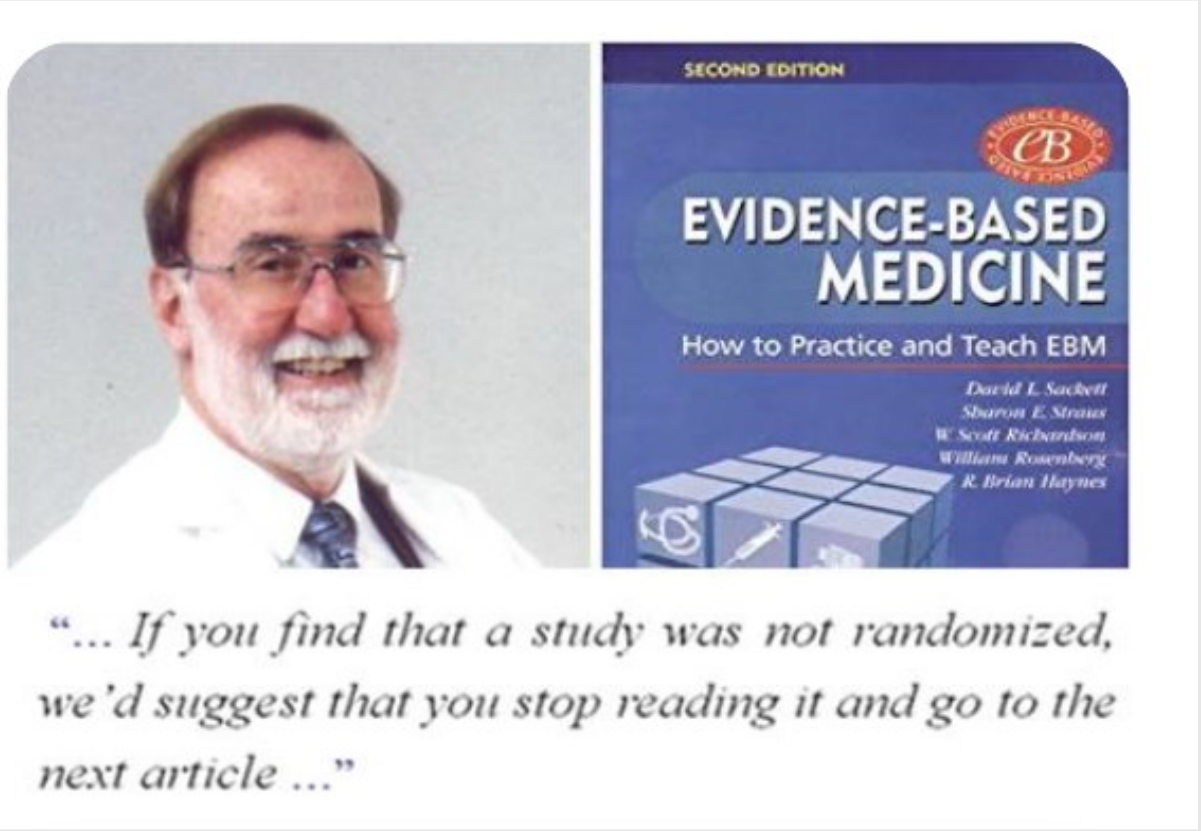
# Study Designs in Epidemiology


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




David L. Sackett
(1934-2015)

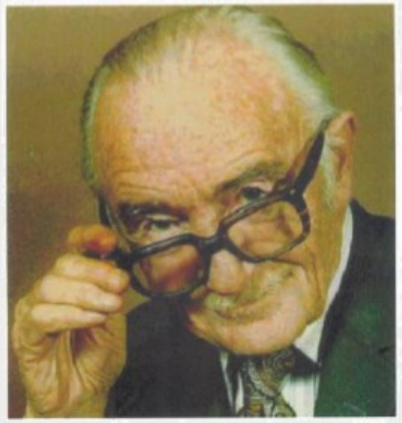
*Evidence Based Medicine: How to Practice and Teach EBM*, 1997, sold 150.000 copies in English and has been translated into numerous languages.

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.

