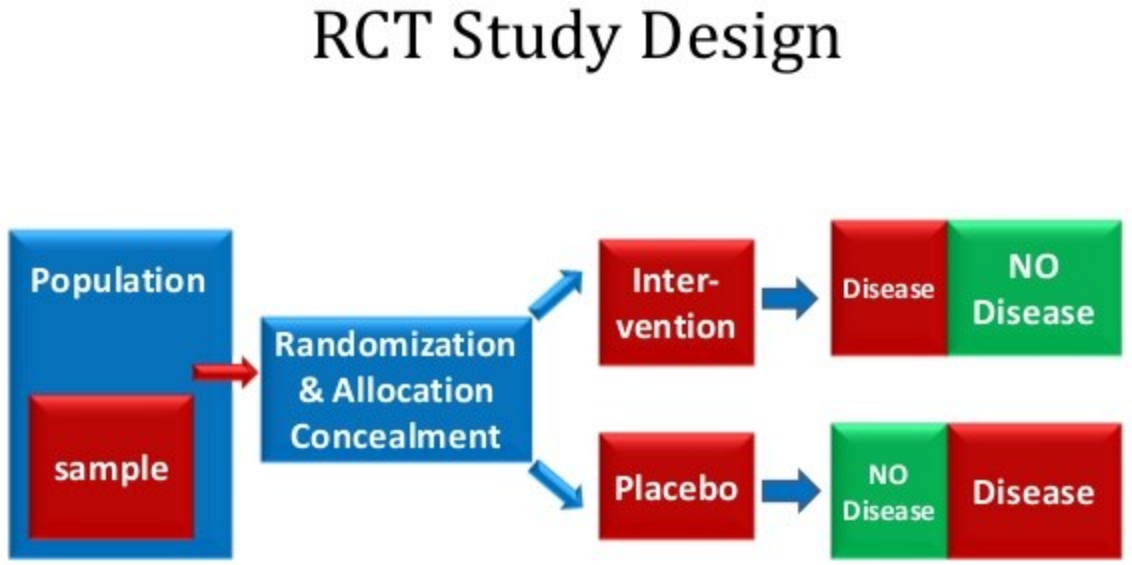
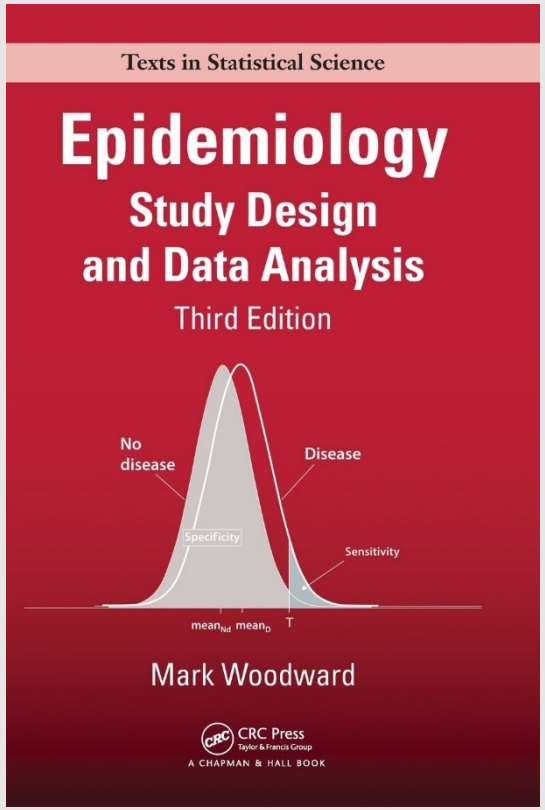
Study Designs in Epidemiology (I)

Randomized Clinical Trials : **RCT**



Science is built of **facts** the way a house is built of bricks: but an accumulation of **facts** is no more science than a pile of bricks is a house.

Henry Poincaré
(1854 – 1912)



To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

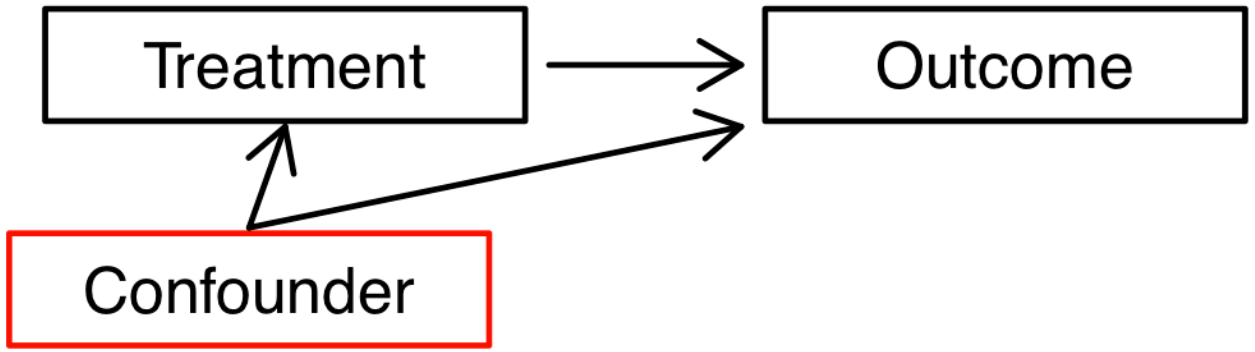
Ronald A. Fisher
(1890-1962)

FOCUS ON *ESTIMATION OF «EFFECT»* STUDIES

(1) Treatment → Outcome

Increasing level of "complexity" of the study design from (1) to (3)*

(2)

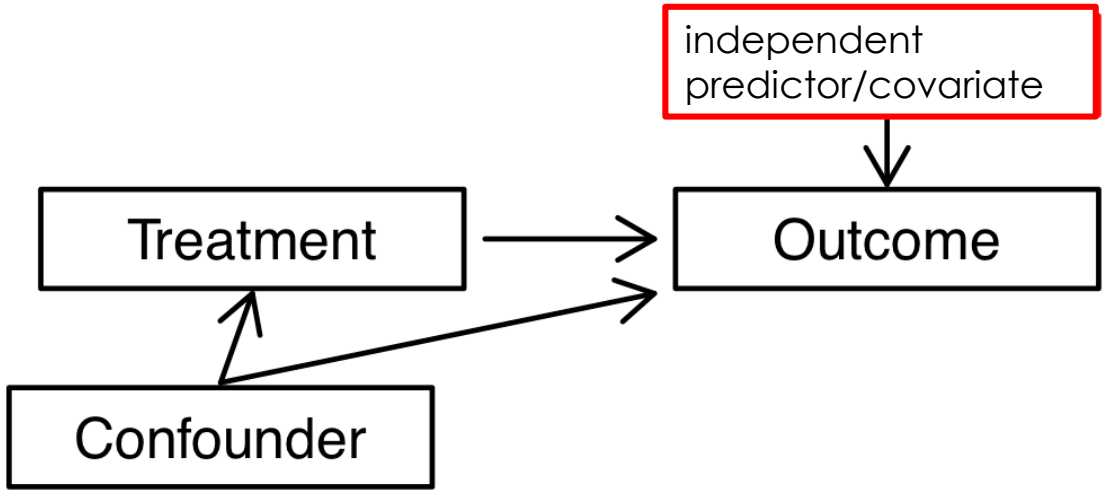


Associated with outcome (but not with the treatment)
ex: age / gender ...

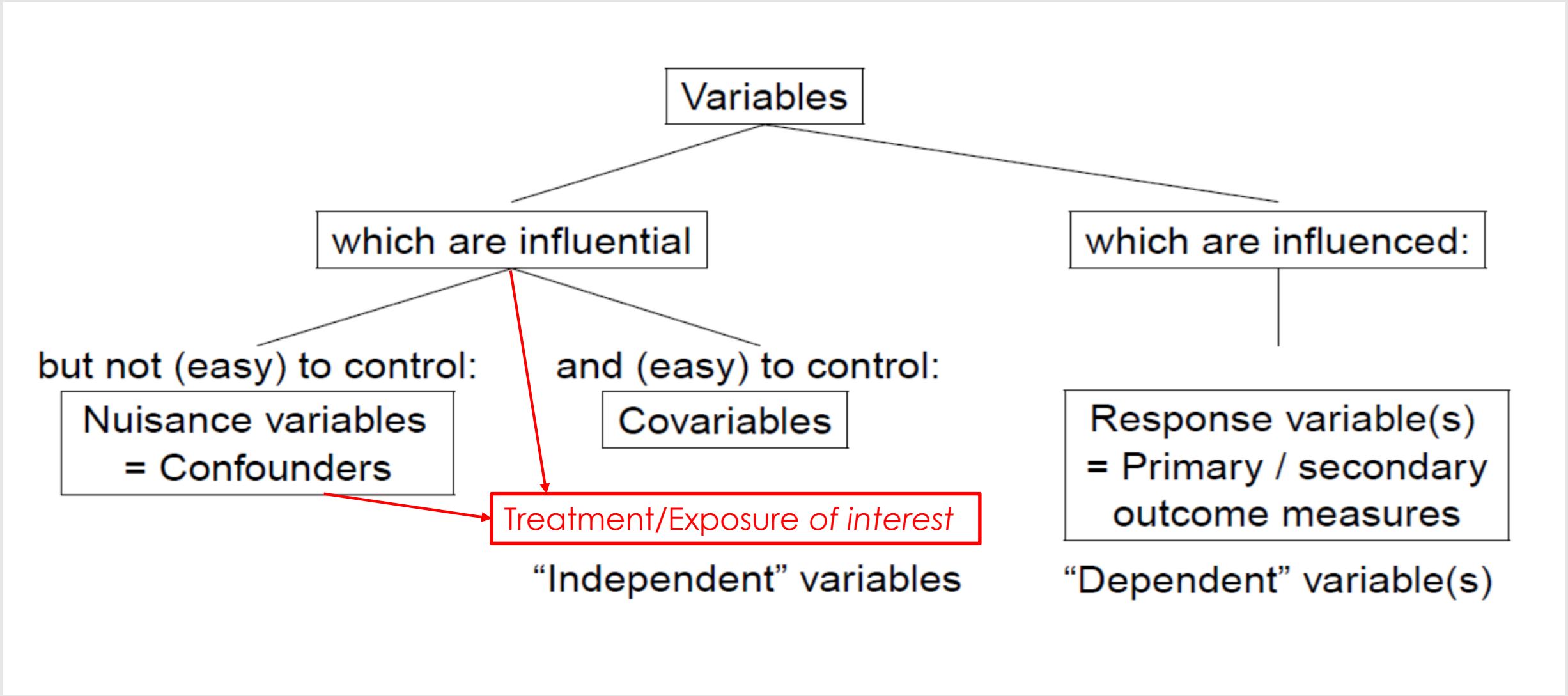
associated with the outcome **and** the treatment:

- clinical indications for the treatment
- Alchool intake w.r.t. exposition to smoke, on oral cancers

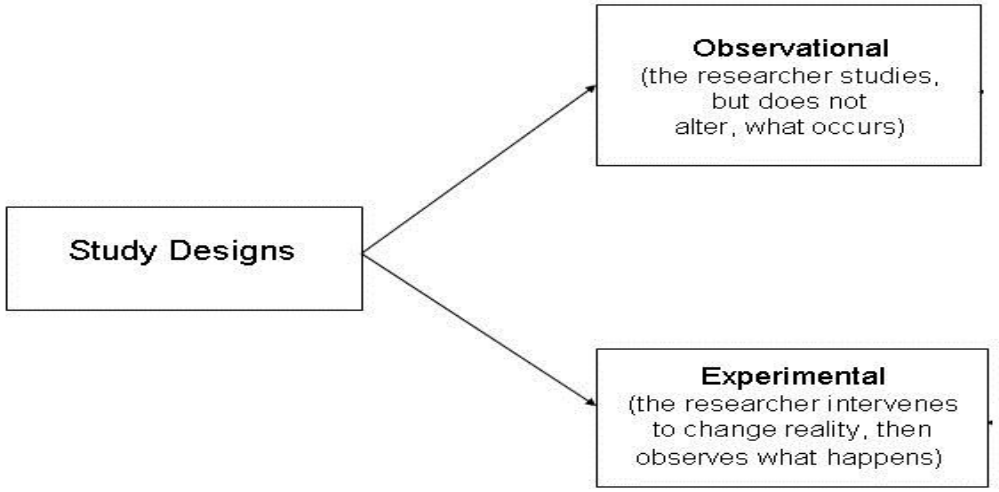
(3)



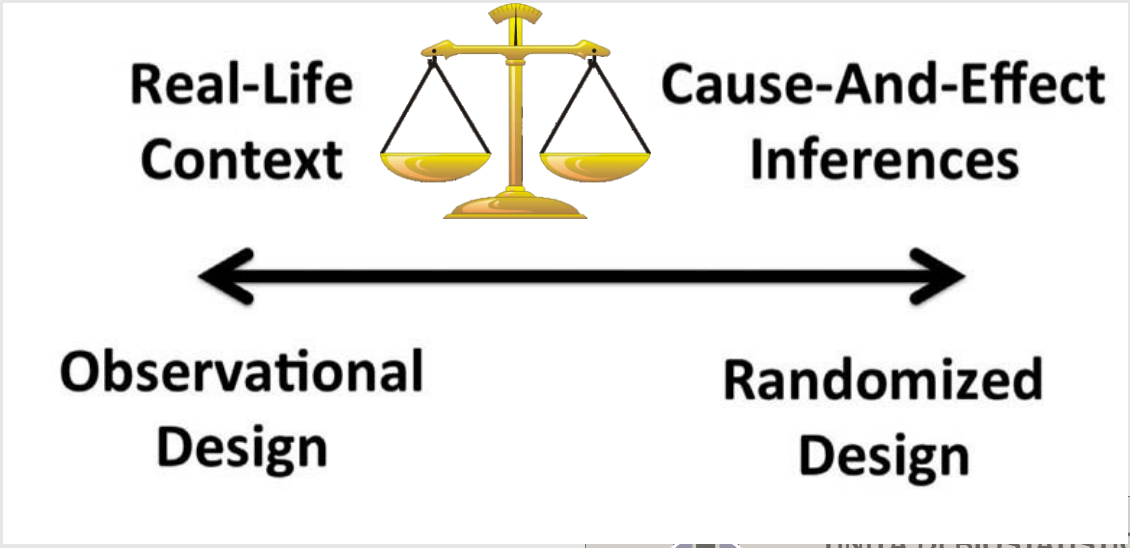
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute


Variables
which are influential
which are influenced:
but not (easy) to control:
and (easy) to control:
Nuisance variables
= Confounders
Covariables
Response variable(s)
= Primary / secondary
outcome measures
Treatment/Exposure of interest
"Independent" variables
"Dependent" variable(s)

What "type" of study : experiment or observation?



The principle of science, the definition, almost, is the following: ***the test of all knowledge is experiment***. Experiment is the *sole judge* of scientific 'truth'.
(R. Feynman,1963)



"Experiment, observation [...] have a **crucial role** for modern therapeutics. ***Arguments about the relative importance of each are an unnecessary distraction***."

