIS



META-ANALYSIS
WORKFLOW



META-ANALYSIS
BIAS

1. **Information bias** : key study variables are inaccurately measured or classified

2. **Interviewer bias** : way an interviewer ask questions or react to responses

3. **Publication bias** : decision to publish research findings is based on their nature or the direction of their results

4. **Research bias** : researcher's beliefs or expectations influence the research design or data collection process

META-ANALYSIS
BIAS

5. **Response bias** : respondents tend to provide inaccurate or false answers to self-report questions

6. **Selection bias** : sampling bias, attrition bias, survivorship bias…

7. **Other bias** : cognitive bias, confirmation bias, framing effect, halo effect, placebo effect…

META-ANALYSIS
EFFECT SIZES

1. Arithmetic mean: most commonly used central tendency measure often presented with standard-error of mean (SE)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad SE_{\bar{x}} = \frac{s}{\sqrt{n}} \text{ with } s \text{ : standard deviation of mean}$$

2. Proportion: outcome measure when we want to examine the prevalence of a disease for example

$$p = \frac{k}{n} \quad \text{and} \quad SE_p = \sqrt{\frac{p(1-p)}{n}}$$

META-ANALYSIS
EFFECT SIZES

3. **Correlations**: effect size which expresses the amount of co-variation between two variables

$$r_{xy} = \frac{\sigma^2 xy}{\sigma_x \times \sigma_y} \qquad \text{and} \qquad SE_{r_{xy}} = \frac{1 - r_{xy}^2}{\sqrt{n-2}}$$

4. **Between-Group Mean Difference**: raw, un-standardized difference in means between two independent groups.

$$MD = \overline{x_1} - \overline{x_2} \qquad \text{and} \qquad SE_{MD} = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_{pooled} = \sqrt{\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

META-ANALYSIS
EFFECT SIZES

5. **Between-Group Standardized Mean Difference**: difference in means between two independent groups, standardized by the pooled standard deviation (often called Cohen's d)

$$SMD_{between} = \frac{\overline{x_1} - \overline{x_2}}{s_{pooled}} \quad \text{and} \quad SE_{SMD} = \sqrt{\frac{n_1+n_2}{n_1 \times n_2} + \frac{SMD_{between}^{\ 2}}{2\times(n_1+n_2)}}$$

6. **Risk-ratio (RR)**: ratio of two risks. Risks are essentially proportions and can be calculated when we are dealing with binary, or dichotomous, outcome data.

	Event	No Event	
Treatment	a	b	n_{treat}
Control	c	d	n_{control}
	n_E	$n_{\neg E}$	

$$p_{E_{treat}} = \frac{a}{a+b} = \frac{a}{n_{treat}} \quad \text{and} \quad p_{E_{control}} = \frac{c}{c+d} = \frac{c}{n_{control}}$$

$RR = \frac{p_{E_{treat}}}{p_{E_{control}}}$, often expressed with **log**

META-ANALYSIS
EFFECT SIZES

RR are presented with $SE_{logRR} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$

7. Odd-ratios (OR): similar to Risk-ratio but focus on **ratio** instead of **probability**

	Event	No Event	
Treatment	a	b	n_{treat}
Control	c	d	n_{control}
	n_E	$n_{\neg E}$	

$OR_{treat} = \frac{a}{b}$ and $OR_{control} = \frac{c}{d}$

$OR = \frac{OR_{treat}}{OR_{control}}$, often expressed with **log**

OR are presented with $SE_{logOR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

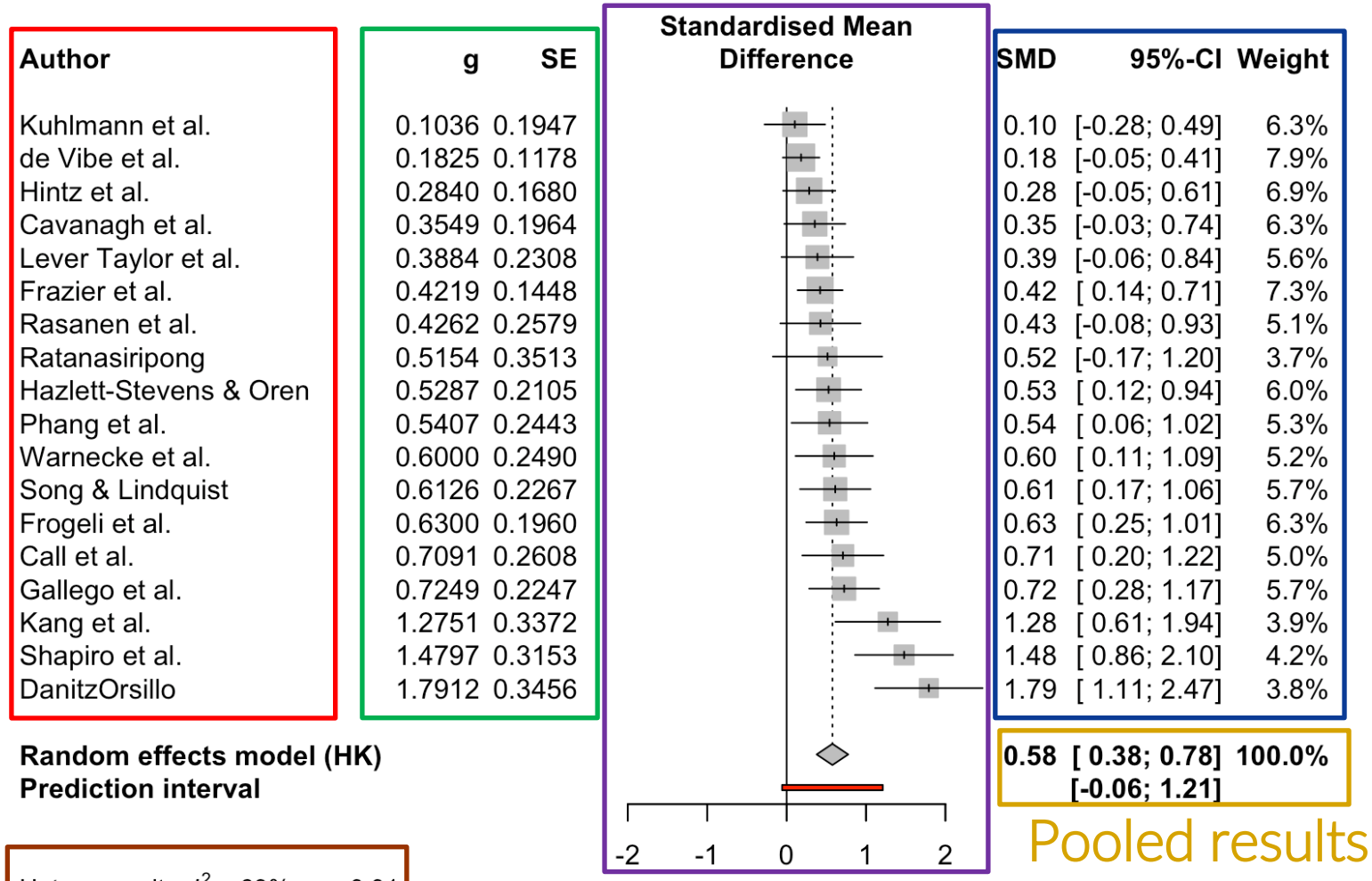
META-ANALYSIS
FOREST-PLOT

Effect size

Plot

Studies ID

Meta-analysis results



Pooled results

Heterogeneity

Heterogeneity: I^2 = 63%, p < 0.01

META-ANALYSIS
HETEROGENEITY

- The more we include studies in a meta-analysis, the more we increase the **variability of measures** (bias…) and then increase **heterogeneity** : **clinical** diversity, **methodological** diversity, **statistical** heterogeneity.