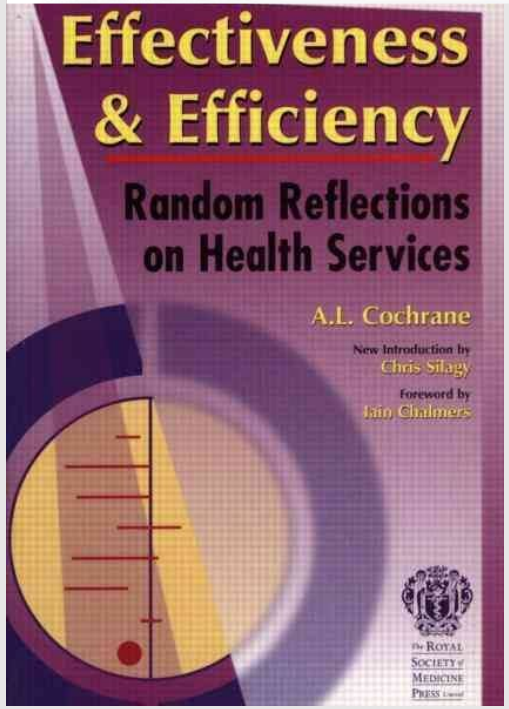
# Archie Cochrane

(1908-1988)