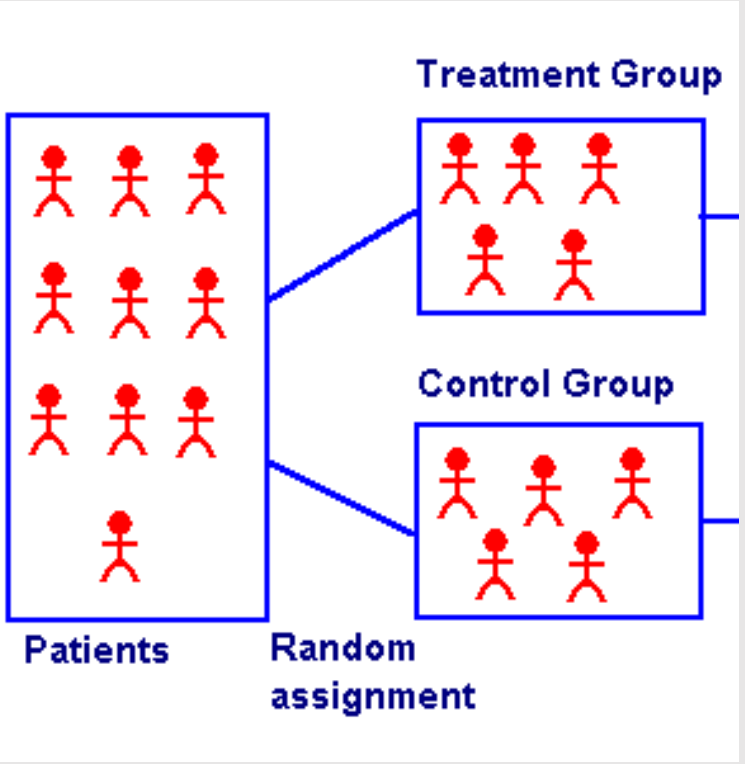
Sir Michael Rawlins, National Institute for Health and Clinical Effectiveness (NICE) ***Lancet* 2008; 372: 2152–61**

Randomized Clinical Trials

RCTs are the **gold standard** in clinical research to compare interventions in humans. The selected population is usually homogeneous with respect to inclusion / exclusion criteria.

To implement an RCT, a number of rules must be respected:

- scientific- methodological
- statistical
- ethical
- legal



The basic **ethical** principle is that of the **clinical equipoise**: uncertainty in the scientific community on the comparative efficacy of the treatments studied in the trial.

RCTs designed to observe outcomes in humans,
under certain experimental conditions created **ad hoc** by the researcher having **randomized*** subjects to the different treatments.

...*note that the general RCT starting sample is not really *random* (criteria...volunteer bias)

Remind: Potential Outcomes Framework

(Rubin-Robins Causal Model)

Each unit (individual) has **two** potential outcomes:

$Y_0(i)$ is the potential outcome had the unit *i* **not** been treated: **control** outcome

$Y_1(i)$ is the potential outcome had the unit *i* been treated: **treated** outcome

Individual treatment effect for subject *i*:

$$ITE_i = Y_1(i) - Y_0(i)$$

Average Treatment Effect*:

$$ATE = E[Y_1 - Y_0] = E[ITE_i]$$

RCTs are the **ideal** study design in which estimate ATE

Potential Outcomes Framework

(Rubin-Robins Causal Model)

$$T_i = \begin{cases} 0 & \text{Untreated} \\ 1 & \text{Treated} \end{cases}$$

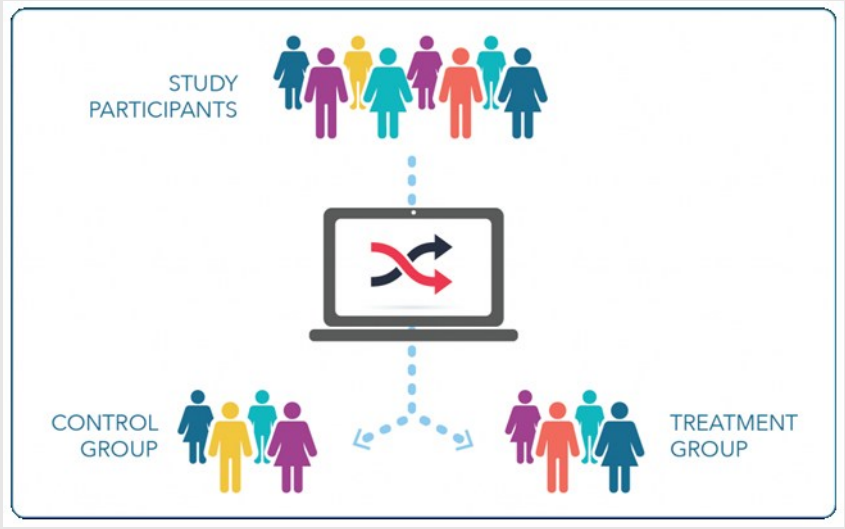
Observed **factual** outcome:

$$y_i = t_i Y_1(i) + (1 - t_i) Y_0(i)$$

Unobserved **counterfactual** outcome:

$$y_i^{CF} = (1 - t_i) Y_1(i) + t_i Y_0(i)$$

one can never **directly** observe **individual** causal effects, because we can never observe **both** potential outcomes for any subject. We need to compare *potential* outcomes, but we only have *observed* outcomes.



Control arm

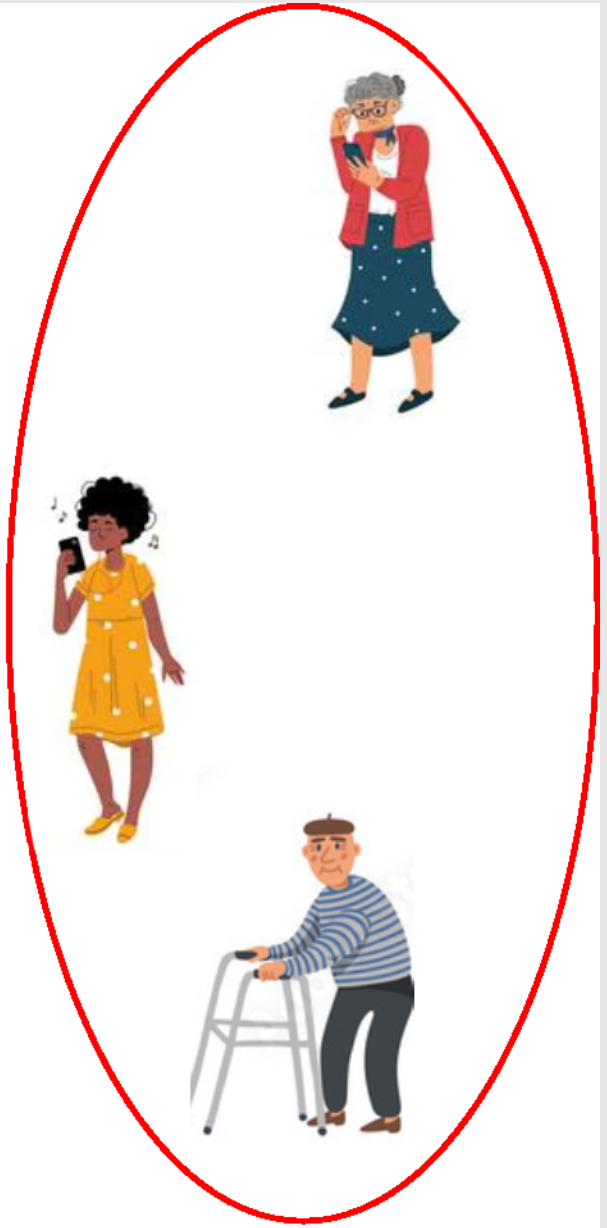
Experimental arm

Exchangeability

Randomization is **fair** with respect to the *potential outcomes* that (on average) **are independent** from the assignment to the treatment (**unbiased** estimator **by design**).

In a randomized experiment the treatment and control group tend to be **similar** in terms of their *observed* and *unobserved* covariates (**exchangeability**).

Under exchangeability: association=causation



Exchangeability

TREATMENT (A)	Y_1	Y_0
1	**100**	*90*
1	**120**	*80*
1	**220**	*170*
1	**200**	*190*
0	*200*	**170**
0	*120*	**90**
0	*220*	**190**
0	*100*	**80**

Observed and **missing** outcomes are *exchangeable*.

$$E[Y_1] = 160 = E[Y|A = 1] \qquad E[Y_0] = 132.5 = E[Y|A = 0]$$

Treatment : policy to increase income
Y : income

RCT at the heart of the implementation of new treatments:

PHASE I: Safety of a pharmacological principle / treatment. Small samples (20-80 pts). MTD = Maximum Tolerated Dose on healthy / sick volunteers; informed consent is crucial.

PHASE II: Therapeutic exploration : sample of volunteers with the disease of interest (100-300 pts). Pharmacokinetic / dynamic assessments, optimal dose, frequency of doses, administration protocols, evaluation of end-points of interest.

PHASE III: Comparative efficacy demonstrate / confirm the efficacy of a treatment and identify any adverse effects in clinical practice (large sample: 500-3000+ pts).

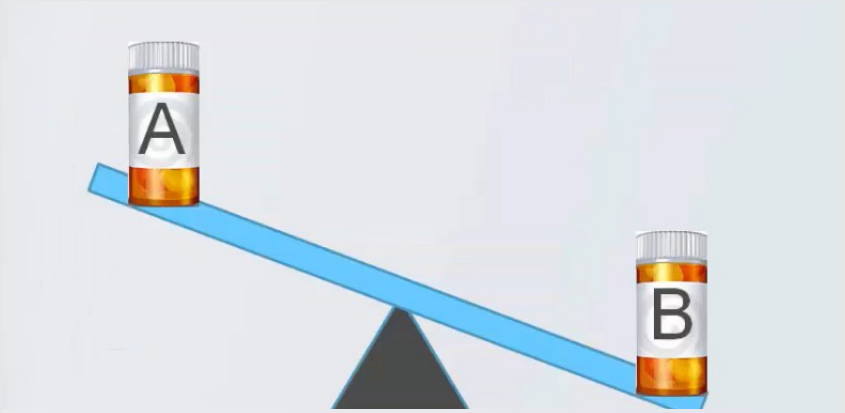
PHASE IV: Post-marketing surveillance: **observational** studies to identify less frequent adverse reactions (longer period); cost-effectiveness in the "real world" population.

about 20% of new drugs are "integrated" by side effects in phase IV studies; about 4% are withdrawn

RCT «Hypothesis Types» (Phase III)*

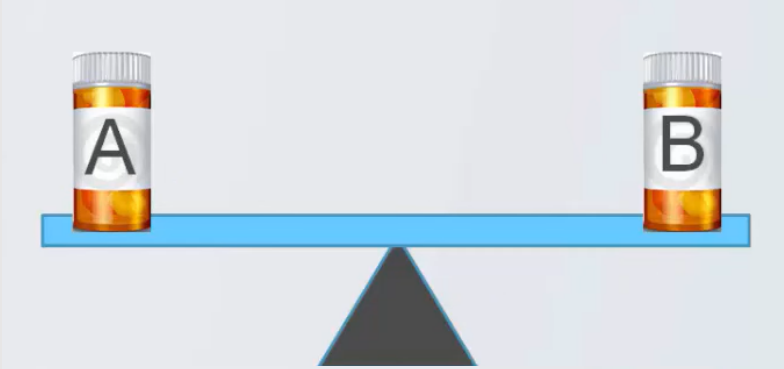
Superiority

$>$



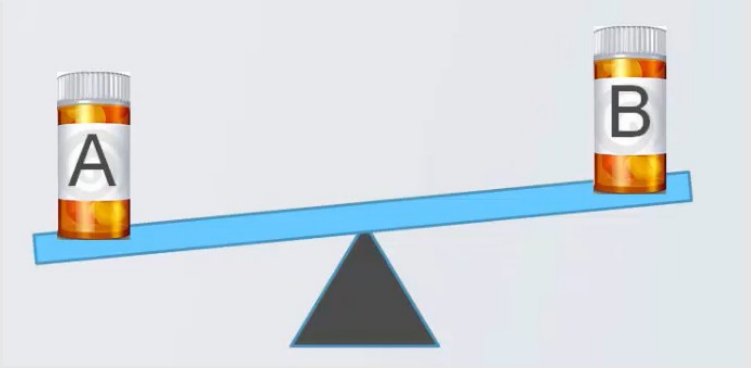
Equivalence

$=$



Non-inferiority

$\not<$



*basic concepts about statistical tests…?

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

The primary objective is to determine the magnitude of increased benefit of the experimental intervention over standard therapy for effectiveness outcomes

Superiority → To show that the treatment is more effective compared to control group

→ Used for establishing new standard of care

- H_0: Treatment *x* is NOT more effective than treatment *y* for given condition
- H_1: Treatment *x* is more effective than treatment *y* for given condition

$$H_0: T_X - T_Y \leq \delta$$

$$H_1: T_X - T_Y > \delta$$

δ is the **minimal clinically relevant «effect size»**

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

2

Equivalence → To show that the treatment is not more effective and is not worse compared to control group

→ Used for establishing generic drugs that are equivalent to their brand formulations

- H_0: Treatment *x* is <u>either</u> worse or better than treatment *y* for given condition by <u>greater than Δ</u>
- H_1: Treatment *x* is <u>NEITHER</u> worse <u>NOR</u> better than treatment *y* for given condition by greater than ±Δ, when Δ is the equivalency margin

$$H_0: |T_X - T_Y| \geq \delta$$

$$H_1: |T_X - T_Y| < \delta$$

δ is called «equivalency margin»

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

The researchers are unconcerned if the experimental treatment is better as long as it is not much worse.

Non-inferiority
- To show that the treatment is not worse compared to control group
- Used for establishing alternative care

- H_0: Treatment *x* is <u>worse</u> than treatment *y* for given condition by <u>greater than Δ</u>
- H_1: Treatment *x* is <u>not worse</u> than treatment *y* for given condition by greater than Δ, when Δ is the non-inferiority margin

δ is called «non-inferiority margin»

$$H_0: T_X - T_Y \leq -\delta$$
$$H_1: T_X - T_Y > -\delta$$

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

The magic of RANDOMIZATION

eliminate systematic confounders , **differences** between groups , allows causal estimates from observed data

- **Simple**: randomly allocates each subject to a treatment independent of previous assignments (head or cross): simple but can create some *imbalances*
- **Block**: force the number of subjects to be allocated to groups of ***k*** in each arm

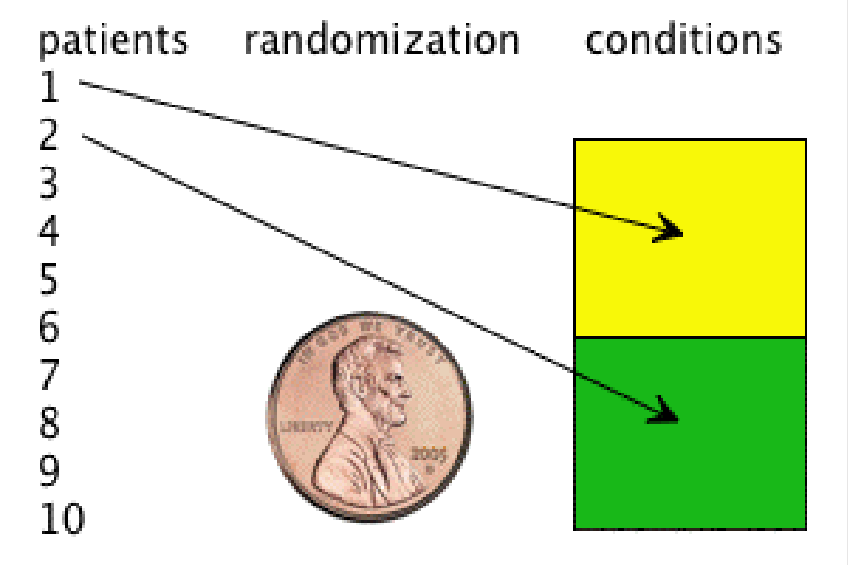
  blocks of size 4 (with two arms) means that in a sequence of 4 subjects 2 subjects are assigned in one arm and 2 in the other [alternate random, in all possible orderings]
  variable size blocks can be used (4, then 8, then 6 ...)