- How to evaluate **heterogeneity** in a meta-analysis :

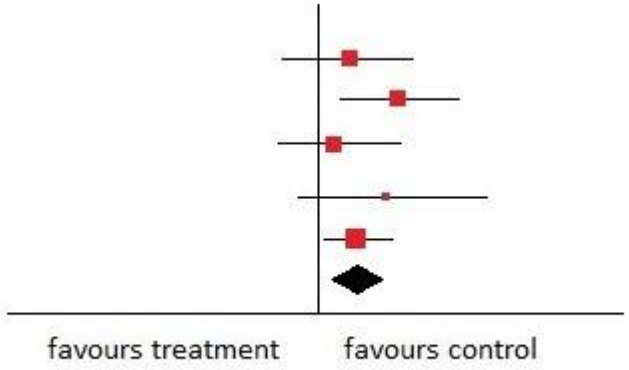
I^2 statistic : $I^2 = \left(\frac{Q-df}{Q}\right) \times 100\%$ with Q : Chi^2 statistic and df = degrees of freedom

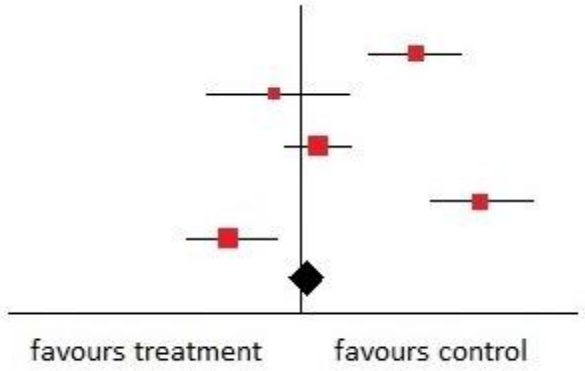
Value	Interpretation
0% to 30%	**Low** heterogeneity
30% to 50%	**Moderate** heterogeneity
50% to 75%	**Substantial** heterogeneity
75% to 100%	**Considerable** heterogeneity

META-ANALYSIS
HETEROGENEITY

Strategies for addressing heterogeneity :

1. Check again that the data are correct
2. Do not do a meta-analysis
3. Explore heterogeneity
4. Ignore heterogeneity
5. Perform a random-effects meta-analysis
6. Reconsider the effect measure
7. Exclude studies



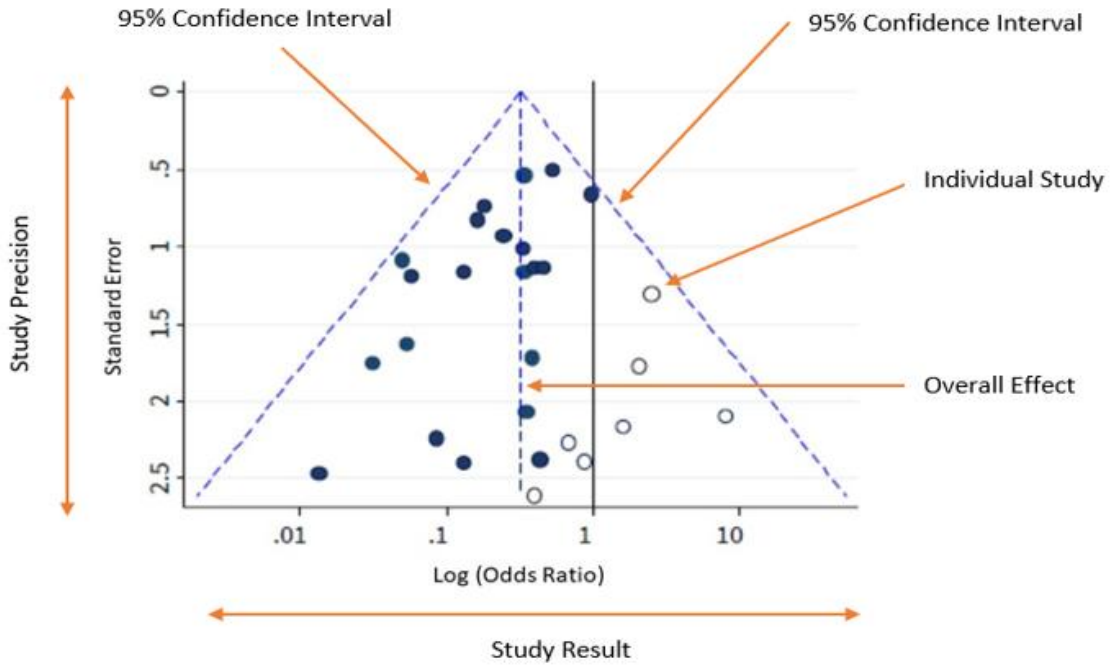


META-ANALYSIS
HETEROGENEITY PLOTS : FUNNEL PLOT

Plots for studying heterogeneity :

Funnel plot :

- Studies are represented with **dots**
- The maximum number of studies should be located **inside the pyramid**
- Studies must be **symmetrically distributed** around the overall effect
- A study with high impact on **heterogeneity** can be identified outside the triangle.

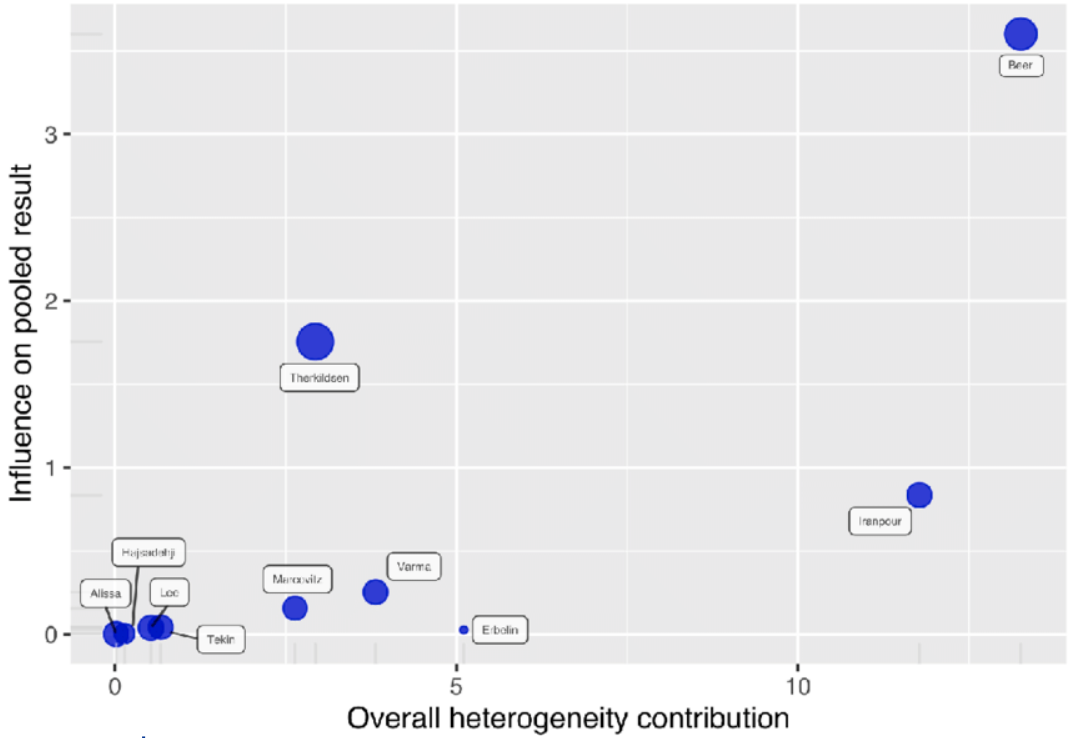


META-ANALYSIS
HETEROGENEITY PLOTS : BAUJAT PLOT

Plots for studying heterogeneity :

Baujat plot :

- Studies are represented with **dots** (size is related to the weight of each study)
- X-axis : overall **heterogeneity** contribution of each study
- Y-axis : influence on **pooled results** of each study
- Removing a study with high levels in both parameters will have a **great impact on results** as well as on heterogeneity of the meta-analysis.