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



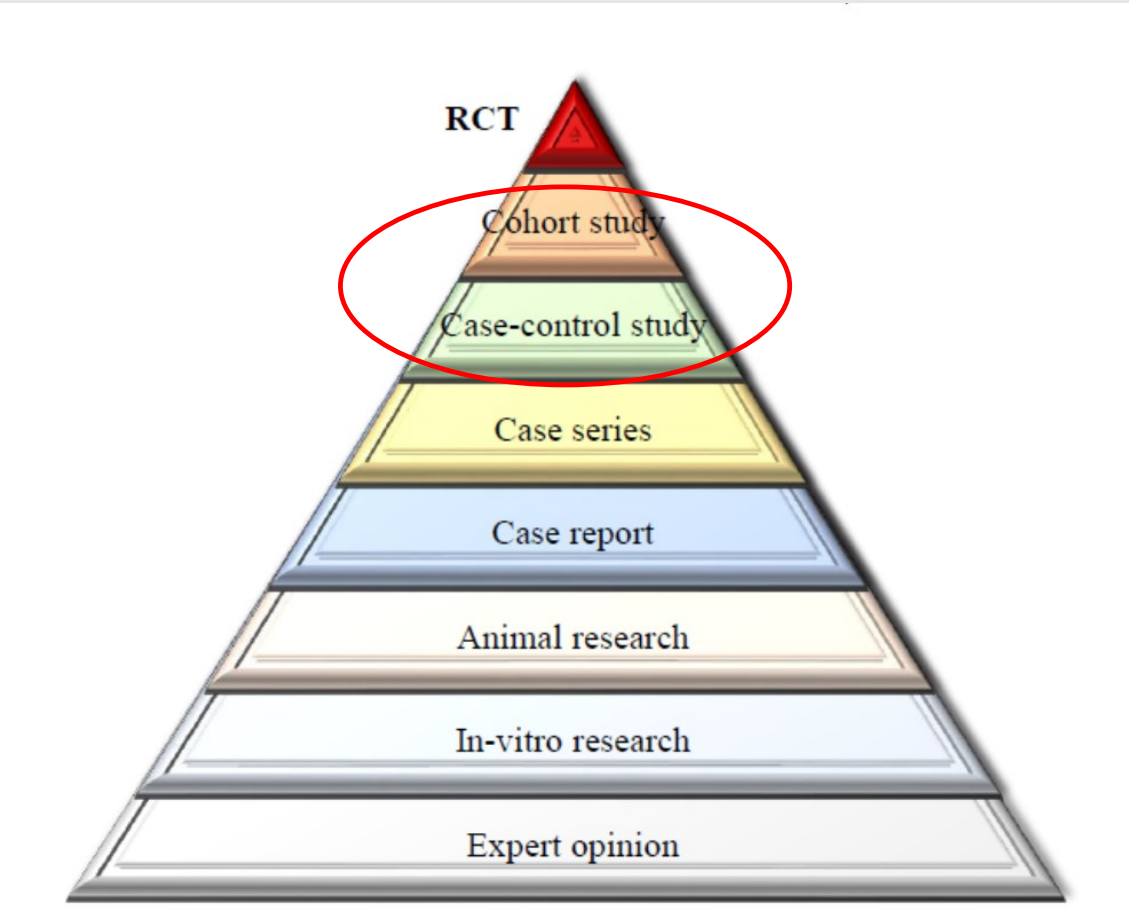
https://www.cochranelibrary.com/

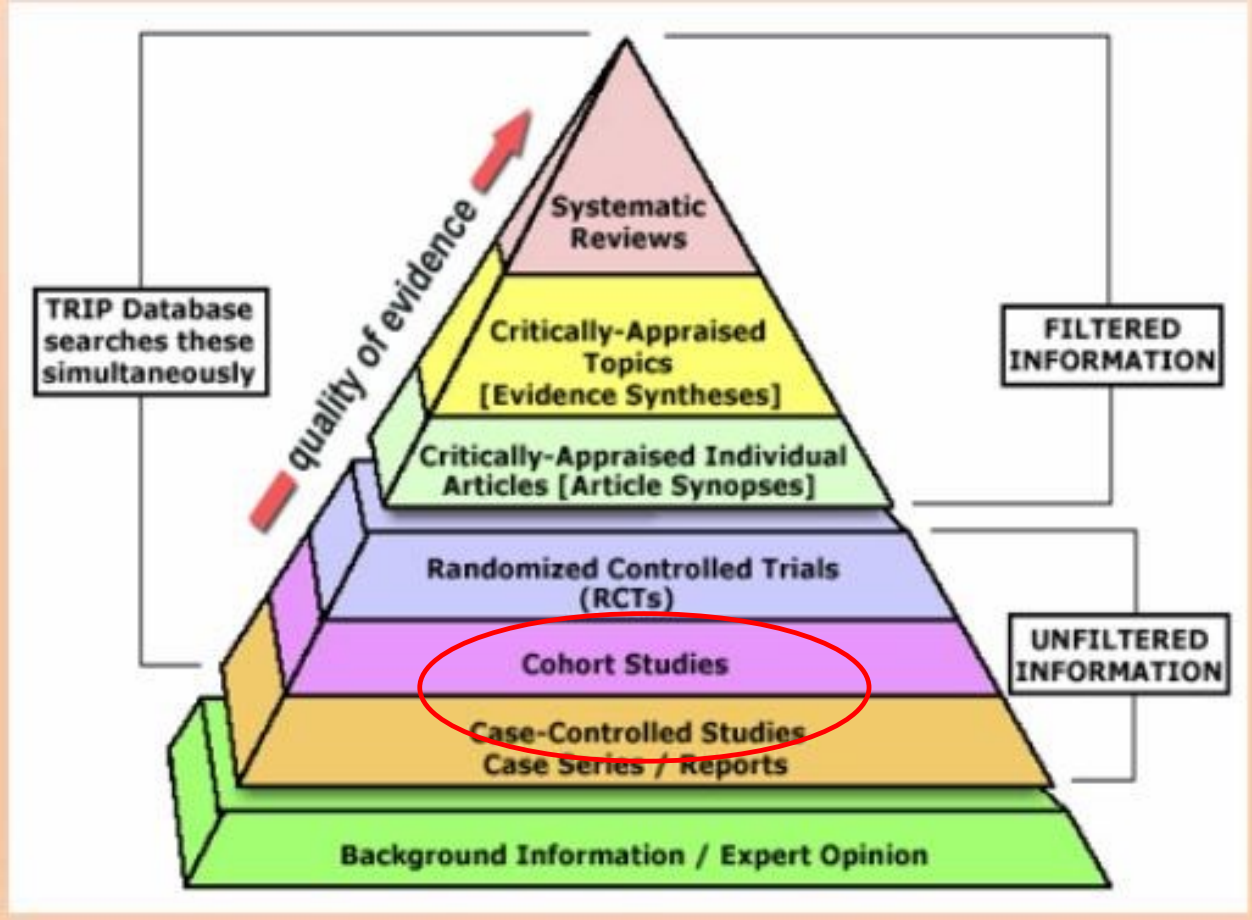




He called for an **international register** of RCTs, for **explicit quality criteria** for appraising published research.
Today, the Cochrane Library is a **collection** of databases that contain different types of high-quality, independent evidence to inform healthcare decision-making (RCTs **and** observational studies).