- **Stratification** (**+ blocks**): stratifies the population with respect to specific covariates to be represented (gender, age ...) and then block randomization from each stratum.

Concealment : hiding of enrollment, those who recruit ***do not have to know*** the random sequence of assignments…

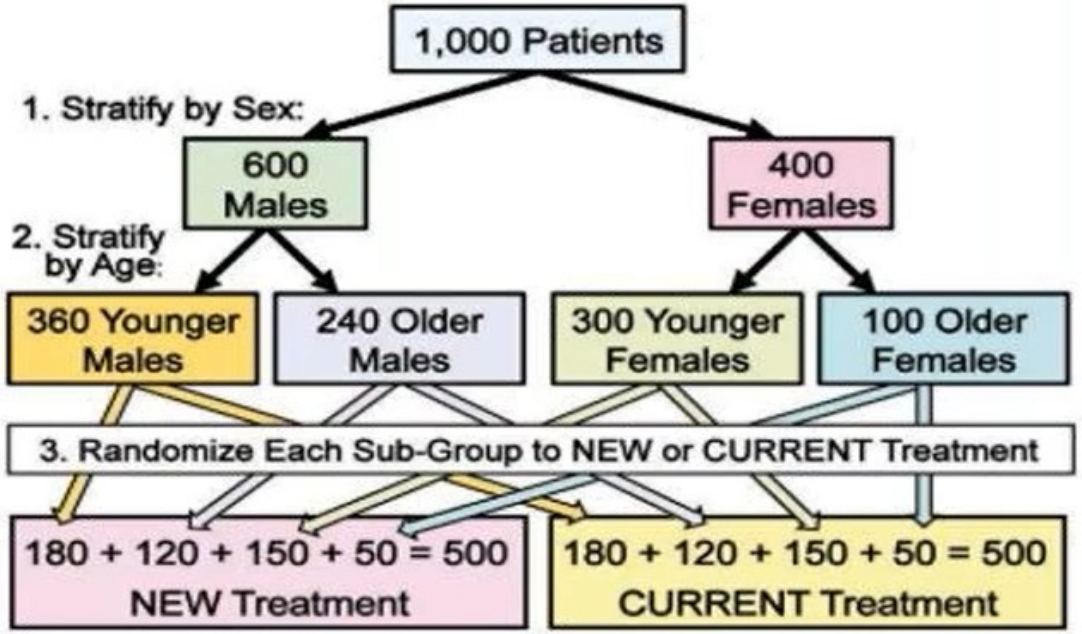
Deciphering the allocation concealment scheme

Block 2.1

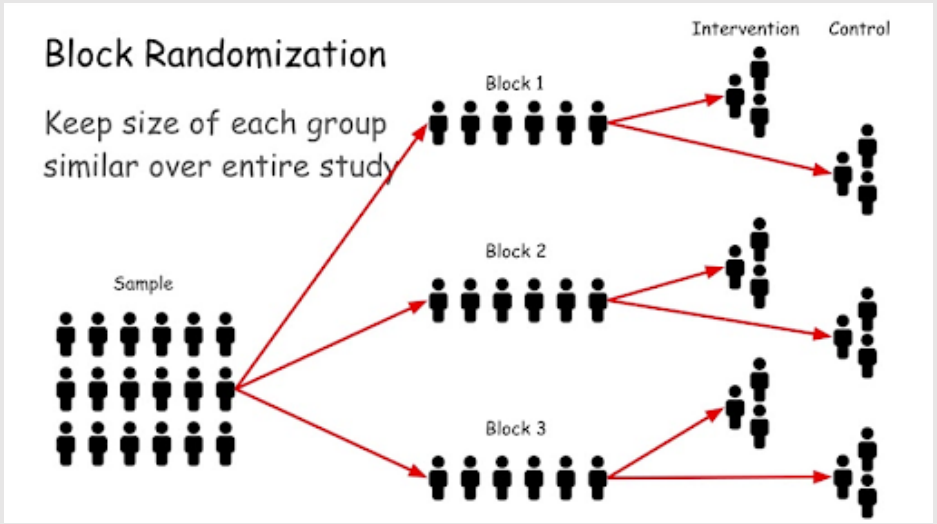


SIMPLE

STRATIFIED



BLOCK



Clinical Trials: BLINDING (MASKING)

To minimize the subjectivity of those who evaluate the pts or analyze the data.

SINGLE BLIND: study subjects do not know which treatment they receive
DOUBLE BLIND: subjects and researchers (doctors / biologists) don't know
TRIPLE BLIND: subjects, the researchers, the statisticians don't know

Figure 1: **The authors: double blinded versus single blinded**

Schulz & Grimes. Lancet 2002

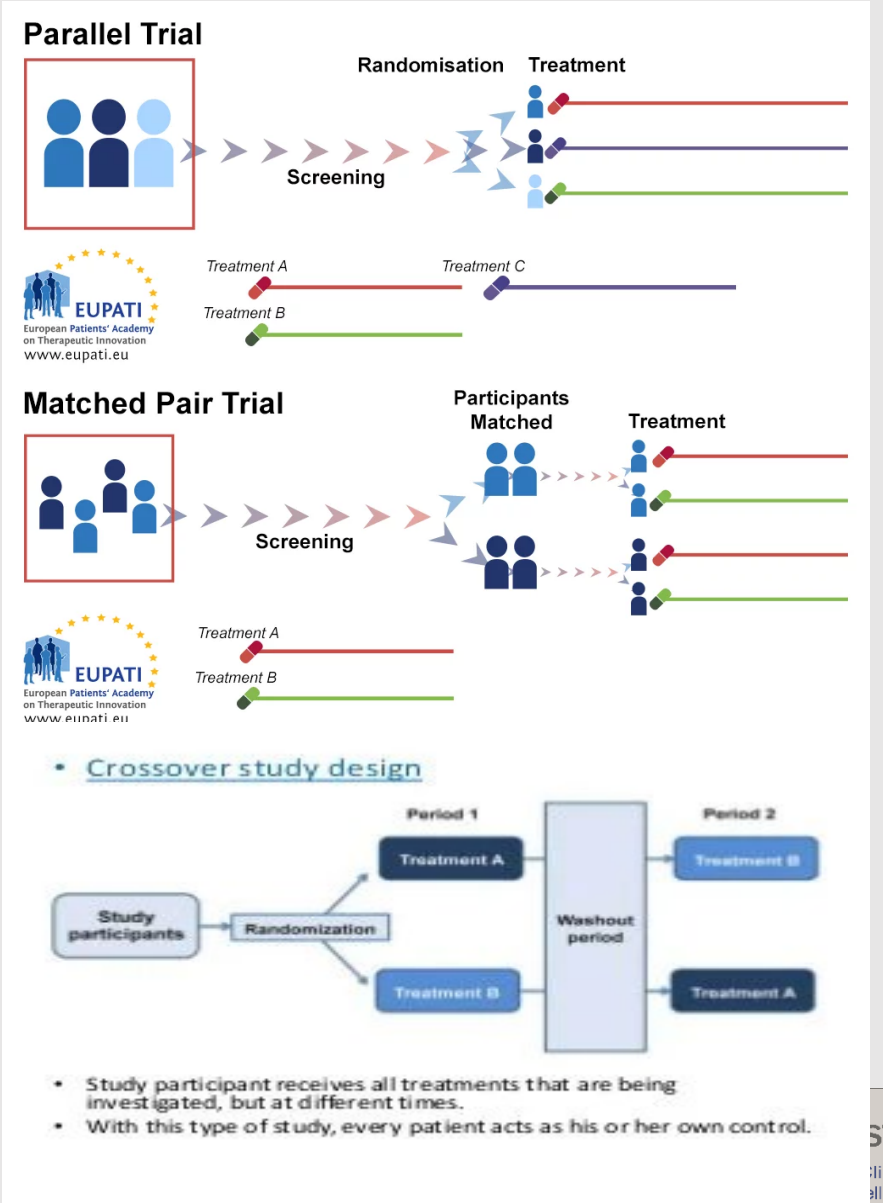
If not done: **OPEN TRIAL**

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

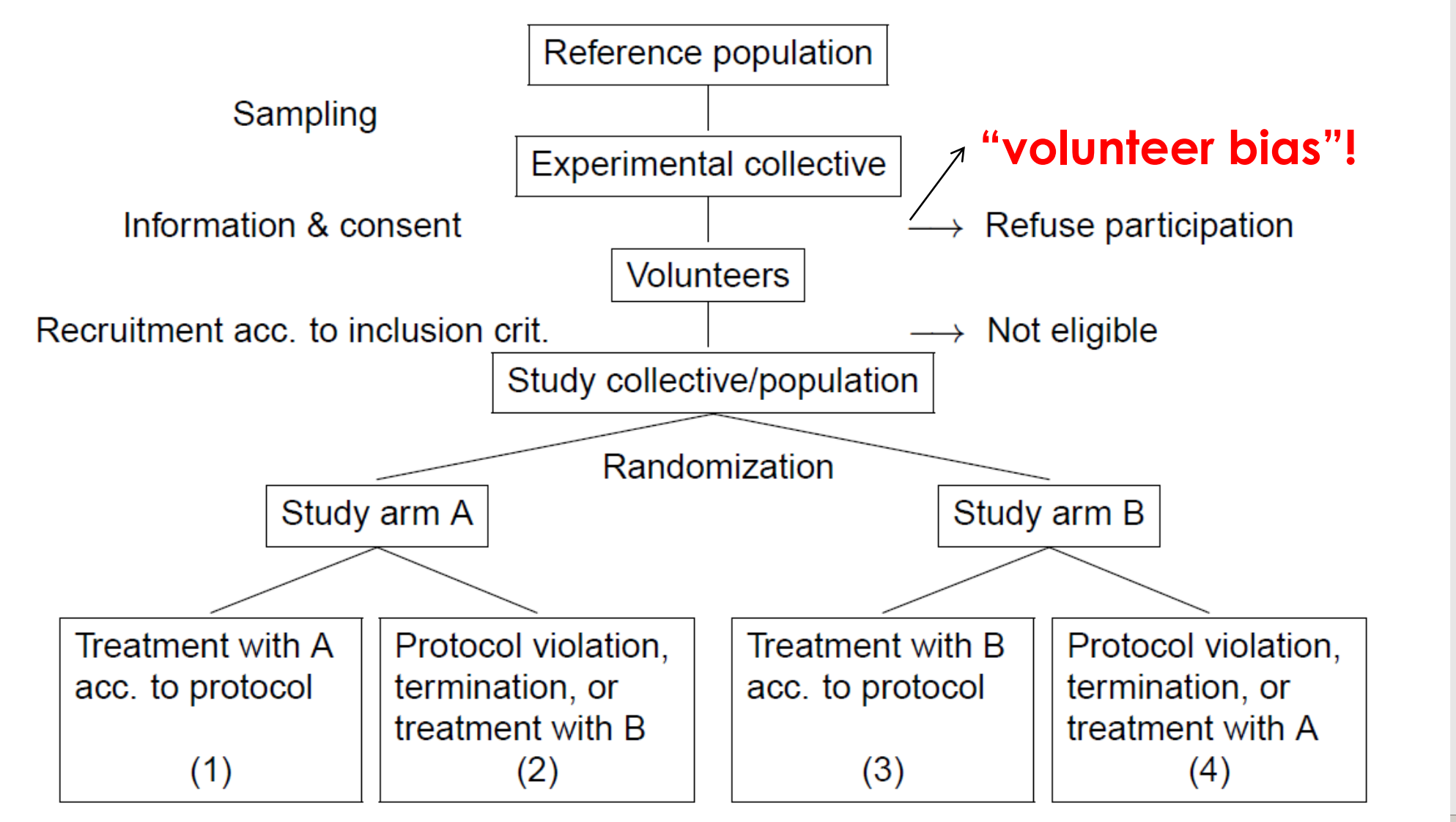
Types of randomised controlled trials

- Parallel
- Matched parallel
- Cross-over
- Sequential
- Cluster
- Umbrella & Basket…

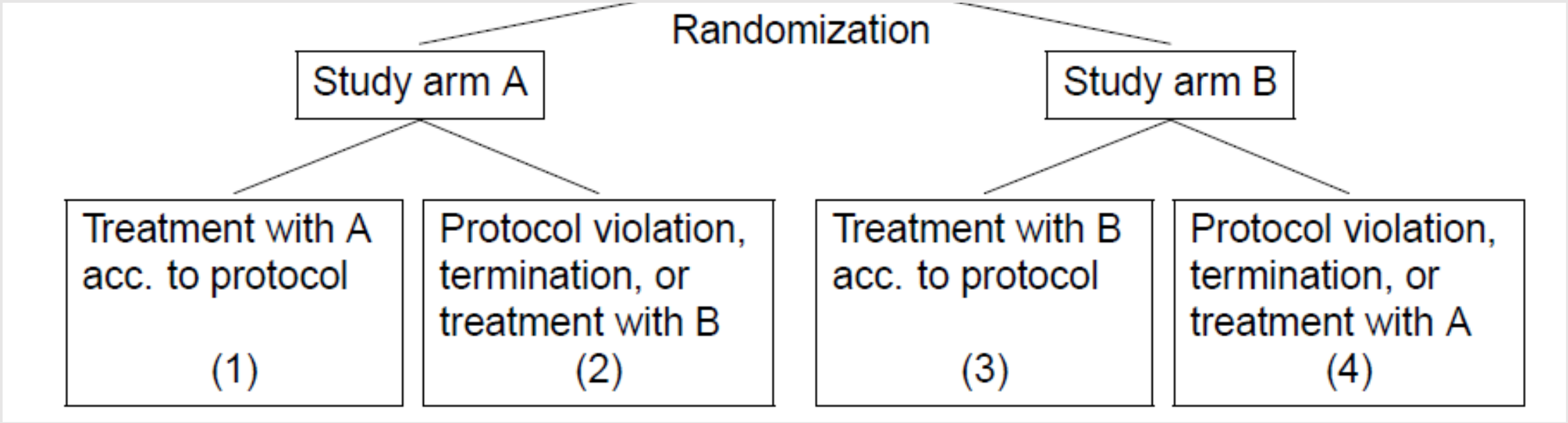
…this is not an exhaustive list!