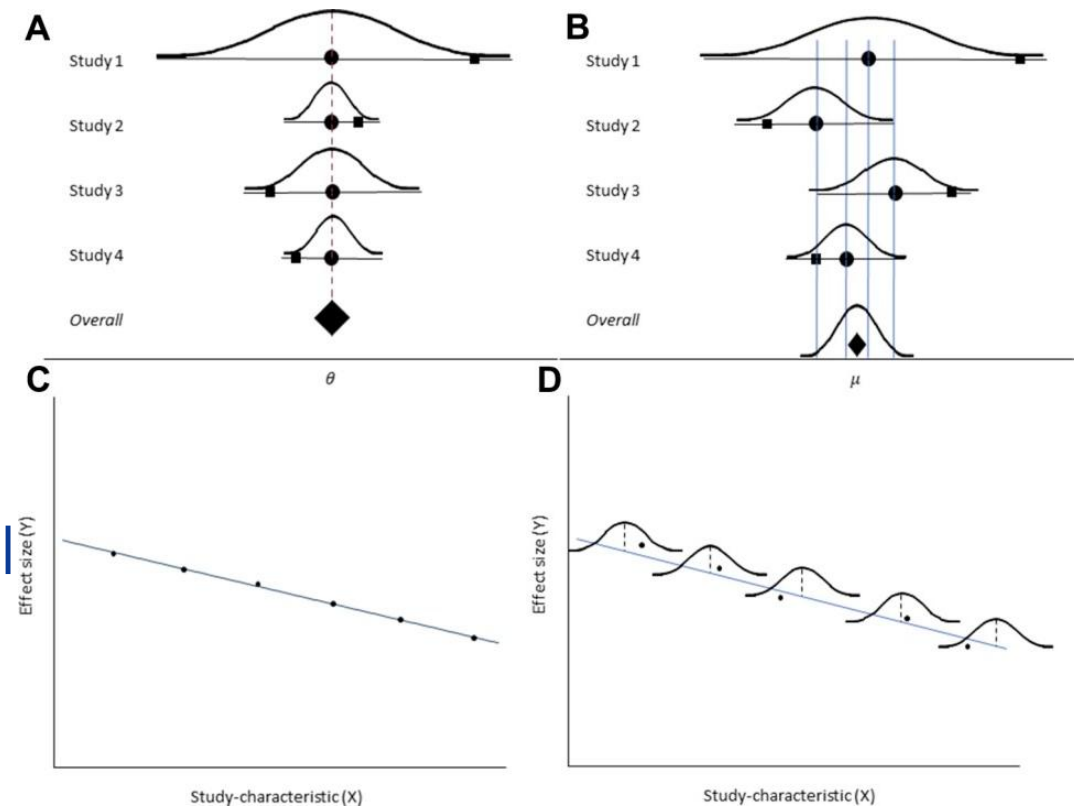


META-ANALYSIS
FIXED VS RANDOM EFFECTS MODELS

Two models available:

- **Fixed effects model** : true effect size across all studies, and estimation of a **common effect size** (known as summary effect size) (plots A and C)

- **Random effects model** : assuming a **normal distribution** of true effect sizes and estimate the mean (also known as summary effect size) and the variance (known as heterogeneity) of this distribution (plots B and D)



META-ANALYSIS
META-REGRESSION

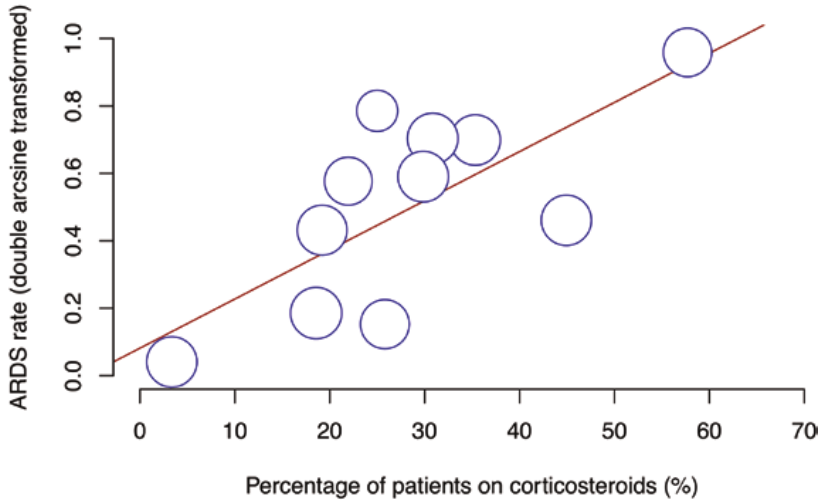
Definition : statistical model built on results of a meta-analysis in order to study impact of **covariates** (also called "***moderators***") on effect-sizes.

Goal : identify covariates which can bring heterogeneity in a meta-analysis

Prerequisites : availability of covariates values (often baseline characteristics : Men/Women ratio, mean age, mean BMI…)

Results can be visualized with a **bubble plot**

A pvalue is provided by the model



META-ANALYSIS
APPLICATION WITH R

Package *meta* allows to run meta-analysis :

- Function **metacont** for Y = continuous variable
- Function **metabin** for Y = binary variable
- Function **metacor** for single correlation
- Function **metameans** for single mean
- Function **metaprop** for single proportions
- Function **metarate** for single incidence rates

Parameters : (example for metacont)

n*, *n.e*, *n.c = sample size of studies (e = experimental group ; c = control group)

mean, mean.e, mean.c = means of outcome of studies

sd, sd.e, sd.c = standard-deviation of outcome of studies

studlab = names of studies

subgroup = variable for subgroups analysis

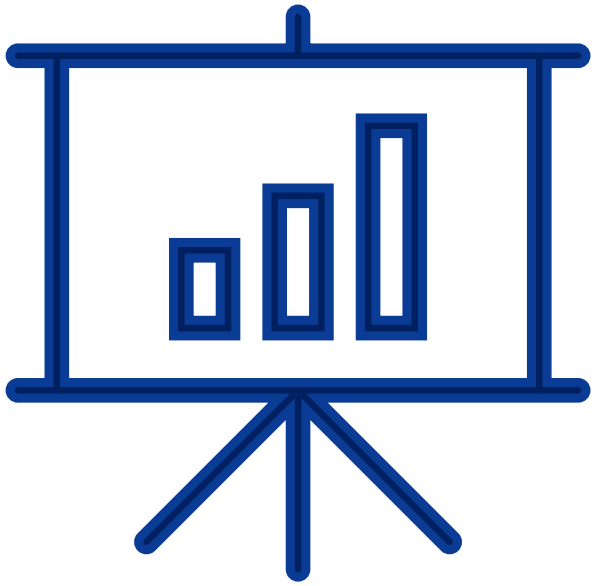
META-ANALYSIS
APPLICATION WITH R

Other functions are available in ***meta*** package :

- ***funnel*** : funnel plot
- ***baujat*** : baujat plot

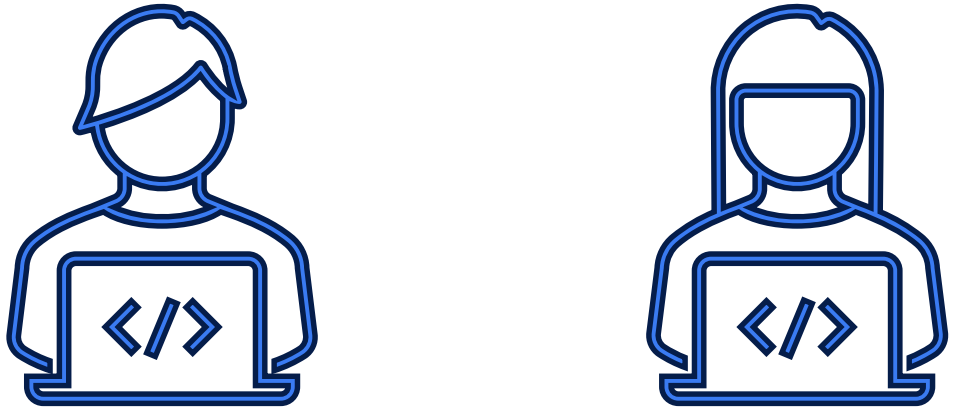
Meta-regressions with ***metareg*** function and ***bubble.metareg*** function for bubble plot

META-ANALYSIS



Live demo

META-ANALYSIS



Time to play !
(10 minutes)

CLINICAL TRIAL DATA

04

CLINICAL TRIAL DATA
DATA COLLECTION HISTORY

- **Census** of population exists since Ancient Egypt and aims for a government to evaluate the size, composition, structure and evolution of a population across time for management of public policies.

- **Data collection** on individuals exists since Humans were able to write and store information in **physical documents** : papyrus, sheets, books, registers…

- With the rising of **computers**, **datacenters**, **Internet** and **numerical supports** (databases, datalakes, clouds, Excel sheets…), data collection, monitoring and storage saw an **exponential increase**.

- With this rise, new challenges and **ethical questions** raised : **data privacy**, **spying** on social networks and medias, **phishing**, data **hacking**…

CLINICAL TRIAL DATA
CLINICAL DATA MANAGEMENT

- **Clinical Data Management (CDM) :** process of collecting, cleaning, analyzing and managing study data in clinical research.

- **Main goal** : gather as much quality study data as possible.

- **CDM** applies across **all three main stages of clinical trials** and even occurs in pre-clinical phases.

- In the pursuit to speed up the drug development process, CDM has become particularly important for **achieving faster development times**. With better data quality comes better reliability. This lets evaluators make quicker decisions with improved efficiency.