# Evidence-Based Practice: Evidence Pyramid

The top of the pyramid represents the strongest evidence.



# Observational 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?

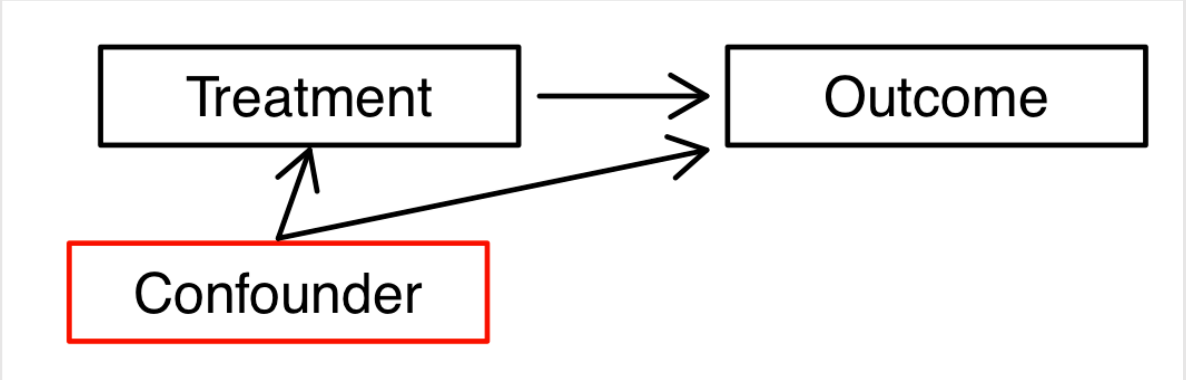
Data:
- **Primary source:** collected **ad hoc** by the researcher for the study
- **Secondary source:** **previously** collected for other purposes (e.g. administrative health data)

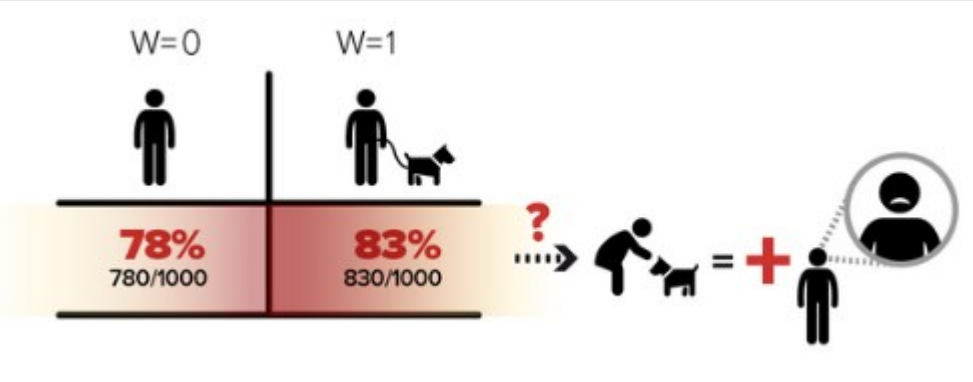
**Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996.**