RCT Flow Chart



RCT data analysis (I)



ITT: intention to treat
(1)+(2) vs (3)+(4)

PP: **per protocol** (1) vs (3)

AT: **as treated** (1)+(4) vs (2)+(3)
(but excluding protocol violations or early drop out)

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

RCT data analysis (II)

Effect estimation + Hypothesis test

1. **Summary trial** results for clinical decision making:

	Ivabradine group (n=3241)	Placebo group (n=3264)	HR (95% CI)	p value
Primary endpoint				
Cardiovascular death or hospital admission for worsening heart failure	793 (24%)	937 (29%)	0·82 (0·75–0·90)	<0·0001

2. **Subgroup** analyses: divide the trial population into groups and examine effects

(Pfizer Trial)

Table 3. Vaccine Efficacy Overall and by Subgroup in Participants without Evidence of Infection before 7 Days after Dose 2.

Efficacy End-Point Subgroup	BNT162b2 (N=18,198)		Placebo (N=18,325)		Vaccine Efficacy, % (95% CI)†
	No. of Cases	Surveillance Time (No. at Risk)*	No. of Cases	Surveillance Time (No. at Risk)*	
Overall	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.0–97.9)
Age group					
16 to 55 yr	5	1.234 (9,897)	114	1.239 (9,955)	95.6 (89.4–98.6)
>55 yr	3	0.980 (7,500)	48	0.983 (7,543)	93.7 (80.6–98.8)
≥65 yr	1	0.508 (3,848)	19	0.511 (3,880)	94.7 (66.7–99.9)
≥75 yr	0	0.102 (774)	5	0.106 (785)	100.0 (–13.1–100.0)
Sex					
Male	3	1.124 (8,875)	81	1.108 (8,762)	96.4 (88.9–99.3)
Female	5	1.090 (8,536)	81	1.114 (8,749)	93.7 (84.7–98.0)
Race or ethnic group‡					
White	7	1.889 (14,504)	146	1.903 (14,670)	95.2 (89.8–98.1)
Black or African American	0	0.165 (1,502)	7	0.164 (1,486)	100.0 (31.2–100.0)
All others	1	0.160 (1,405)	9	0.155 (1,355)	89.3 (22.6–99.8)
Hispanic or Latinx	3	0.605 (4,764)	53	0.600 (4,746)	94.4 (82.7–98.9)
Non-Hispanic, non-Latinx	5	1.596 (12,548)	109	1.608 (12,661)	95.4 (88.9–98.5)
Country					
Argentina	1	0.351 (2,545)	35	0.346 (2,521)	97.2 (83.3–99.9)
Brazil	1	0.119 (1,129)	8	0.117 (1,121)	87.7 (8.1–99.7)
United States	6	1.732 (13,359)	119	1.747 (13,506)	94.9 (88.6–98.2)

RCT data analysis (III)

If trial population has substantial **variation in baseline outcome risk** or there are **interactions between treatment and features : regression models !**

3. **Heterogeneity of Effects (HTE)**

$$P(outcome) = f(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p)$$ Without the intervention

$$lp = \beta_1 x_1 + \cdots + \beta_p x_p$$

$$P(outcome) = f(\alpha + \boldsymbol{\beta_{tx} tx} + \beta_{lp} lp + \boldsymbol{\delta_{int} tx * lp})$$

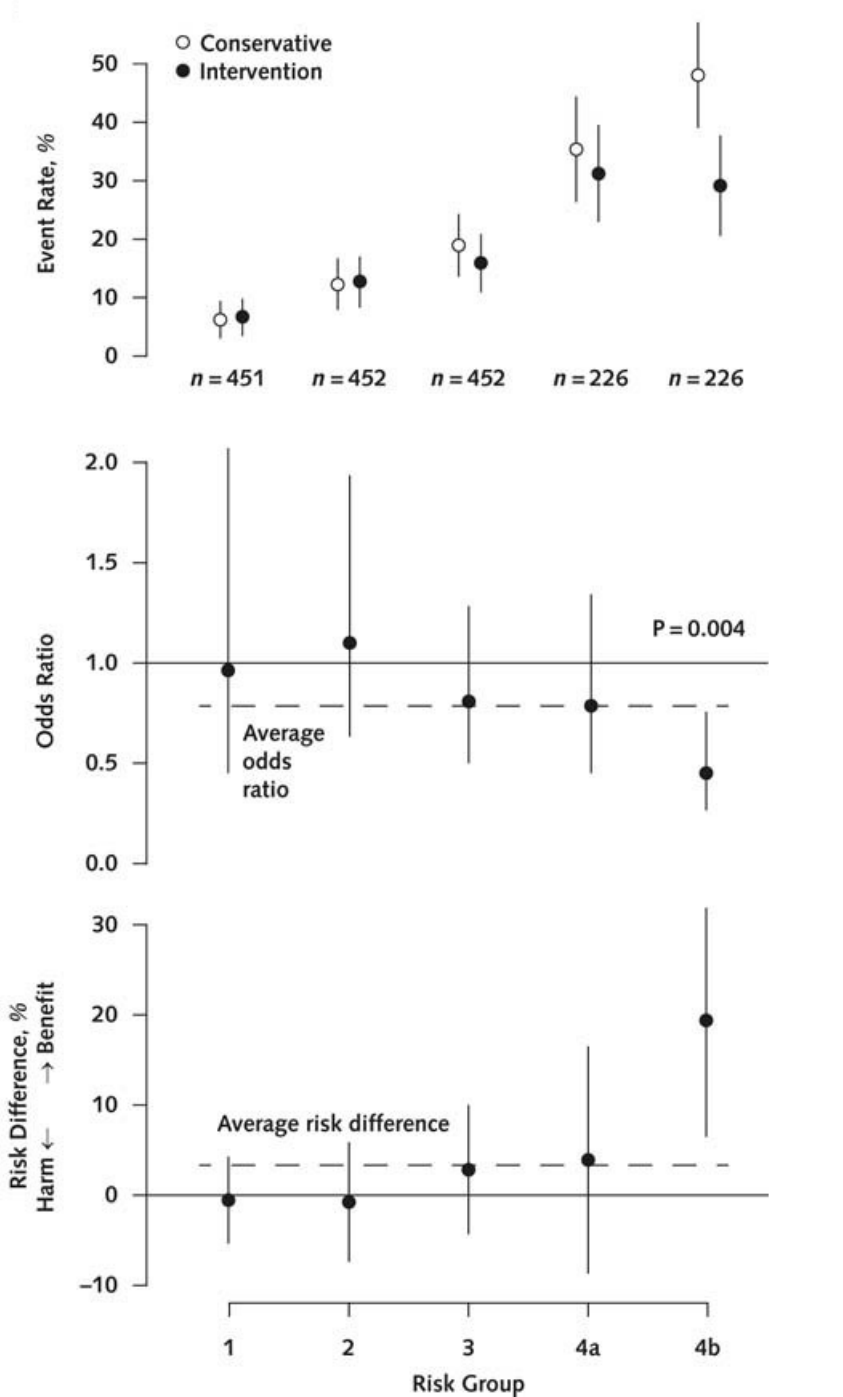
$$P(outcome) = f(\alpha + \boldsymbol{\beta_{tx} tx} + \beta_1 x_1 + \cdots + \beta_p x_p)$$

Risk oriented

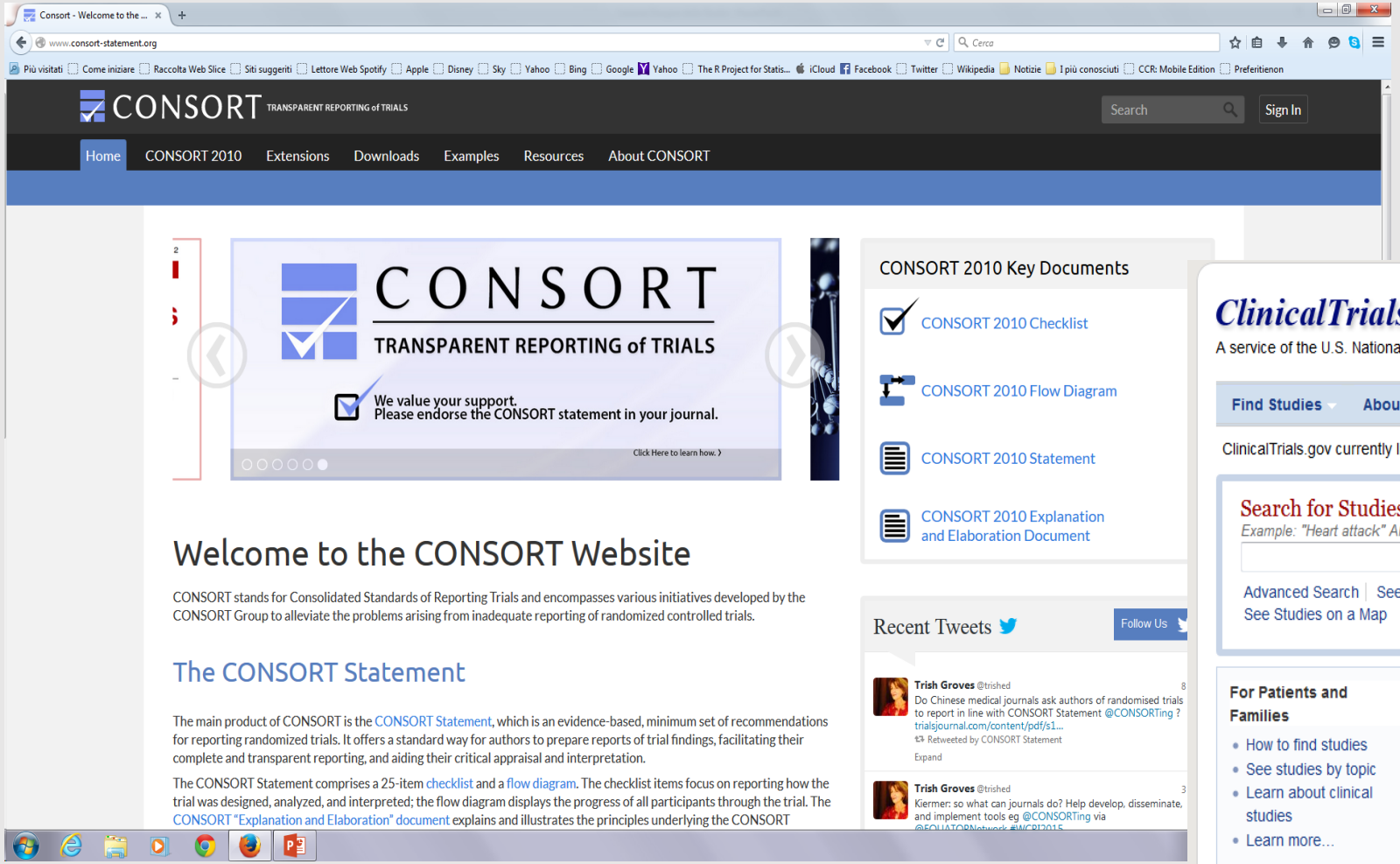
Effect oriented

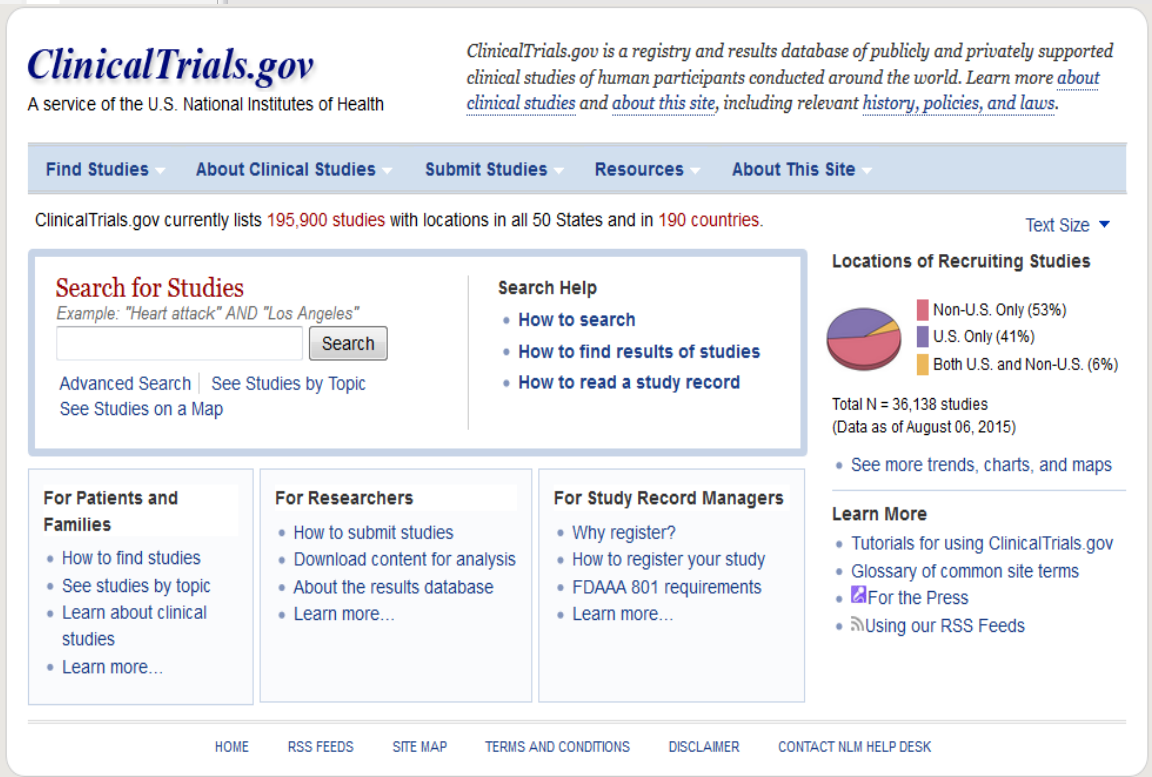
$$P(outcome) = f(\alpha + \boldsymbol{\beta_{tx} tx} + \beta_1 x_1 + \cdots + \beta_p x_p + \underbrace{\boldsymbol{\beta_{1int} x_1 tx} + \cdots + \boldsymbol{\beta_{pint} x_p tx}})$$

Patient-centered estimates of treatment effects: patients have many attributes that simultaneously affect the outcome of interest and the benefits of treatment.



Guidelines for carrying out an RCT: it also includes a **25**-point checklist to be filled in when reporting the results





RCTs: limitations (I)

- RCTs are performed in **selected populations** of pts usually for **short periods** of time.

- In clinical practice, intervention / treatment is generally applied in a **heterogeneous** population of pts - often with multiple comorbidities - and usually for **longer** periods.

- RCTs can describe the most common and early adverse reactions*, but they could be unable to identify those that are less common or have longer latency.