CLINICAL TRIAL DATA
CLINICAL DATA MANAGEMENT

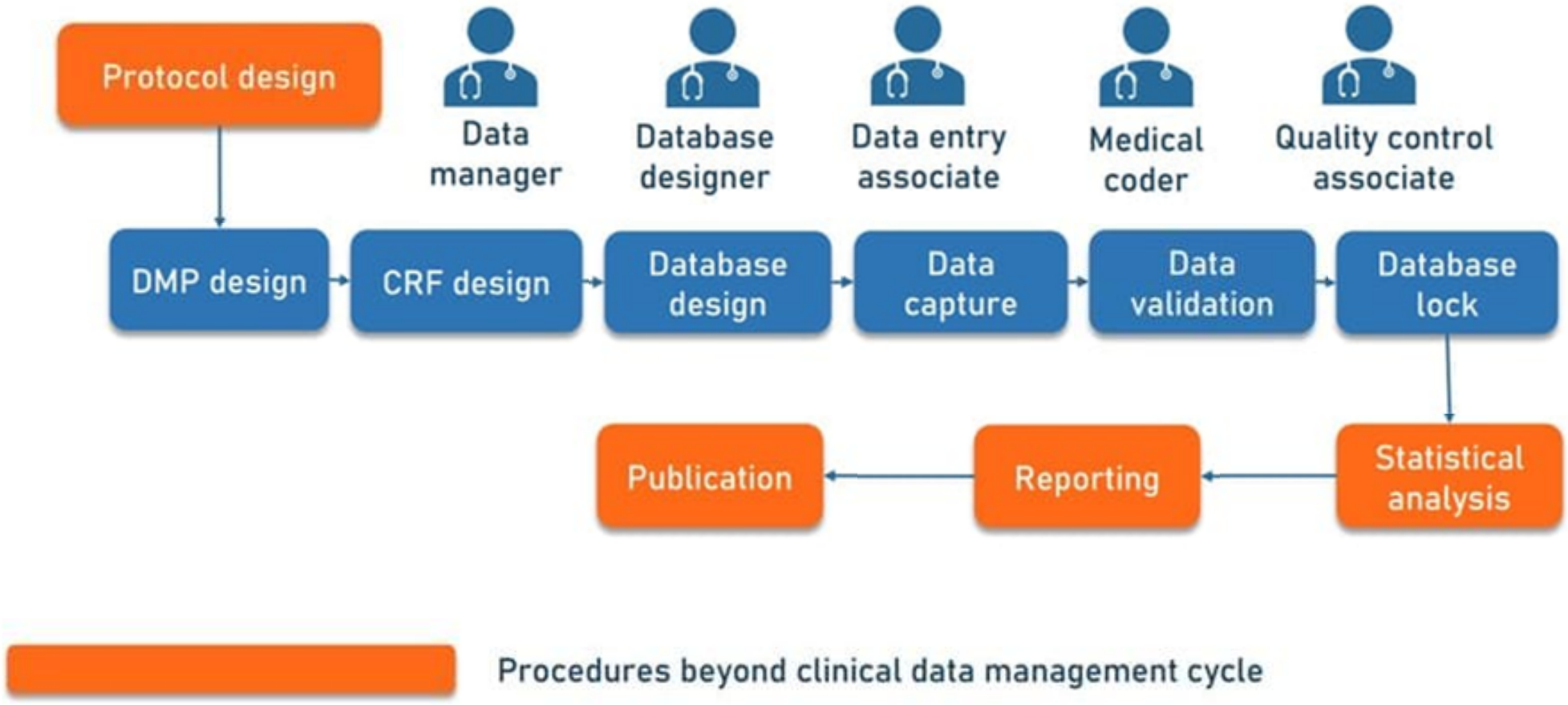
- CDM ensures :

  - ✓ Quality data collection
  - ✓ Compliance with regulatory standards
  - ✓ Data security
  - ✓ Efficient clinical research process

- Three main pillars of data quality that clinical data management processes aim to achieve:

  - ✓ Accuracy : data that is true to life and error free.
  - ✓ Completeness : datasets that contain the full-extent of the required data.
  - ✓ Consistency : data entries that follow consistent formatting throughout.

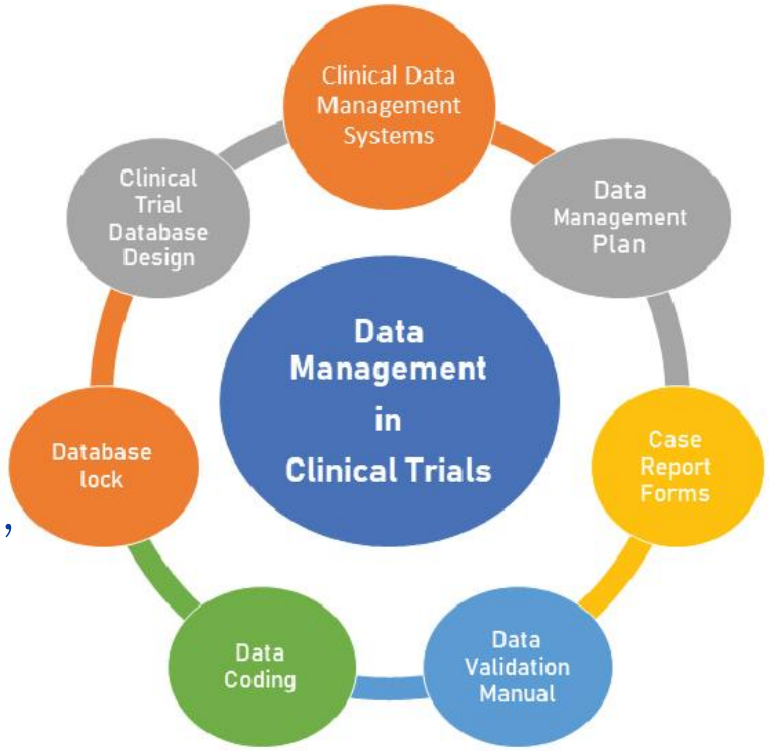
CLINICAL TRIAL DATA
CLINICAL DATA MANAGEMENT WORKFLOW



CLINICAL TRIAL DATA
DATA MANAGEMENT PLAN

DMP describes the following aspects :

- Data to be gathered from trial participants,
- Existing data that can be integrated,
- Data formats,
- Metadata and its standards,
- Storage and backup methods,
- Security measures to protect confidential information,
- Data quality procedures,
- Responsibility assignments across team members,
- Access and sharing mechanisms and limitations,
- Long-time archiving and preservation procedures,
- The cost of data preparation and archiving, and
- Compliance with relevant regulations and requirements.



CLINICAL TRIAL DATA
CASE REPORT FORM

CRFs collect only data necessary for the clinical study, avoiding any redundancy. The fields to be filled in may include :

- **Demographics** (age, gender)
- **Basic** measurements (height, weight)
- **Vital signs** (blood pressure, temperature, etc.) captured at various time points
- **Lab exams**
- **Medical history**
- **Adverse events**
- Other **parameters**, based on the research requirements.

CRF can be **physical** (printed documents) and/or **electronic** (eCRF).

CLINICAL TRIAL DATA
CLINICAL DATA INTERCHANGE STANDARDS CONSORTIUM

- Clinical Data Interchange Standards Consortium (CDISC) is the official internation organization dedicated to the development and maintenance of standards in clinical data collection, storage and sharing.

- Founded in 1997 and hosted in Austin, Texas (USA).

- Since 2016, CDISC standards are mandatory for FDA submission.

- Study Data Tabulation Model (SDTM) : harmonize data structure in human clinical trial for an easy data re-use.