Haynes B. ***Can it work? Does it work? Is it worth it?*** *BMJ 1999*

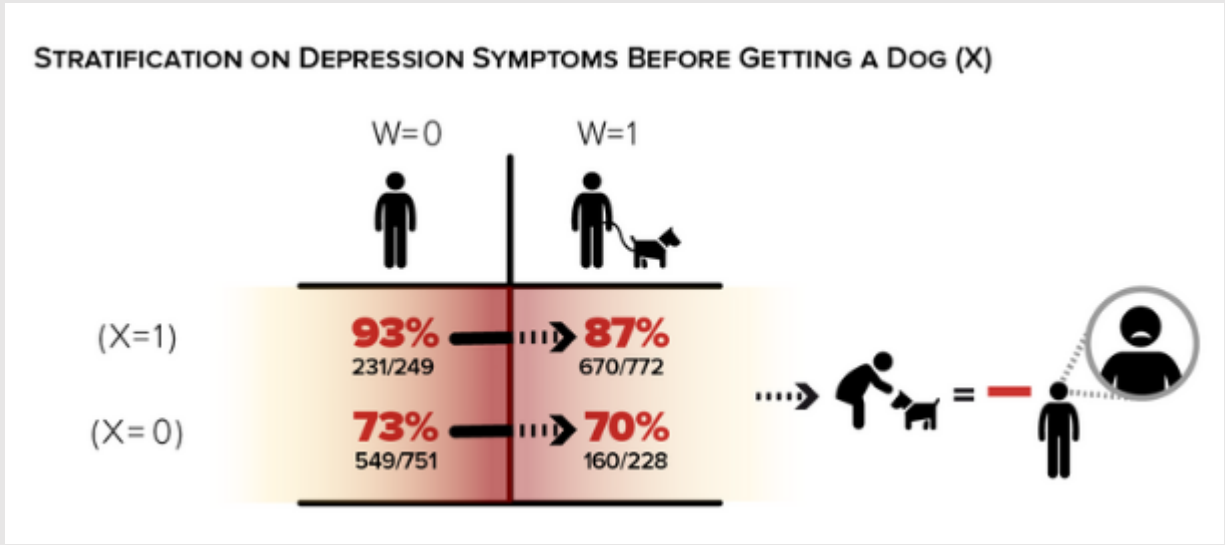
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

**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:

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

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

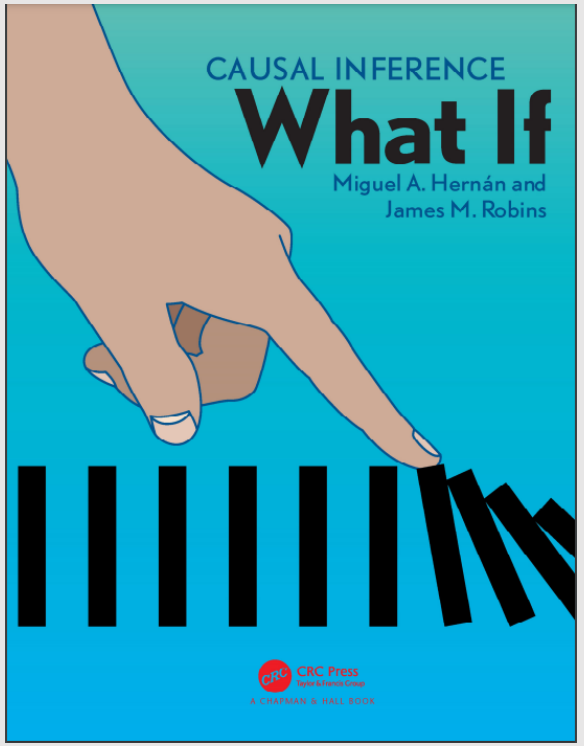
→ ***propensity*** of adopting a dog conditional to the level of depression symptoms pretreatment

A key feature of RCTs is instead 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.

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

Solution?