* 500-3000 subjects: 6/1000 - 1/1000 adverse events (95% CI)

	Potential problems
Patients	
Age	Effectiveness in younger or older patients
Sex	Effectiveness generally
Severity of the disease	Effectiveness in mild or severe forms of the condition
Risk factors	Effectiveness in patients with risk factors for the condition (eg, smokers)
Comorbidities	Influence of other conditions on effectiveness
Ethnicity	Effectiveness in other ethnic groups
Socioeconomic status	Effectiveness in disadvantaged patients
Treatment	
Dose	Too high a dose used in RCTs
Timing of administration	Influence on adherence (compliance) to treatment regimens
Duration of therapy	Effectiveness during long-term use
Comedication	Adverse interactions
Comparative effectiveness	Effectiveness in comparison with other products used for the same indication
Setting	
Quality of care	Prescription and monitoring by less specialist (expert) healthcare providers

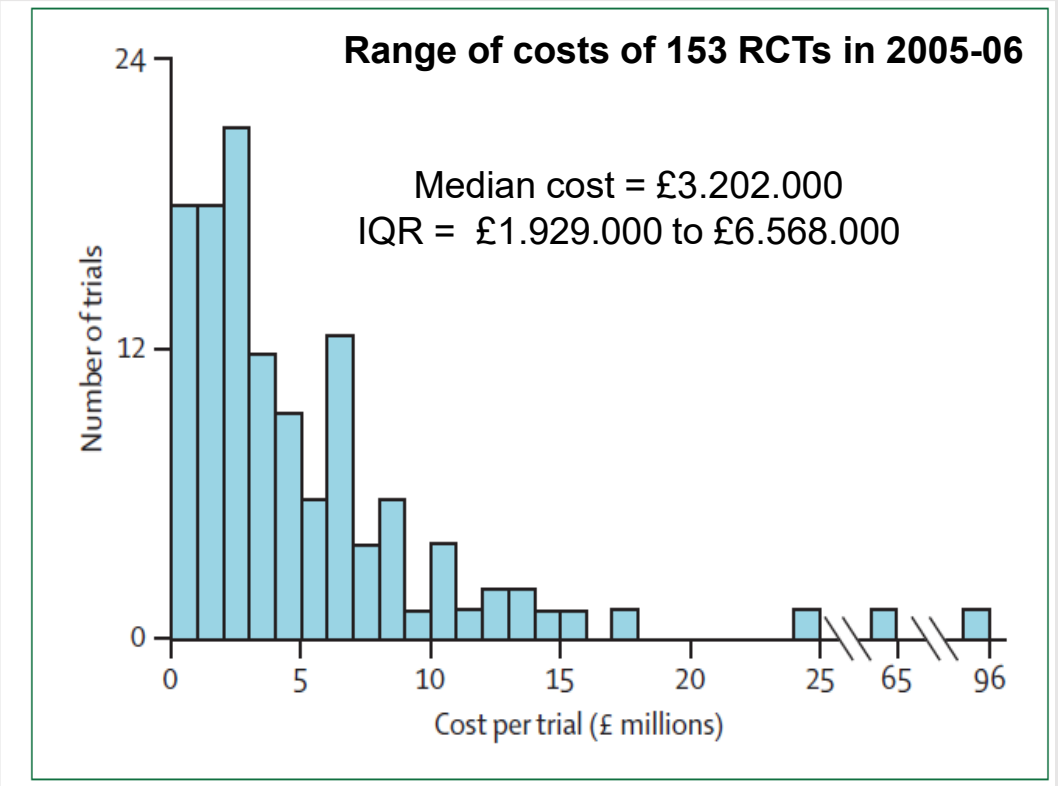
RCTs: limitations (II)

Califf RM. Clinical trials bureaucracy: unintended consequences of well-intentioned policy. *Clin Trials* 2006.

- RCT **costs** are high and show a growing trend

- **Volunteer bias**: study eligibility, compliance, geographical proximity, socio-economic status, health status ...

Homogeneous population selected:
- internal validity (= low variability)
- limit for external validation of results



EFFICACY ≠ EFFECTIVENESS

Success of a treatment in a "artificial" context

Success of a treatment in a "Real world" context

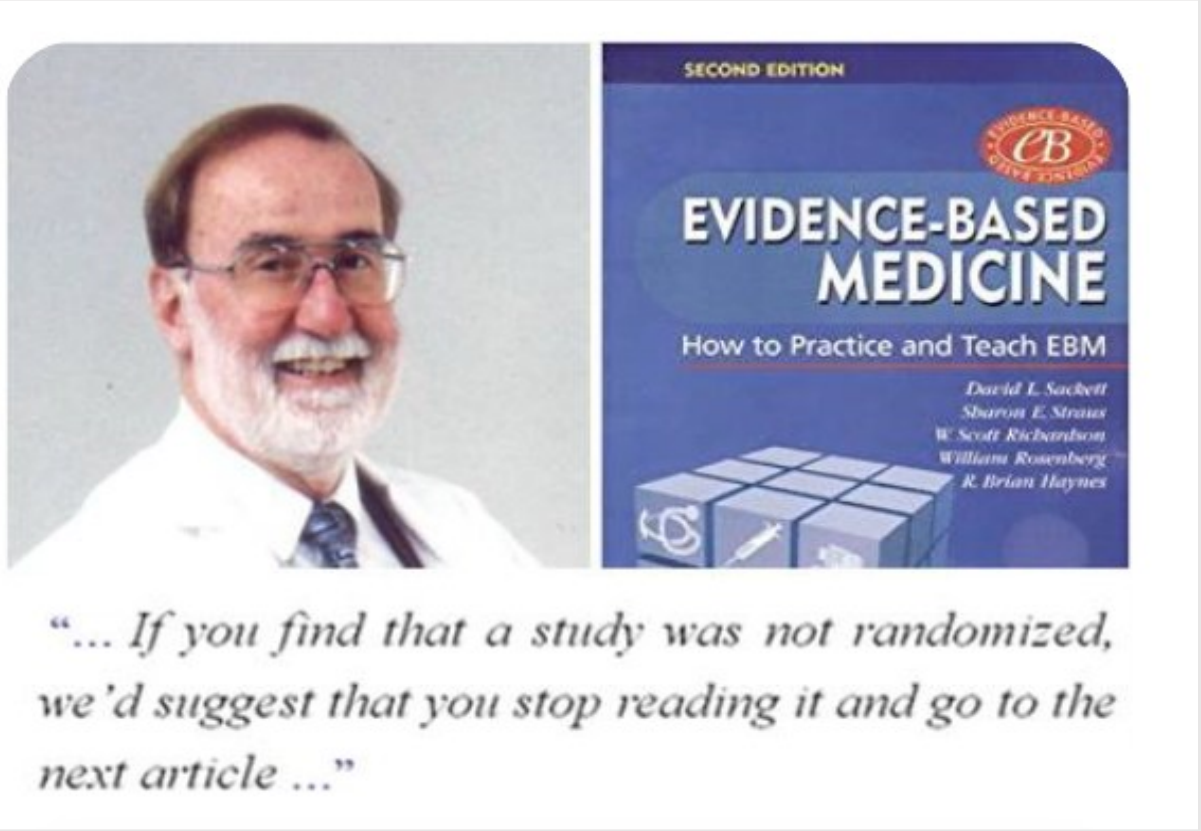
Whilst randomization remains a key design feature for **confirmatory trials aimed at new drug approval**, many innovative non-randomized study designs have been proposed and used in recent years.

The availability of **large databases** of **real world evidence (RWE)** has led to questioning the necessity to carry out expensive and lengthy randomized clinical trials, if **sophisticated statistical techniques** (e.g., **based on causal inference**) could inform the choice between therapeutic options…

Study Designs in Epidemiology (II)

- Experiment ***vs*** Observation [again !]
- Population-based studies
- ***Exposure***-based sampling
- ***Disease***-based sampling

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



David L. Sackett
(1934-2015)

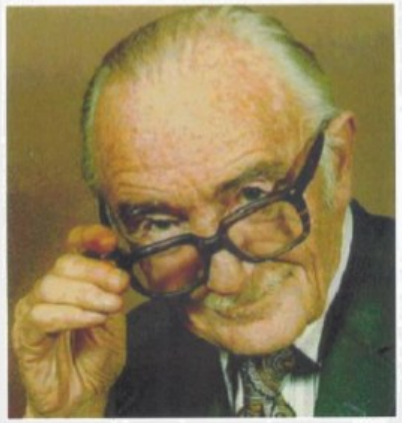
Evidence Based Medicine: How to Practice and Teach EBM, 1997, sold 150.000 copies in English and has been translated into numerous languages.

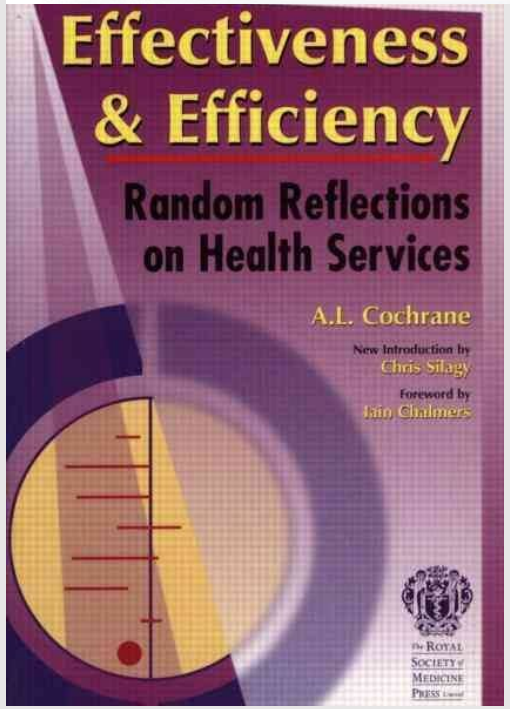
EBM is the conscientious, explicit and judicious use of **current best evidence** in making decisions about the care of individual patients.
The practice of EBM means integrating *individual clinical expertise* + *best available* external clinical evidence from systematic research.

Archie Cochrane

(1908-1988)

"Between measurements based on randomised controlled trials and benefit in the community there is a gulf which has been much under-estimated".





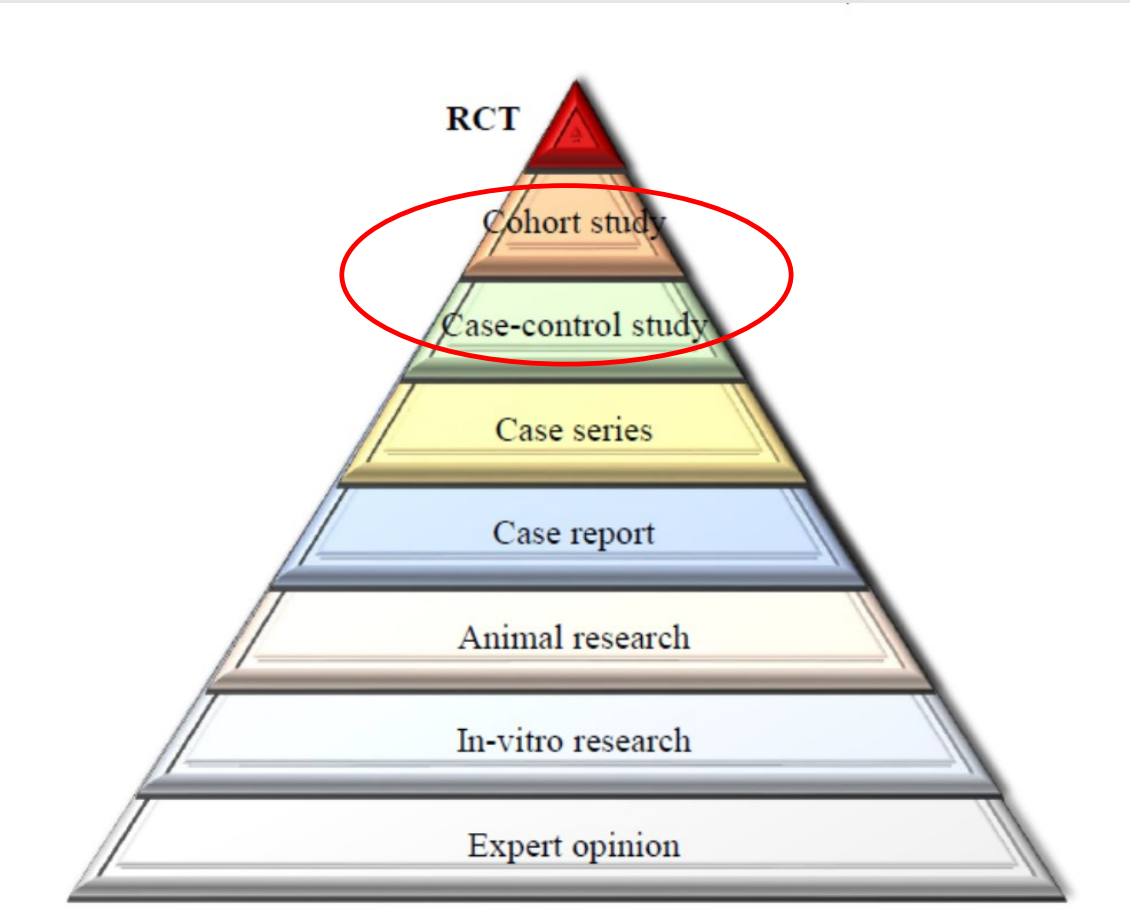
https://www.cochranelibrary.com/

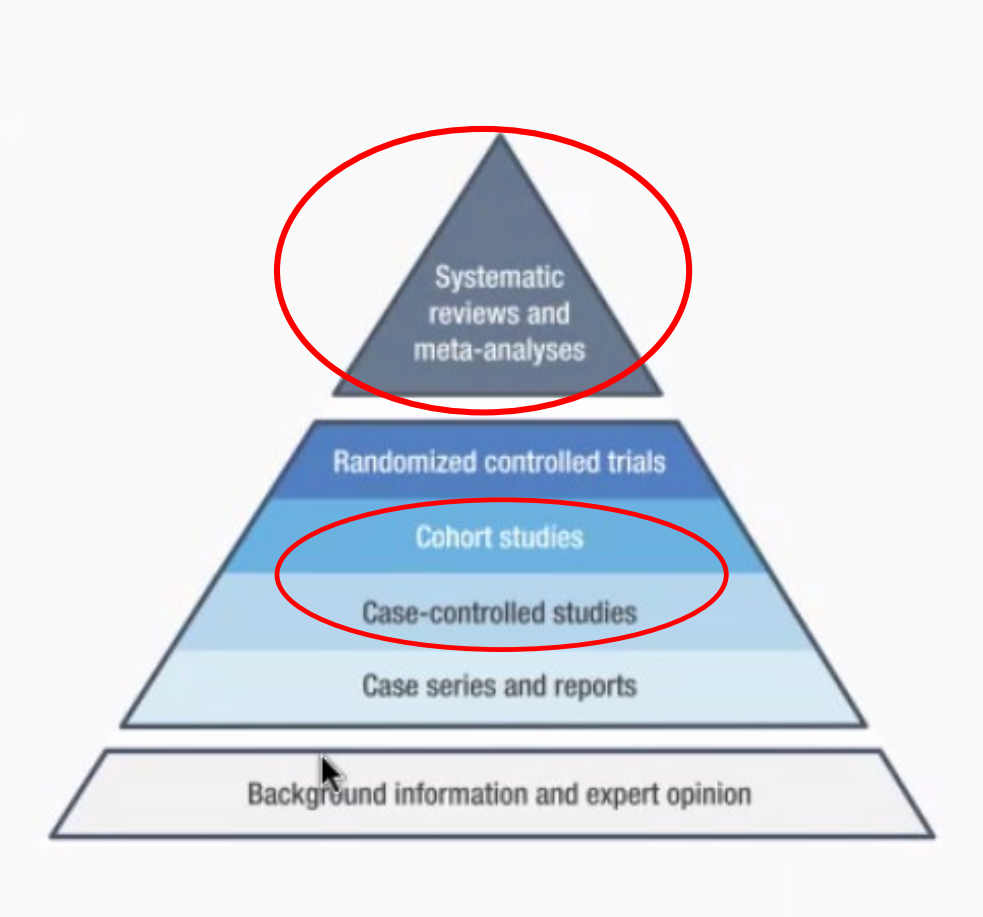
He called for an **international register** of studies, for **explicit quality criteria** for appraising published research.