CLINICAL TRIAL DATA
CASE REPORT FORMS REVIEW

10 minutes

LONGITUDINAL DATA ANALYSIS

05

LONGITUDINAL DATA ANALYSIS
ARTICLE REVIEW

10 minutes

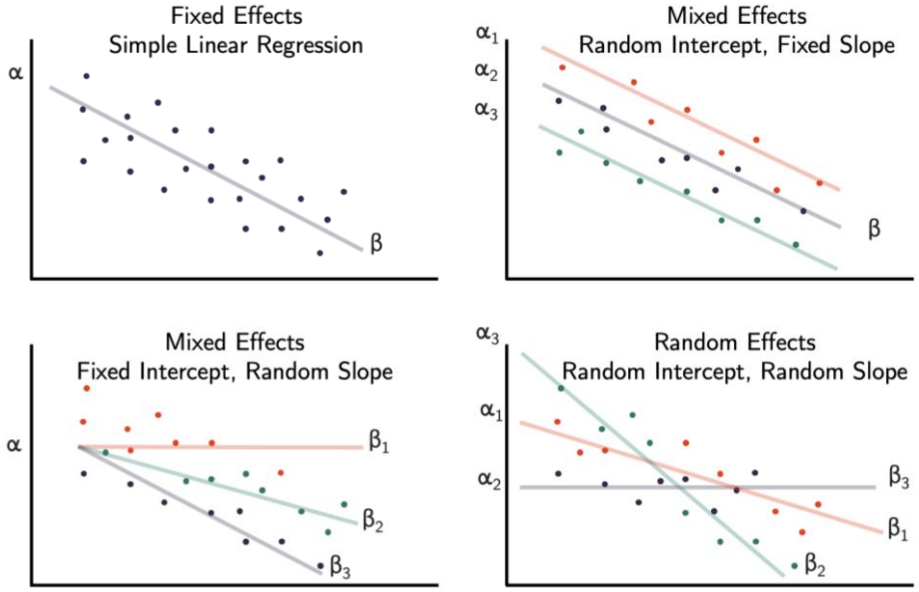
LONGITUDINAL DATA ANALYSIS
LONGITUDINAL DATA OVERVIEW

- **Longitudinal data** : collection of data for an individual across time.
- During **clinical trials** (except retrospective trials), data is collected, monitored and stored in **secured database** for each patient included in the trial.
- This data can have **various forms** : **biomarkers** (in blood, urine, stood…), **electronic health records** (heart rate, respiratory rate, temperature…), quality of life **questionnaires**…
- Each patient is anonymized with **a unique Patient ID**.
- Data collection, treatment and storage is **strictly controlled during a clinical trial**.

LONGITUDINAL DATA ANALYSIS
FIXED EFFECTS VS MIXED EFFECTS

- Fixed effect :
  - Estimate separate levels with no **relationship** assumed between the levels.
  - **Easily interpretable** effect (p-value)
  - **Most common** in a linear model.

- Random effect :
  - Each level can be thought of as a **random variable** from an underlying process or distribution
  - **Complex interpretable** effect (often not studied)
  - **Less common** in a linear model.



LONGITUDINAL DATA ANALYSIS
LINEAR MIXED MODEL DEFINITION

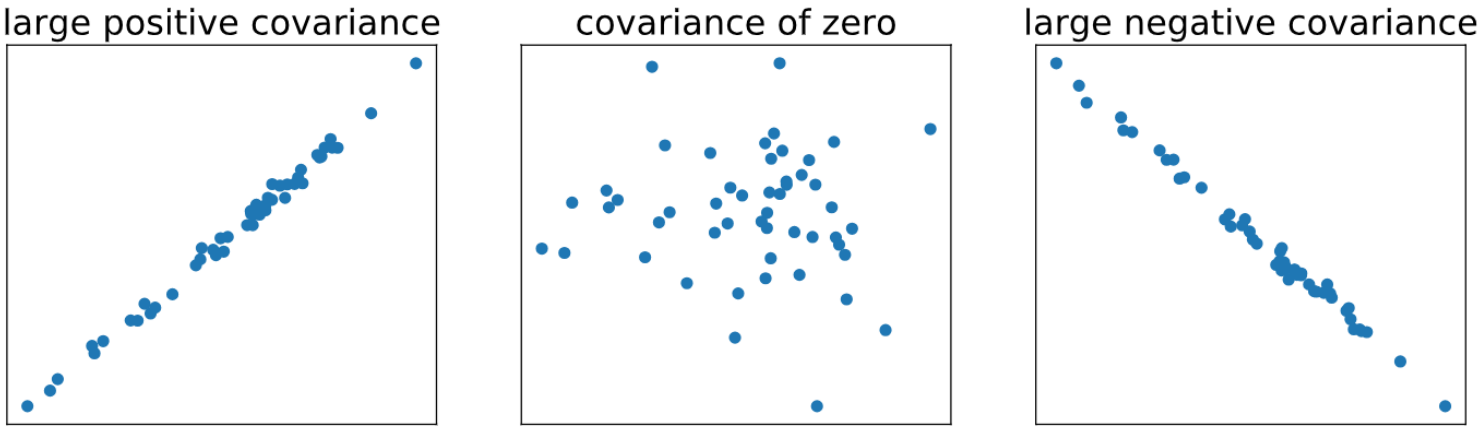
- Linear model which contains **fixed** and **random effects**

$$\boldsymbol{Y = X\beta + Zu + \varepsilon}$$

With
$\boldsymbol{Y}$: vector of observations (continuous)
$\boldsymbol{X}$: design matrix of fixed effects
$\boldsymbol{\beta}$: unknow vector a fixed effects
$\boldsymbol{Z}$: design matrix of random effects
$\boldsymbol{u}$: unknow vector a random effects
$\boldsymbol{\varepsilon}$: residuals

- Often used for analysis of **longitudinal clinical data** : many timepoints / patient

LONGITUDINAL DATA ANALYSIS
COVARIANCE STRUCTURES



- Introducing random effect implies to use a specific **covariance matrix.**

- **Overall shape :**
  (example of a matrix for 3 patients evaluated each at 3 timepoints)

$$\begin{bmatrix} R_1 & 0 & 0 \\ 0 & R_2 & 0 \\ 0 & 0 & R_3 \end{bmatrix}$$

LONGITUDINAL DATA ANALYSIS
COVARIANCE STRUCTURES

- Many designs for **R** :

  - **General form :**

matrix of **variances** of each timepoint on the diagonal and covariances of timepoints elsewhere

$$R_i = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_2^2 \end{bmatrix}$$

  - **Compound symmetry :**

matrix of **random term variances** with the same **total variance** of the diagonal (default matrix in *lme* function)

$$R_i = \begin{bmatrix} \sigma^2 & \theta & \theta \\ \theta & \sigma^2 & \theta \\ \theta & \theta & \sigma^2 \end{bmatrix}$$

LONGITUDINAL DATA ANALYSIS
COVARIANCE STRUCTURES

- Many designs for R :

  - First order autoregressive :

matrix with global variance on the diagonal and correlations between timepoints with an index

$$R_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

  - Toeplitz :

matrix of random term variances with the same total variance of the diagonal

$$R_i = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

LONGITUDINAL DATA ANALYSIS
APPLICATION WITH R

lme function (*nlme* package)

Parameters : *formula* = Y ~ fixed effects
random = ~ random effects
data = dataset
subset = subset of the rows of the dataset to keep
method = « REML » (for Restricted Log-Likelihood method)

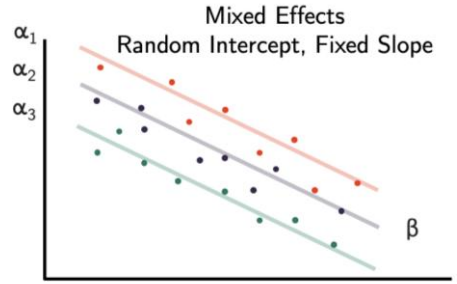
Outputs : *lme* object which contains results :
Coefficients of the equation
Residuals
Fitted values
Covariance matrix
REML value

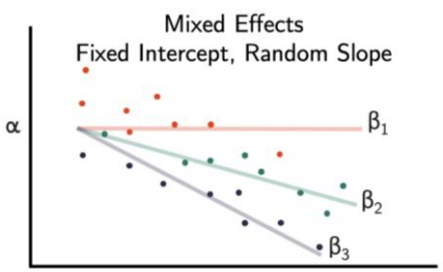
LONGITUDINAL DATA ANALYSIS
APPLICATION WITH R

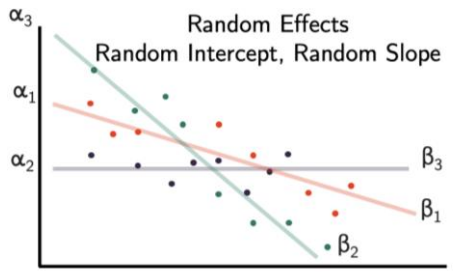
getVarCov function (*nlme* package) : **variance** and **correlation components** from a random-effect model

How to write a **random effect** in *lme* :

- ~ 1 | Patient_ID : random intercept
- ~ Time | Patient_ID : random slope
- ~ (1 + Time) | Patient_ID :
  random intercept + random slope