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

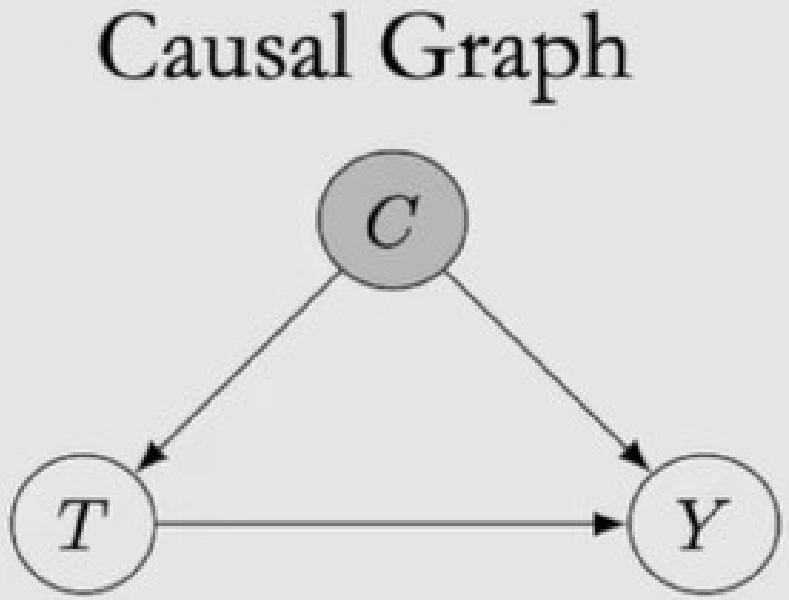


# A basic example:

Two COVID treatments A and B are compared
(outcome: % of subjects died)

Which works better ??

| Treatment | Condition: Mild | Condition: Severe | Total |
|---|---|---|---|
| A | 15% (210/1400) | 30% (30/100) | **16%** (240/1500) |
| B | **10%** (5/50) | **20%** (100/500) | 19% (105/550) |
| | $\mathbb{E}[Y\|t, C=0]$ | $\mathbb{E}[Y\|t, C=1]$ | $\mathbb{E}[Y\|t]$ |



Naive

$$\frac{1400}{1500}(0.15) + \frac{100}{1500}(0.30) = 0.16$$

$$\frac{50}{550}(0.10) + \frac{500}{550}(0.20) = 0.19$$

| Treatment | Condition: Mild | Severe | Total |
|---|---|---|---|
| A | 15% (210/1400) | 30% (30/100) | **16%** (240/1500) |
| B | **10%** (5/50) | **20%** (100/500) | 19% (105/550) |
| | $\mathbb{E}[Y\|t, C=0]$ | $\mathbb{E}[Y\|t, C=1]$ | $\mathbb{E}[Y\|t]$ |

The **naif** approach is to evaluate the global effect of treatment **weigthing it with the observed proportions of mild and severe patients in each of the two groups**

$$\sum_c E(Y|t,c)P(C = c|T)$$

But: treatment B is given to subject that are mostly in the *Severe* condition, instead treatment A to subjects that are mostly in the *Mild* condition…

MILD: Treat A: 1400/1500 w.r.t Treat B: 50/550

SEVERE: Treat A: 100/1500 w.r.t Treat B: 500/550

$$\frac{1450}{2050}(0.15) + \frac{600}{2050}(0.30) \approx 0.194$$

$$\frac{1450}{2050}(0.10) + \frac{600}{2050}(0.20) \approx 0.129$$

In the **causal** approach we compute the global effect of A and B **weigthing it** with respect to the observed proportions of the baseline condition in the ***overall*** population

$$\sum_c E(Y|t,c)P(C=c)$$

(G-formula*)

**Condition**

| Treatment | Mild | Severe | Total | Causal |
|---|---|---|---|---|
| A | 15% (210/1400) | 30% (30/100) | **16%** (240/1500) | 19.4% |
| B | **10%** (5/50) | **20%** (100/500) | 19% (105/550) | **12.9%** |
| | $\mathbb{E}[Y\|t,C=0]$ | $\mathbb{E}[Y\|t,C=1]$ | $\mathbb{E}[Y\|t]$ | $\mathbb{E}[Y\|\text{do}(t)]$ |

In this way we *weight* each treatment effect as if *all were treated* versus *all were untreated*…[standardized to the overall distribution of the counfounder].