The Cochrane Library is a **collection** of databases that contain **different types** of high-quality, independent evidence to inform healthcare decision-making (RCTs **and** observational studies).

Evidence-Based Practice: Evidence Pyramid

The top of the pyramid represents the **strongest** [*causal*] evidence.





Observational descriptive studies

A **descriptive** observational study is one that is designed to describe the distribution of one or more variables, without regard to any specific causal hypothesis.

Examples :

- **estimating** the prevalence/incidence of a certain disease in a certain population in a certain time period …in function of age/sex…

- **describe** the distribution of values of a specific biomarker in a population …

Data:

- **Primary source:** collected **ad hoc** by the researcher for the study
- **Secondary source:** **previously** collected for other purposes (e.g. administrative health data)

Observational analytical studies

Observational *analytical* studies analyze the effect of an exposure or a treatment or intervention on subjects.

There is **no randomization to the treatment**; no manipulation by the researcher.

Direct observation of the "real world" ...

Confounders: is there an alternative explanation to the observed results?

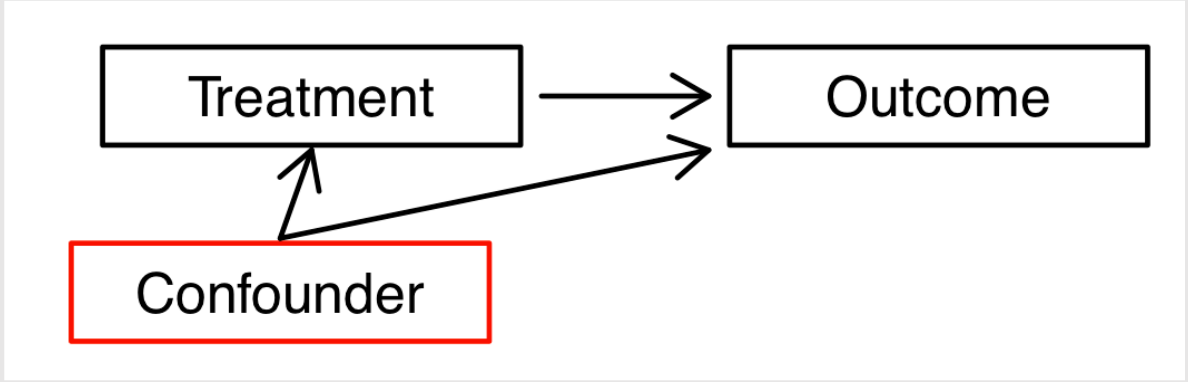
Data:

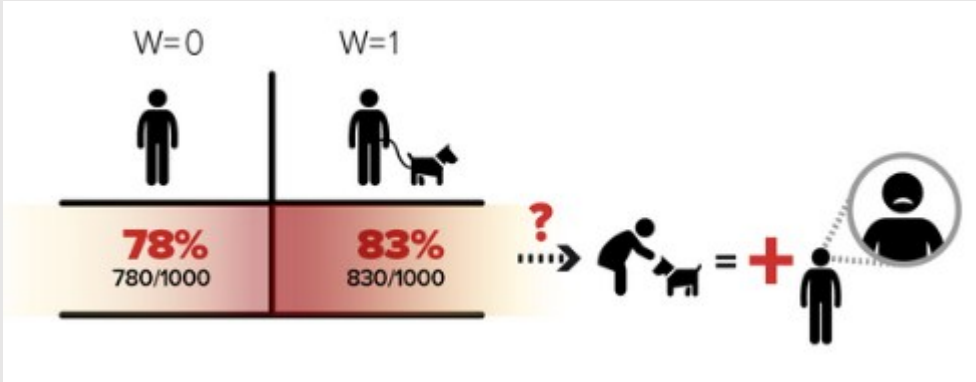
- **Primary source:** collected **ad hoc** by the researcher for the study
- **Secondary source:** **previously** collected for other purposes (e.g. administrative health data)

(In RCTs data are always from primary source)

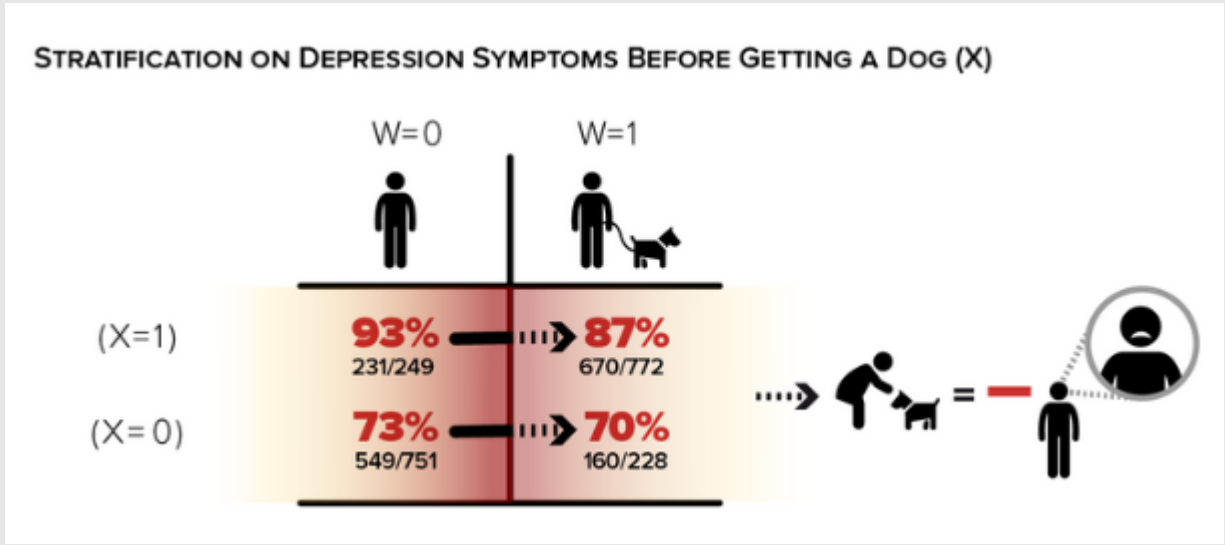
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Confounding is a key challenge when estimating ***causal*** effects from observational data:





W=1 : adopt a dog
W=0 : not adopt a dog
Outcome: severe depression (Y=1/0)



a key potential confounder is the degree of severity of mild/severe depression (X=1/0) **before the dog adoption**…

[Simpson's paradox]

https://hdsr.mitpress.mit.edu/pub/hdzphsk6/release/5

The paradox occurs because people with severe depression symptoms before "treatment assignment" *are more likely* to adopt a dog:

$$P(W_i = 1 \,|X_i = 1) = \frac{772}{772 + 249} = 0.76$$

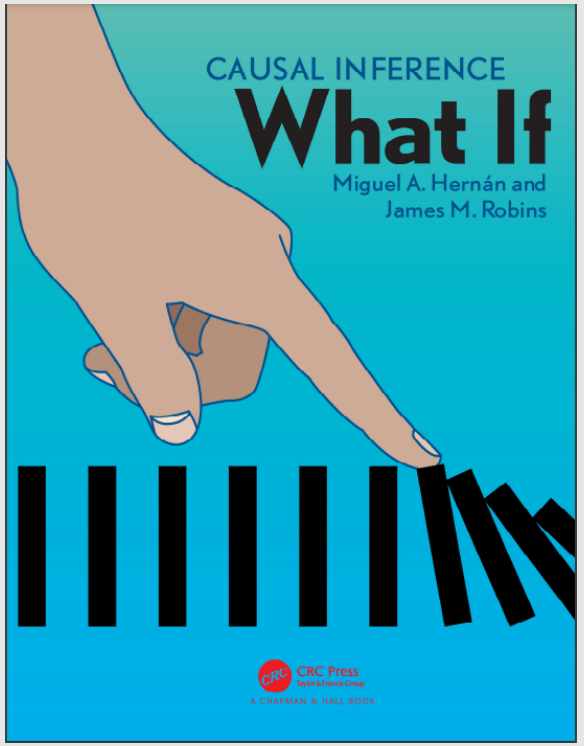
$$P(W_i = 1 \,|X_i = 0) = \frac{228}{228 + 751} = 0.23$$

→ ***propensity*** of adopting a dog conditional to the level of depression symptoms pretreatment

A key feature of RCTs is that **the probability of getting the treatment or the placebo** is ***known*** – under the experimenter's control – and it **does not depend** on (un)observed characteristics of the study subjects.

Solution?

To apply ***causal inference*** approaches to the analysis of data coming from observational studies !



$$P(W_i = \mathrm{a} \,|X_i = \mathrm{b}) = P(W_i = \mathrm{a})$$

measured/unmeasured

Remind: Exchangeability !!

TREATMENT (A)	Y_1	Y_0
1	**100**	*90*
1	**120**	*80*
1	**220**	*170*
1	**200**	*190*
0	*200*	**170**
0	*120*	**90**
0	*220*	**190**
0	*100*	**80**

Observed and **missing** outcomes are exchangeable.

$$E[Y_1] = 160 = E[Y|A = 1] \qquad E[Y_0] = 132.5 = E[Y|A = 0]$$

Treatment : policy to increase income
Y : income

Lack of exchangeability !!

TREATMENT (A)	Y_1	Y_0
1	**100**	
1	**120**	
1	**220**	
1	**200**	
0		**170**
0		**90**
0		**190**
0		**80**

More educated
(higher **pre**-policy incomes)

Less educated
(lower **pre**-policy incomes)

Treatment ***users*** are more educated than non-treated subjects

Treatment : policy to increase income
Y : income

Lack of exchangeability !!

Bad estimate !
Observed too high…

Bad estimate !
Observed too low…

TREATMENT (A)	Y_1		(Before)	Y_0
1	**100**		*90*	*120*
1	**120**		*80*	*110*
1	**220**		*170*	*200*
1	**200**	*(Before)*	*190*	*220*
0	*170*	*200*		**170**
0	*90*	*120*		**90**
0	*190*	*220*		**190**
0	*70*	*100*		**80**

More educated
higher **pre**-policy incomes

Less educated
lower **pre**-policy incomes

$$E[Y_1] = 145 \neq E[Y|A = 1] = \mathbf{160}$$

$$E[Y_0] = 147.5 \neq E[Y|A = 0] = \mathbf{132.5}$$

Imagine that we observe **mortality** among subjects who received A versus B in an observational study:

	Death	Alive	Tot
Treat A	240	1260	1500
Treat B	105	445	550
Tot	345	1705	2050

$$P(Death|A) = 16\% < P(Death|B) = 19\%$$

Treatment A seems better than Treatment B

Now: Medical doctors then give us some additional data about the conditions of subjects ***before starting*** the treatment (their ***baseline*** condition).

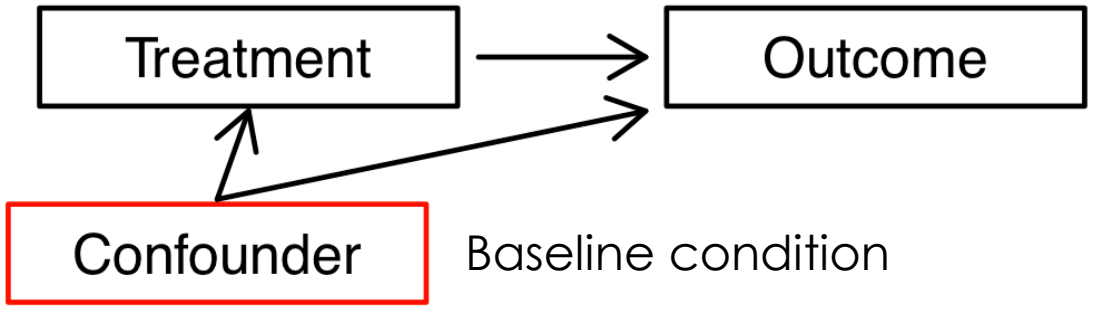
Condition	Death	A	B	Tot
Mild	Yes	210	5	215
	No	1190	45	1235
	Tot	1400	50	1450
Severe	Yes	30	100	130
	No	70	400	470
	Tot	100	500	600
Total		1500	550	2050

If we now compute the probability of death under Treatment A and B as follows:

$$\sum_C P(D|A,c) * P(C=c|A) = \frac{210}{1400} * \frac{1400}{1500} + \frac{30}{100} * \frac{100}{1500} = 0.16$$

$$\sum_C P(D|B,c) * P(C=c|B) = \frac{5}{50} * \frac{50}{550} + \frac{100}{500} * \frac{500}{550} = 0.19$$

We are ignoring that Mild and Severe condition are *distributed differently* between A and B !



$$CATE_x = E[Y_1 - Y_0 | X = x]$$

Condition	Death	A	B	Tot
Mild	Yes	210	5	215
	No	1190	45	1235
	Tot	1400	50	1450
Severe	Yes	30	100	130
	No	70	400	470
	Tot	100	500	600
Total		1500	550	2050

The **conditional** treatment effect indicate that B is better than A !!

$$P(D|mild, A) = 15\% > P(D|mild, B) = 10\%$$

$$P(D|severe, A) = 30\% > P(D|severe, B) = 20\%$$

How to estimate an **average treatment effect** ??