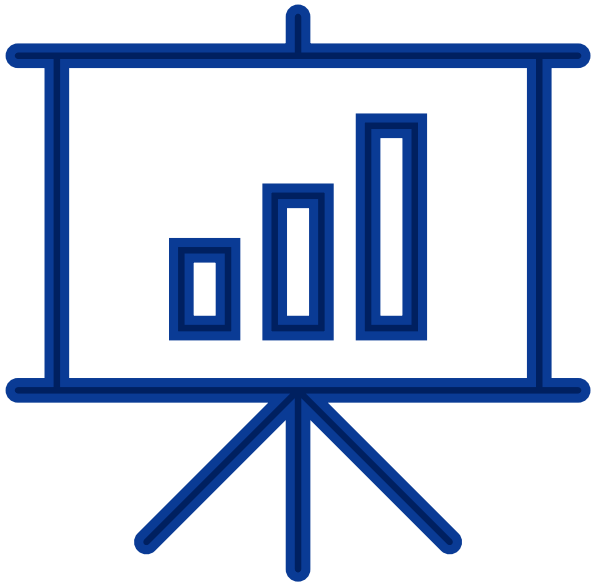
LONGITUDINAL DATA ANALYSIS
APPLICATION WITH R

r2 function (***performance*** package) : get the R^2 of mixed models (also called *Nakagawa's R^2*)

Two R^2 are computed :

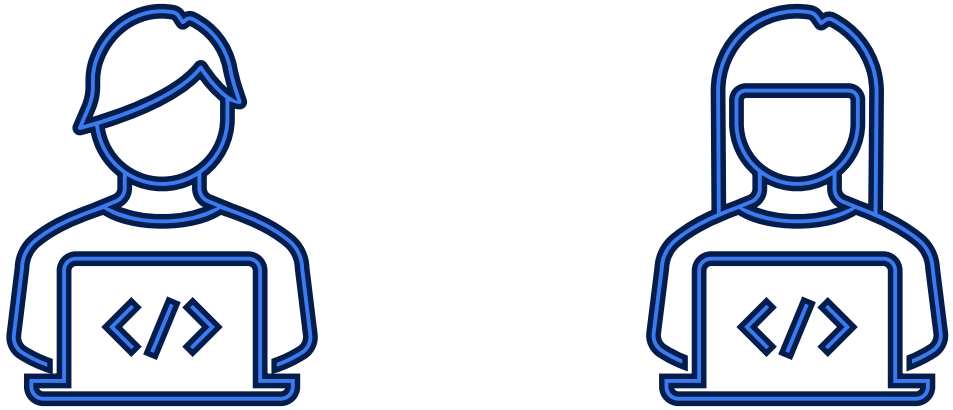
- **Conditional R^2** : R^2 of the overall model (fixed + random effects)
- **Marginal R^2** : R^2 of fixed effects

LONGITUDINAL DATA ANALYSIS



Live demo

LONGITUDINAL DATA ANALYSIS



Time to play !
(15 minutes)

SURVIVAL ANALYSIS

06

SURVIVAL ANALYSIS
ARTICLE REVIEW

10 minutes

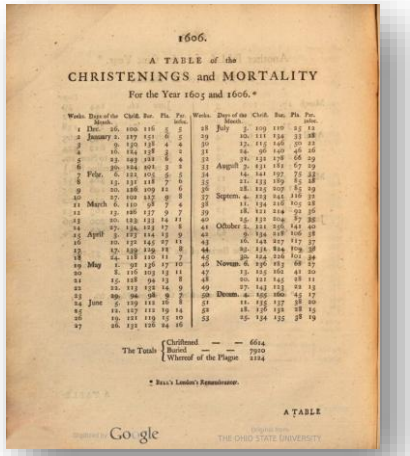
SURVIVAL ANALYSIS
BRIEF HISTORY

- Survival analysis emerged during the 17th century with the **rising of statistics in medicine.**

- **John GRAUNT** (1620 – 1674), English founder of demography and epidemiology.


Captain John Graunt

He produced a **life table** which shows, for each age, the probability that a person of that age will die before their next birthday ("*probability of death*").


1606.

A TABLE of the

CHRISTENINGS and MORTALITY

For the Year 1605 and 1606.*

He used it during the **bubonic pledge** in London in 1665-1666 (100.000 deaths in 18 months, almost ¼ of the city population)

SURVIVAL ANALYSIS
OVERVIEW

- **Definition** : branch of statistics dedicated to the analysis of expected duration of time **until one event occurs**

- Many **events** to model : death, cancer remission, heart attack…

- Also used in **economics / sociology** fields

- Used in several ways :

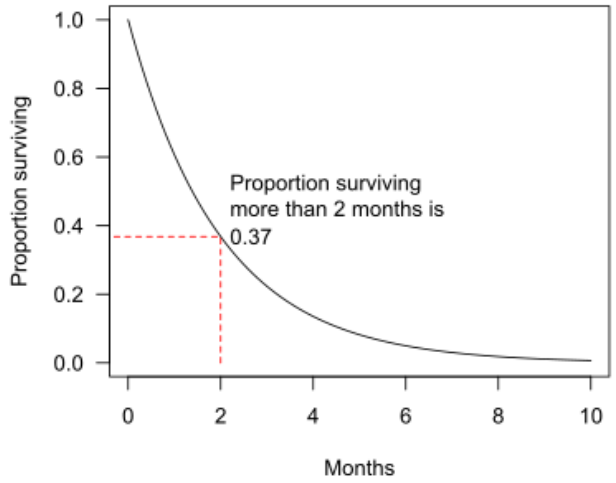
    - **Describe** the survival times of members of a group
    - **Compare survival** time of two or more groups
    - Describe the **effect of categorical or quantitative variables** on survival

SURVIVAL ANALYSIS
KAPLAN-MEIER ESTIMATOR

- **Definition** : Non-parametric statistic used to estimate the **survival function** from lifetime data.


- Invented by **Edward L. KAPLAN** (1920 – 2006) and **Paul MEIER** (1924 – 2011), two American statisticians

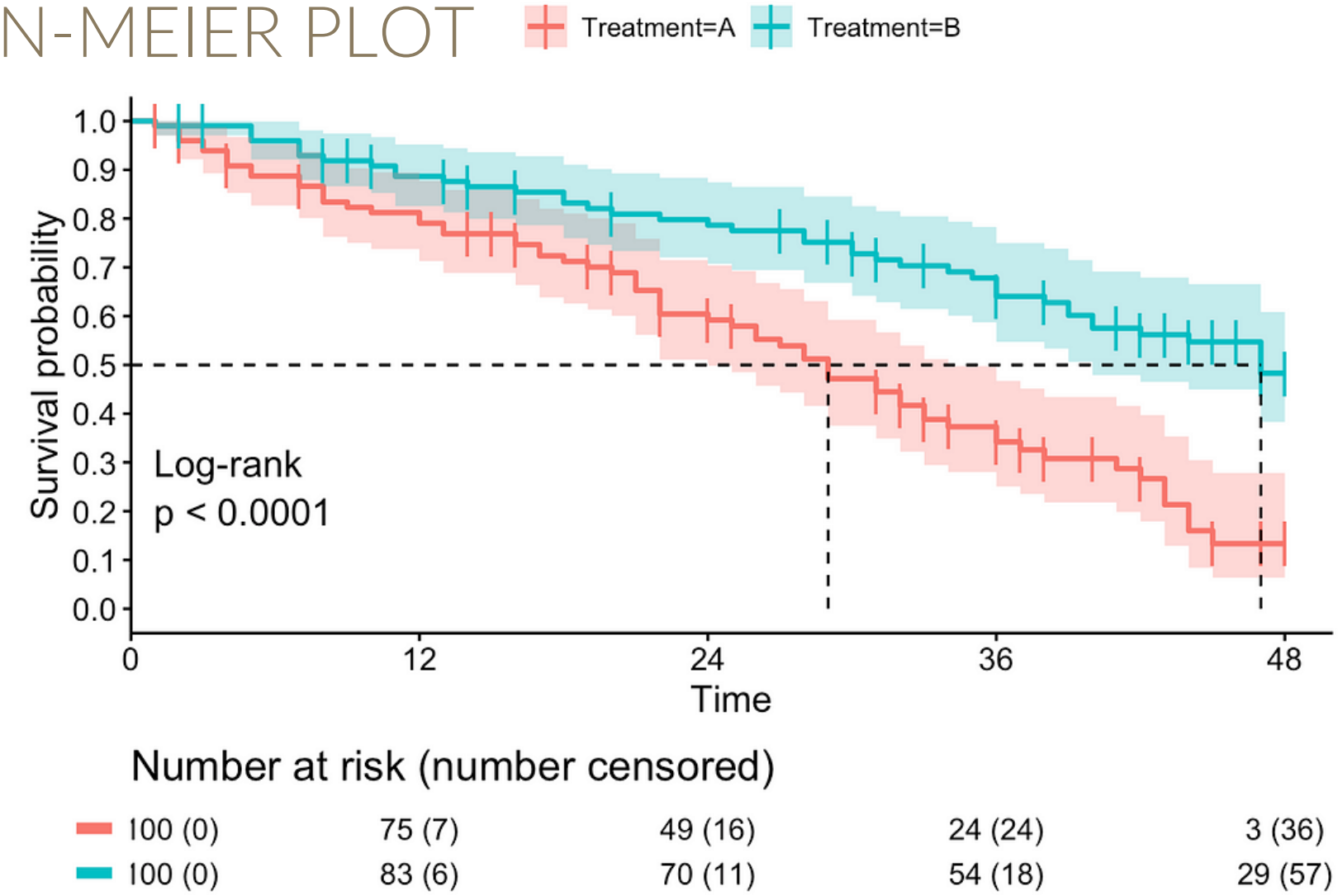
- **Survival function** : $S(t) = \mathrm{P}(T > t)$