The g- formula

Condition	Death	A	B	Tot
Mild	Yes	210	5	215
	No	1190	45	1235
	Tot	1400	50	1450
Severe	Yes	30	100	130
	No	70	400	470
	Tot	100	500	600
Total		1500	550	2050

$$\sum_C P(D|A,c) * P(C=c) = \frac{210}{1400} * \frac{1450}{2050} + \frac{30}{100} * \frac{600}{2050} = 0.194$$

All treated with A

B is better than A !

$$\sum_C P(D|B,c) * P(C=c) = \frac{5}{50} * \frac{1450}{2050} + \frac{100}{500} * \frac{600}{2050} = 0.129$$

All treated with B

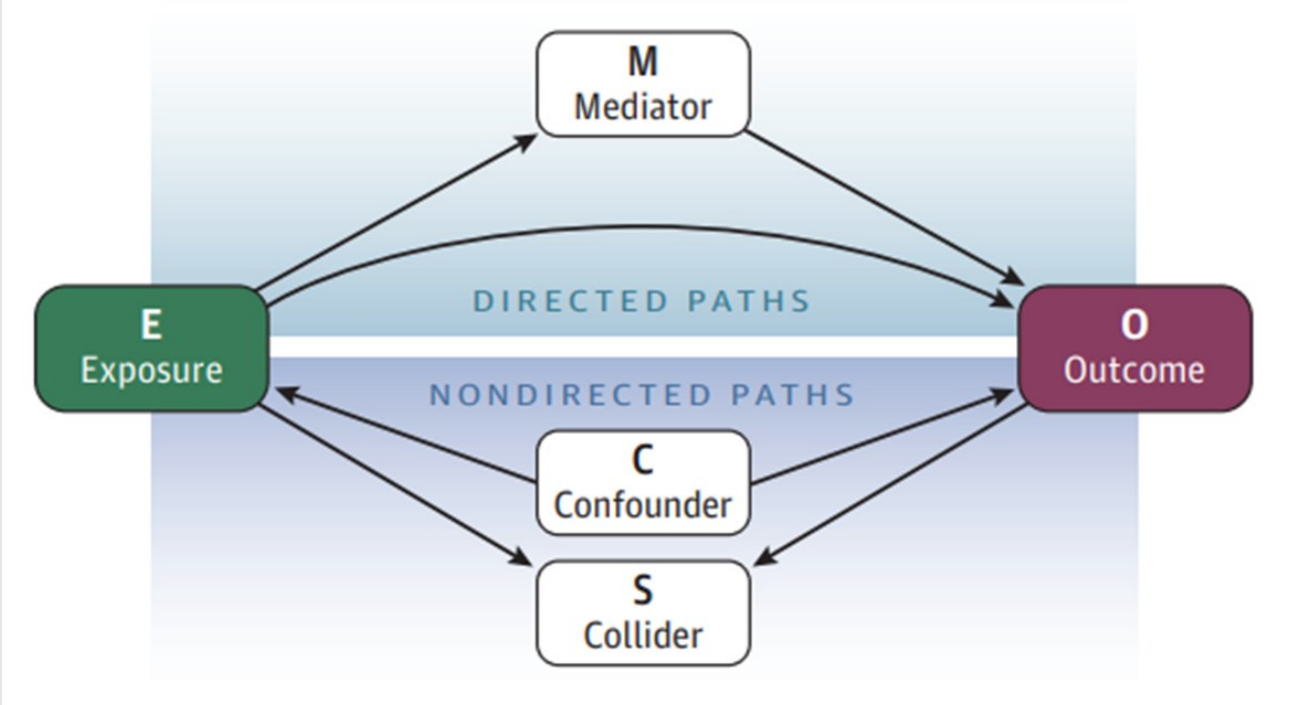
Again: the fundamental problem of causal inference

We will never observe a potential outcome under a condition *other than* the one that actually occurred (**counterfactual**), so that we will never observe an **individual** causal effect. This is the reason why we estimate **average/conditional** treatment effects (**both in RCTs and in observational studies**).

In RCTS **randomization ensure exchangeability**, in observational studies we do not expect it.

In order to obtain (*conditional*) exchangeability, we need to introduce **several assumptions** that essentially embed **subject matter expert knowledge**.

We will see other methods in Block 3 !



We should define **relationships** between variables **before** the statistical analysis is performed…

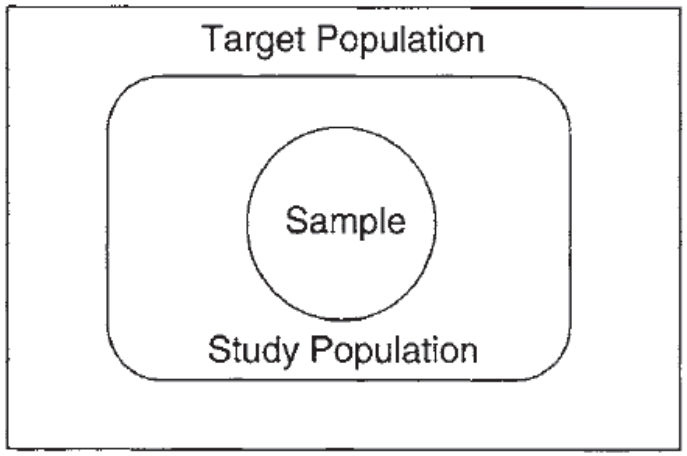
Observational design: (re)-fresh of key-definitions

Target Population: the population to which we would like to apply our estimates regarding the relationship between disease and exposure.

Sometimes, it can be difficult to sample **directly** from the Target Population; in such cases, there is often a ***convenient subgroup*** of the population for which appropriate sampling frames are available.

We call this subgroup the **Study Population**, the population from which ***we are able*** to sample.

Sample : the actual sampled individuals from the Study Population for whom **we collect data**.



If Target Population ≠ Study Population → **selection bias**

Study Population is not *representative* of the Target Population with regard to the disease-exposure relationship of concern.

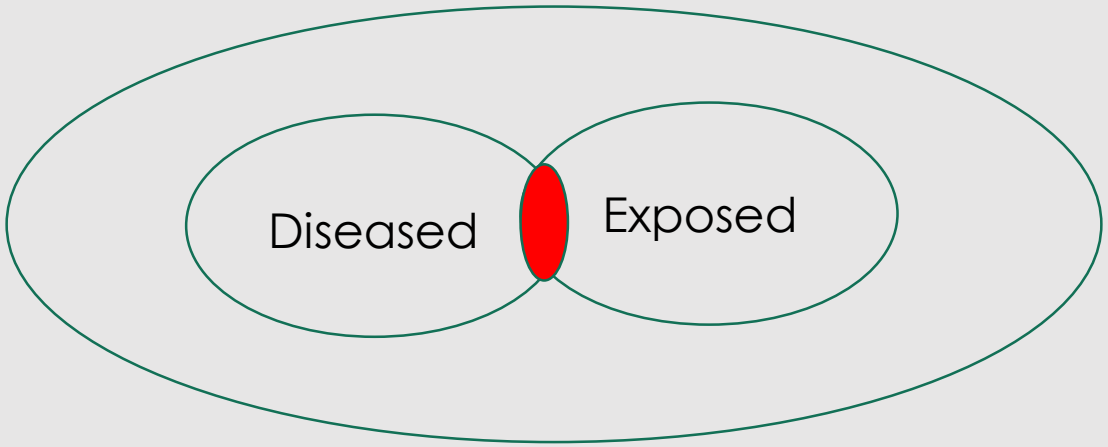
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

If the study sample **is not selected randomly***, we can treat the data in the same manner but without the same (statistical) confidence in the calculations.

Substantial **bias** can be introduced if factors, often *unmeasured* or *unknown*, influencing the sample selection are associated with exposure and disease.

How do we usually obtain a *random* study sample from the Study Population?

3 forms of *sampling schemes* are most commonly used in observational studies



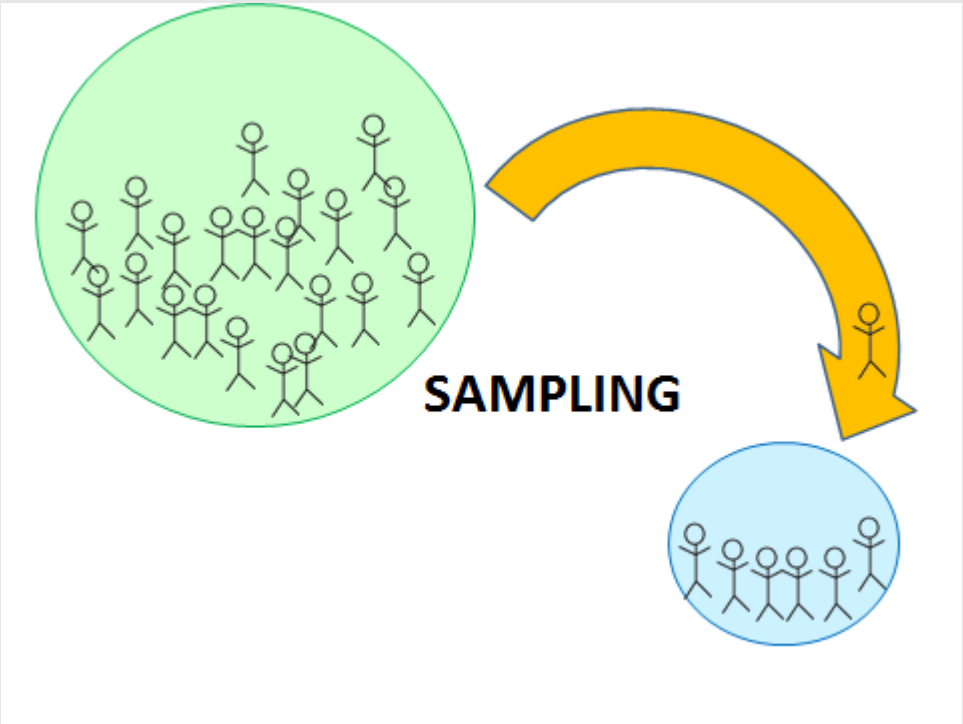
* Details about sampling processes in the R examples in Moodle !!!

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

1. ***Population**-based* studies

2. ***Exposure**-based* sampling

3. ***Disease**-based* sampling



	D	Not D
E	a	b
Not E	c	d

Population-based studies

The main steps of a population-based design are:

1. Take a **simple random sample** of size *n* from the Study Population

2. Subsequently, measure the presence/absence of both D and E **for all sampled** individuals

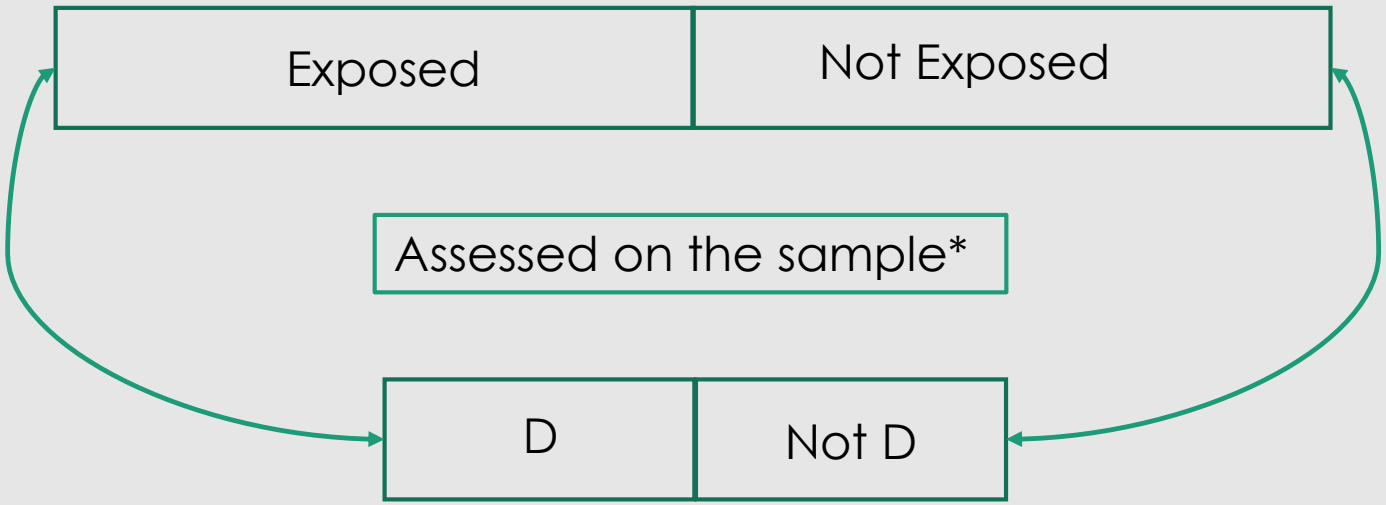
"*subsequently*" refers to the order of **a) sampling** individuals and **b) measuring** the factors, D and E.

A further **classification** in the observational studies is often used to differentiate the **timing** of measurements on D and E:

- A **prospective** study as one in which *measurement of exposure* is made on an individual ***prior*** *to the occurrence* of disease (**primary/secondary data source**).

- In a **retrospective** study, *measurement of [past] exposure* occurs ***after*** an individual's disease status has been already determined.

Population-based* studies



Population-based studies measure exposures, confounders and outcome of **a (unique) sample from the** population.

* ***cross-sectional***: information on exposure will be collected by the investigator and **at the same time** information on disease prevalence

* ***cohort***: exposure will be measured by the investigator at ***baseline*** and then in ***follow up*** information on disease incidence is collected

The various types of population probabilities that may be of interest:

- **Joint** probabilities: $P(D\&E), P(D\&\bar{E}), P(\bar{D}\&E), P(\bar{D}\&\bar{E})$
- **Marginal** probabilities: $P(D), P(E), P(\bar{D}), P(\bar{E})$
- **Conditional** probabilities: $P(D|E), P(D|\bar{E}), P(E|D), P(E|\bar{D})$

Each of these probabilities **can be** estimated using data from a population-based sample: estimates are given by the **observed proportion** of the simple random sample that corresponds to the population probability of interest.

Back to our example : the **outcome** now is the low birthweight:

Population-based study of mother's marital status and low birthweight (**SAMPLE!**)		Birthweight		
		Low	Normal	
Marital status at birth	Unmarried	7	52	59
	Married	7	134	141
		14	186	200

The population probability P(D&E) is estimated by the observed proportion of the sample:

$$P(E\&D) = \frac{7}{200} = 0.035$$

$$P(D|E) = \frac{7}{59} = 0.12$$

$$P(D) = \frac{14}{200} = 0.07$$

$$P(D|\bar{E}) = \frac{7}{141} = 0.05$$

$$\widehat{RR} = \frac{7/59}{7/141} = 2.39$$

$$\widehat{OR} = \frac{(7/59):(52/59)}{(7/141):(134/141)} = 2.58$$

We can estimate all these measures from the population-based study sample !

$$\widehat{ER} = \frac{7}{59} - \frac{7}{141} = 0.069$$

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

Exposure-based* sampling studies

Sampling is carried out separately at different *exposure levels*

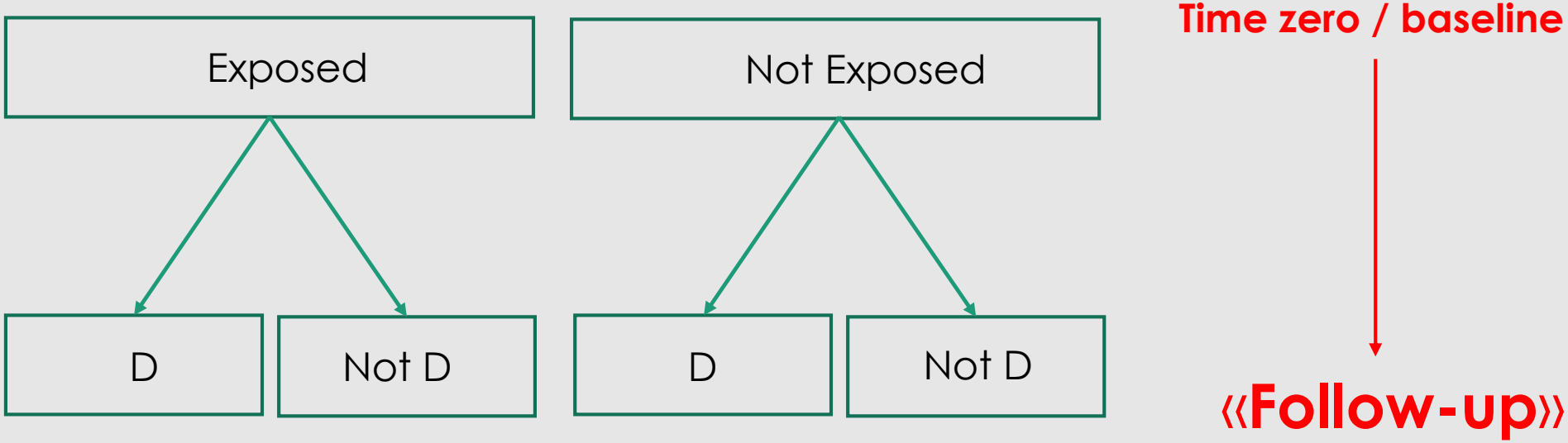
1. Identify **two** subgroups of the population on the basis of the presence or absence of E
2. Take a **simple random sample from each group** (E and $\bar{E}$) separately, of sizes n_E and $n_{\bar{E}}$
3. Measure ***subsequently*** the presence/absence of D for individuals in both random samples

Chronological timing of the two factors D and E are not pertinent to the *sampling* characteristics of the design (but related to *causality* assumptions...)

The **key statistical property** is the **separate** identification and sampling of the exposure groups

*Note that also this design could be also called **Cohort** study

Exposure-based sampling



Data collection could be based on **primary data**, actively follow up study cohort to observe outcomes) or based on *already available* collected data (**secondary data**).

Pre-specify the **sample sizes** for the separate samples taken from the exposure groups.

This division is important in determining the amount of information on the disease-exposure relationship (**sample size** considerations).

For an extreme example: if one exposure group is allocated a **very small** sample size, then there will be little information available on the disease-exposure relationship.

2 random samples*, size 100, from the population of **unmarried** mothers and from **married** mothers.

		Birthweight		
		Low	Normal	
Marital status at birth (Exposure)	Unmarried	12	88	100
	Married	5	95	100
		17	183	200

What quantities can we estimate from these data?

*This design assumes that, prior to sampling, one is able to divide the population by marital status into **two distinct sampling frames**

Exposure-based sampling

1. Joint probabilities **cannot** be estimated: frequencies of joint characteristics are **artificially influenced** by the **pre-specified** number of unexposed/exposed subjects sampled

2. Marginal probabilities **are not estimable** for the same reason

3. Only **conditional** probabilities **that condition on exposure status** can be estimated !!

The conditional probability estimates provide essentially the ***same*** picture as those from the previous population-based study of the same population:

$$P(D|E) = \frac{12}{100} = 0.12$$

$$P(D|\bar{E}) = \frac{5}{100} = 0.05$$

		Birthweight		
		Low	Normal	
Marital status at birth	Unmarried	12	88	100
Exposure	Married	5	95	100
		17	183	200

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

Exposure-based sampling

$$\widehat{RR} = \frac{12/100}{5/100} = 2.40$$

$$\widehat{OR} = \frac{(12/100):(88/100)}{(5/100):(95/100)} = 2.59$$

Conditional probabilities as we know are the basic building blocks of the two most used measure of effect, i.e. the Relative Risk and the Odds Ratio.

These estimates are compatible with those provided by the population-based data from the same population.

		Birthweight		
		Low	Normal	
Marital status at birth	Unmarried	12	88	100
Exposure	Married	5	95	100
		17	183	200

Disease-based sampling – [case-control studies]

A case-control study has the same specifications as an exposure-based sampling study, except that the roles of E and D **are reversed**.

Separate samples are selected from cases (D) and non diseased individuals or **controls** ($\overline{D}$).

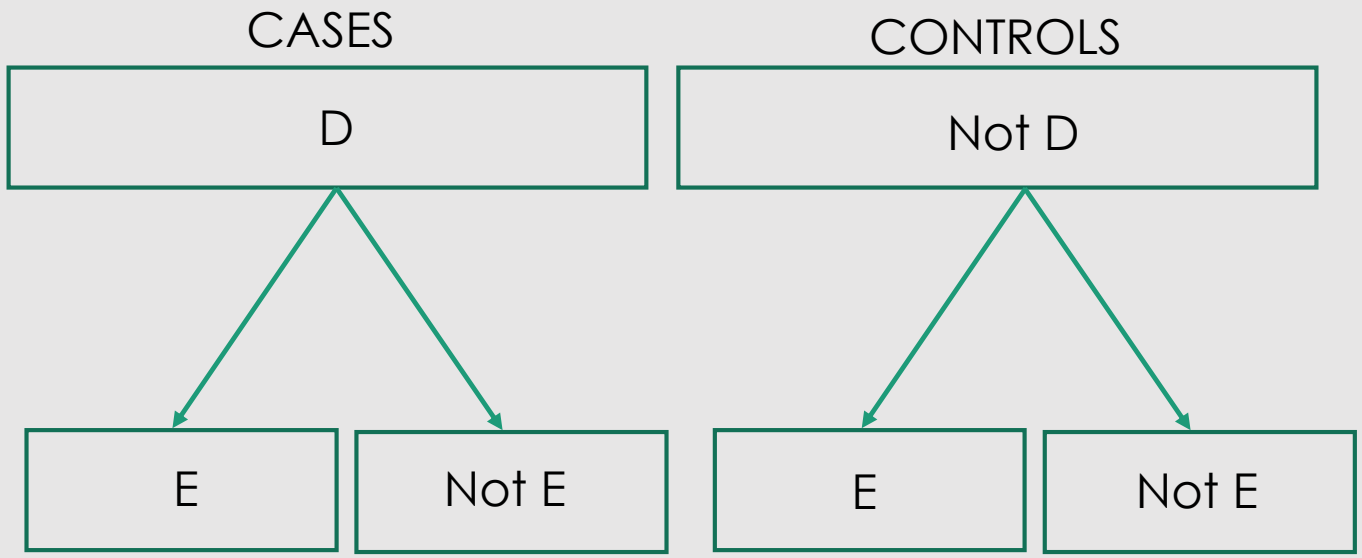
1. Identify **two subgroups** of the population on the basis of the presence or absence of D.

2. Take a simple random sample from each (D and $\overline{D}$) separately, of sizes n_D and $n_{\overline{D}}$

3. Measure subsequently the presence and absence of E for individuals in both random samples.

As for exposure-based designs, the investigator must **pre-specify** the number of cases and controls in the two separate random samples.

Disease-based sampling – [case-control studies]



"classic" case-control design, selecting all cases that accrue in the population in a given time interval and a random sample of those who remain disease free [**exclusive sampling** of controls].

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

Disease-based sampling - case-control studies

Data from a **case-control** study of a mother's marital status and low birthweight

		Birthweight		
		Low	Normal	
Marital status at birth (Exposure)	Unmarried	50	28	78
	Married	50	72	122
		100	100	200

2 sampling frames based on disease presence/absence, are accessible to the investigator

- Joint probabilities **cannot be estimated**: frequencies of joint characteristics are again artificially influenced by the exact allocation of the number cases/controls sampled
- Marginal probabilities **are not available** for the same reason
- Only **conditional** probabilities that **condition on outcome status**, can be estimated !!

$$P(E|D) = \frac{50}{100} = 0.50$$

$$P(E|\bar{D}) = \frac{28}{100} = 0.28$$

At first glance, it seems unlikely that we can estimate *any* measure of association from a case-control design…

Disease-based sampling - case-control studies

		Birthweight		
		Low	Normal	
Marital status at birth (Exposure)	Unmarried	50	28	78
	Married	50	72	122
		100	100	200

This is partly true: **it is impossible to estimate the Relative Risk with case-control data** with exclusive sampling of the controls.

However, we can estimate the **Odds Ratio** for **E associated with D** and for symmetry the **reverse**:

$$OR = \frac{P(E|D)}{P(\bar{E}|D)} : \frac{P(E|\bar{D})}{P(\bar{E}|\bar{D})} \quad \Longleftrightarrow \quad OR = \frac{P(D|E)}{P(\bar{D}|E)} : \frac{P(D|\bar{E})}{P(\bar{D}|\bar{E})}$$

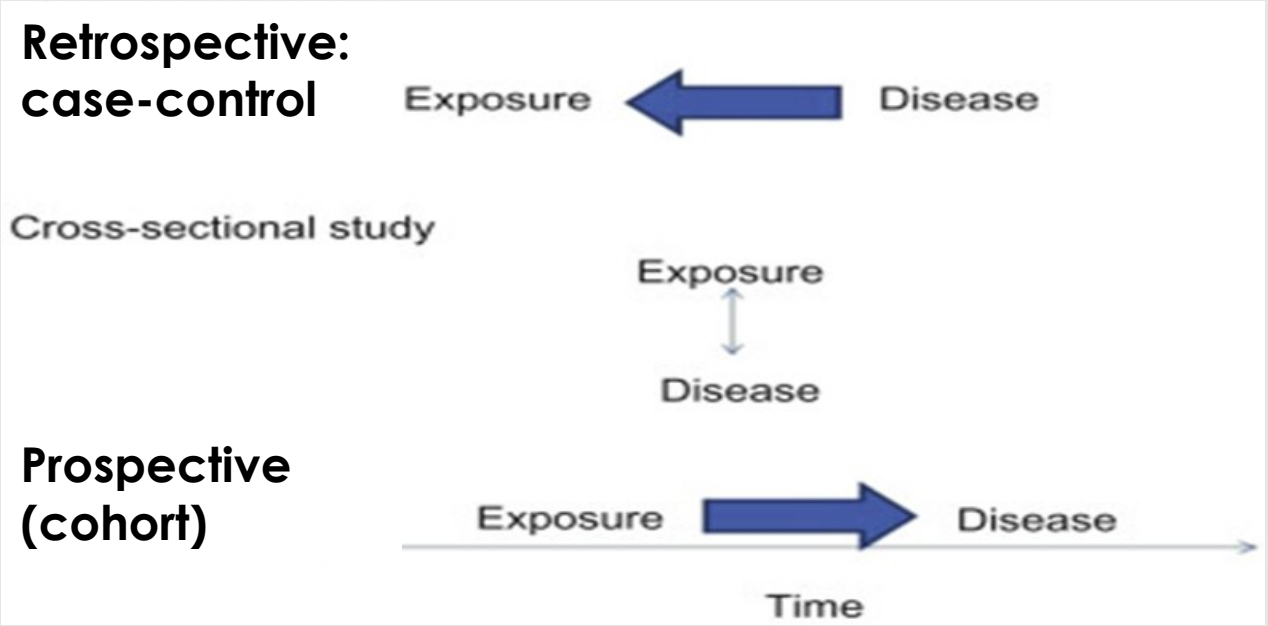
$$OR = \frac{\frac{50}{100} : \frac{50}{100}}{\frac{28}{100} : \frac{72}{100}} = 2.57$$

compatible with the estimates provided by the population-based and cohort data

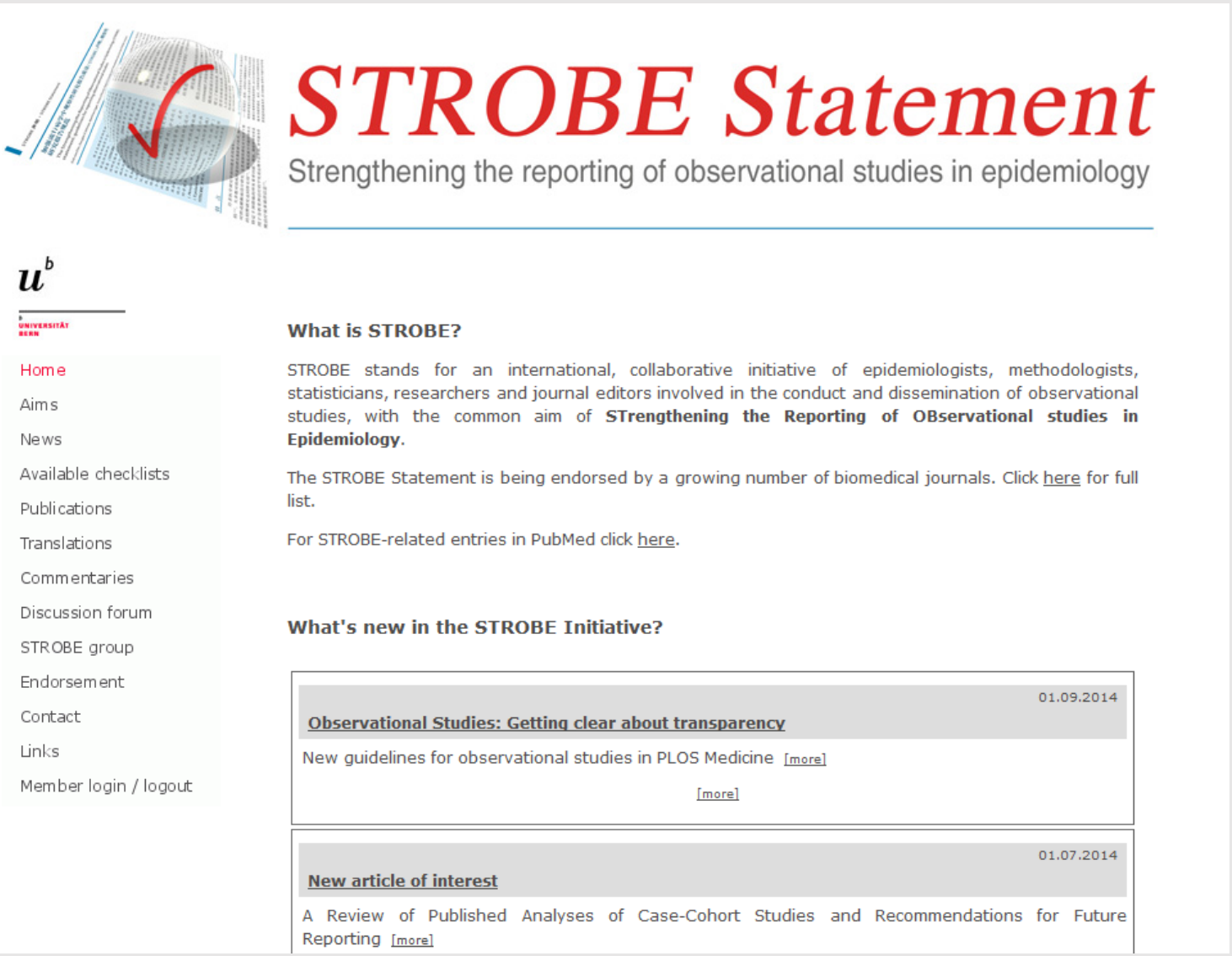
The ***time*** factor

Prospective vs Retrospective **data collection** may have considerable influence on the **quality** and **validity** of exposure measurement and on the ascertainment of ***causal*** relationship (easier in prospective studies!).

On the other side *prospective* measurement of D may require a 10 - or 20 -year **follow-up** period …[easier with **secondary** data sources].



Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)



22-point checklist:

- to consult before plan the study
- as guidelines for writing the results

https://www.strobe-statement.org/index.php?id=strobe-home