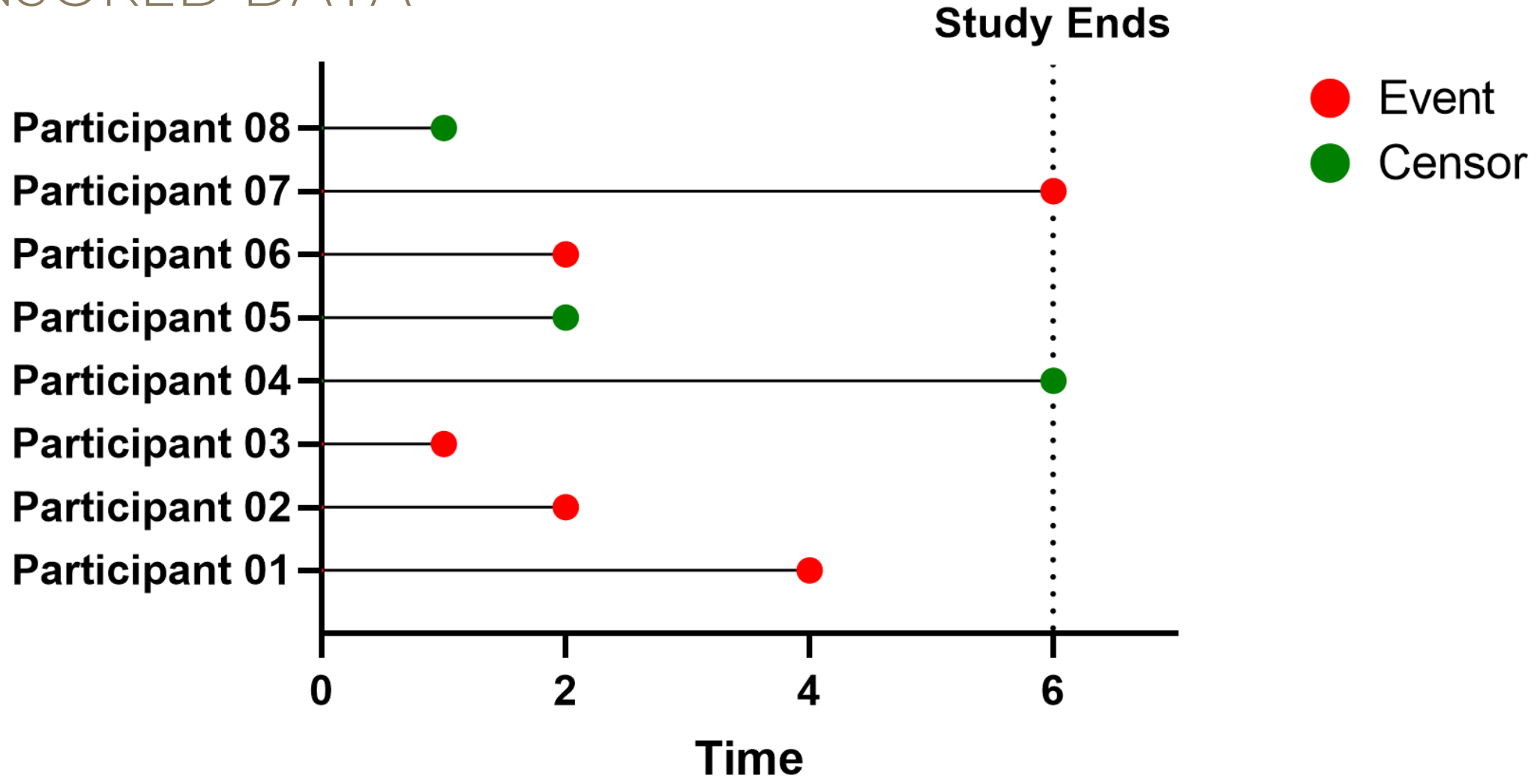
- **Kaplan-Meier plot** : illustrates the evolution of **overall survival or within groups across time**.

- **Log-rank test** : compare the **survival of two groups** with a pvalue (also called Cochran-Mantel-Haenszel test).

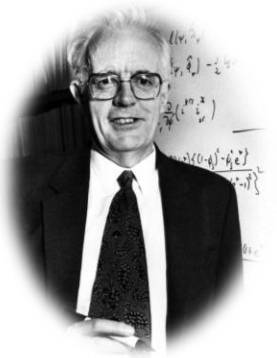
SURVIVAL ANALYSIS
KAPLAN-MEIER PLOT



SURVIVAL ANALYSIS
CENSORED DATA



SURVIVAL ANALYSIS
PROPORTIONAL HAZARDS MODEL (COX MODEL)



- Invented by **Sir David COX** (1924 – 2022), British statistician in 1972. He is also the inventor of **logistic regression** in 1958.

- **Goal** : study the **impact of one or more parameters** (continuous or categorical) on the **survival**.

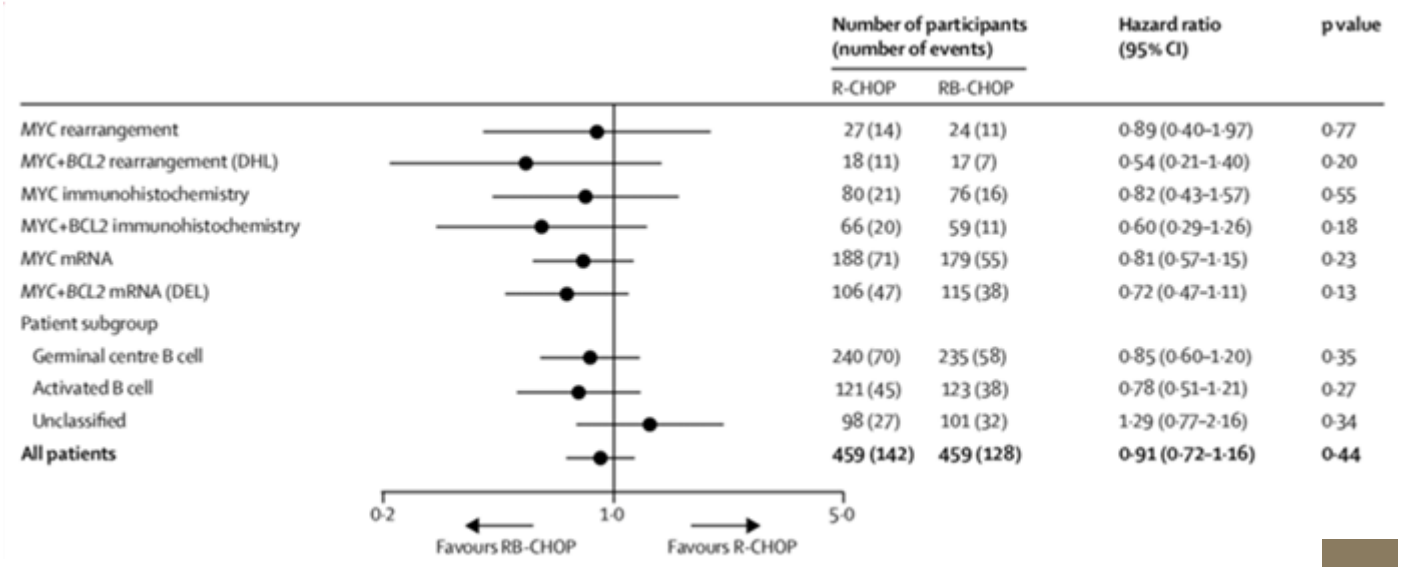
- Uses **hazard-function** :

$$\lambda(t|X_i) = \lambda_0(t)\exp\left(\beta_1 X_{i1} + \cdots + X_{ip}\right)$$

with : t : time
$\lambda_0(t)$: baseline risk at time t
X_{ip} : covariate p value for individual i

SURVIVAL ANALYSIS
HAZARDS RATIOS

- Hazard-ratio : $\mathrm{HR} = \frac{\lambda(t|x+1)}{\lambda(t|x)}$ (values : 0 to +∞)
- **Assumption** : assumes the hazards for any two individuals have the same **proportion at all times**
- Illustration with **forest-plot**
- HR = exp(β)



SURVIVAL ANALYSIS
APPLICATION WITH R

survfit function (*survival* package) : Kaplan-Meier curve

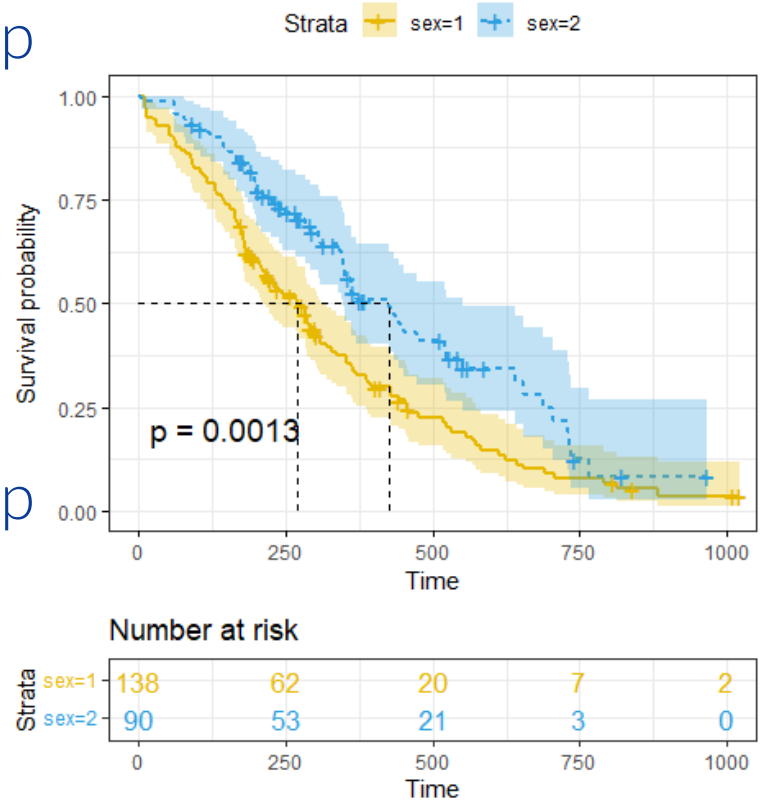
Parameters : *formula* = Surv(time, outcome) ~ group
data = dataset

Outputs : *Kaplan-Meier curve*

survdiff function (*survival* package) : log-rank test

Parameters : *formula* = Surv(time, outcome) ~ group
data = dataset

Outputs : *log-rank test*



ggsurvplot function (*survminer* package) : customized plots

SURVIVAL ANALYSIS
APPLICATION WITH R

coxph function (*survival* package) : Cox model

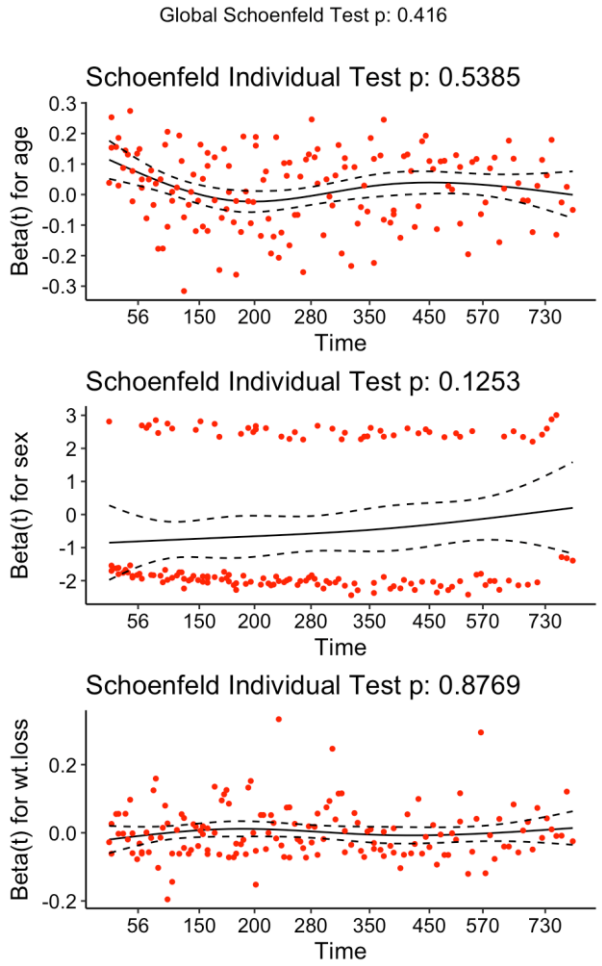
Parameters : *formula* = Surv(time, outcome) ~ covariates
data = dataset

Outputs : Cox model

cox.zph function (*survival* package) : test for proportional hazard assumption

Parameters : *model* = Cox model object

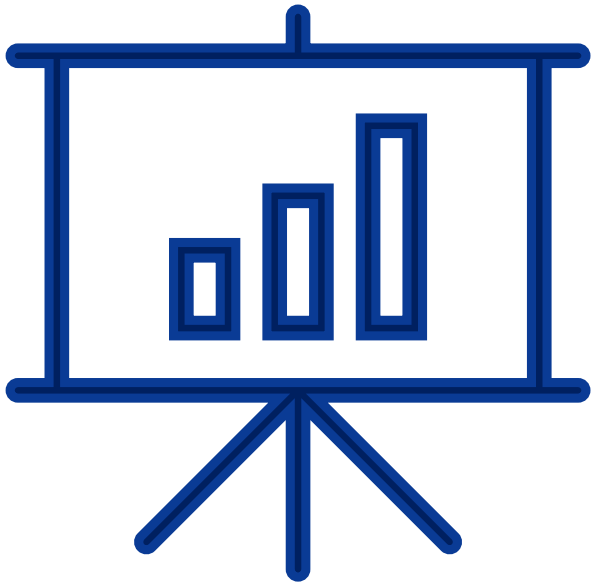
Outputs : pvalues
plots



SURVIVAL ANALYSIS
APPLICATION WITH R

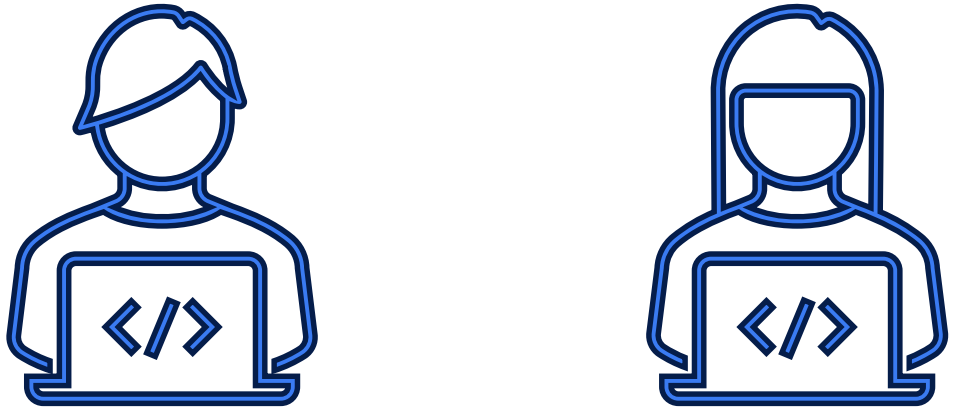
ggforest function (*surminer* package) : Forest-plot of Hazard-ratios

SURVIVAL ANALYSIS



Live demo

SURVIVAL ANALYSIS



Time to play !
(15 minutes)

QUESTIONS

07

THANK YOU FOR YOUR ATTENTION

NOVEMBER 2025