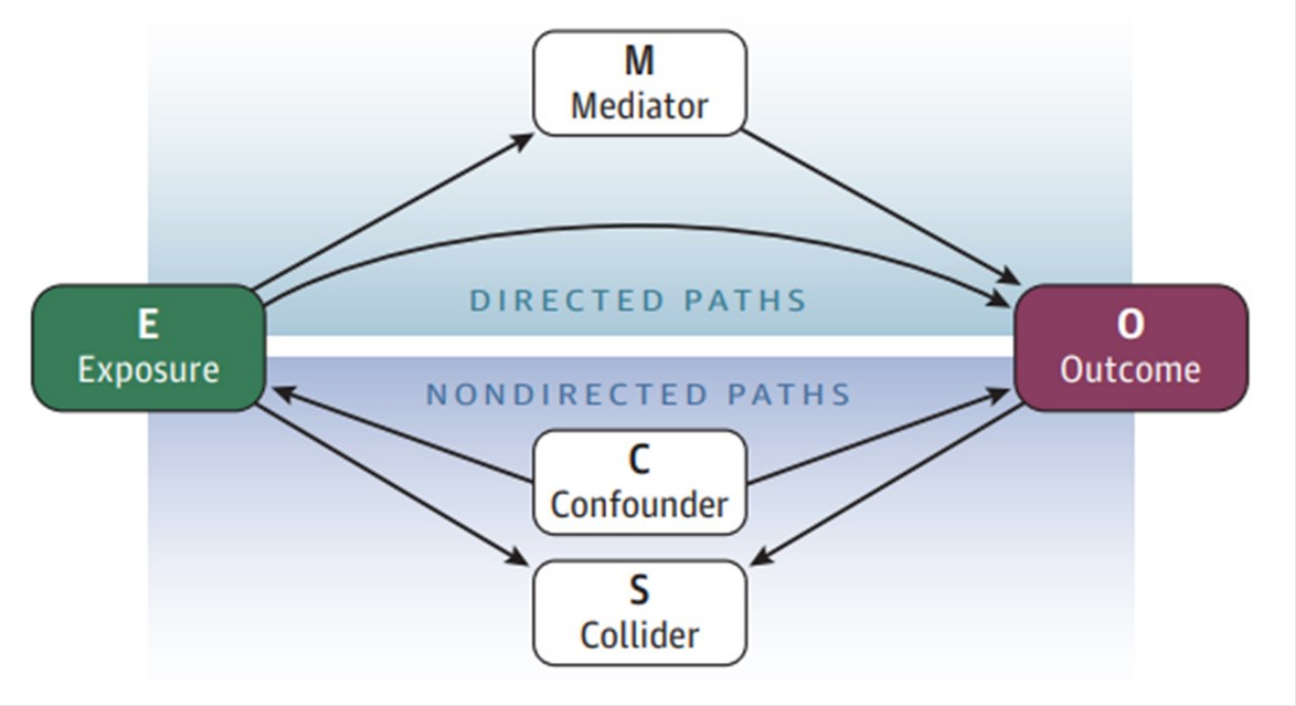
[*non-parametric approach]

# *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 (***conditional***) **average** treatment effects (**both in RCTs and in observational studies**).

Data alone **are not sufficient** to predict the counterfactual outcome. We need to introduce **several assumptions** that essentially embed **subject matter expert knowledge**.

**We will see some specific methods in Block 3 !**




To estimate causal effects, we should define **relationships** between variables **before** the statistical analysis is performed…

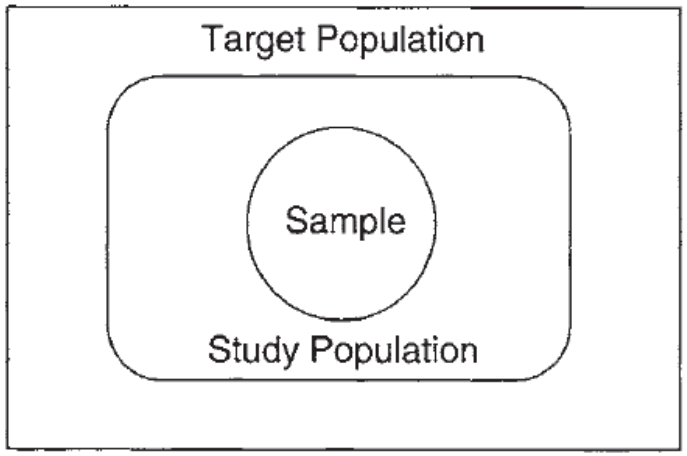
# Observational design: (re)-fresh of key-definitions

**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.

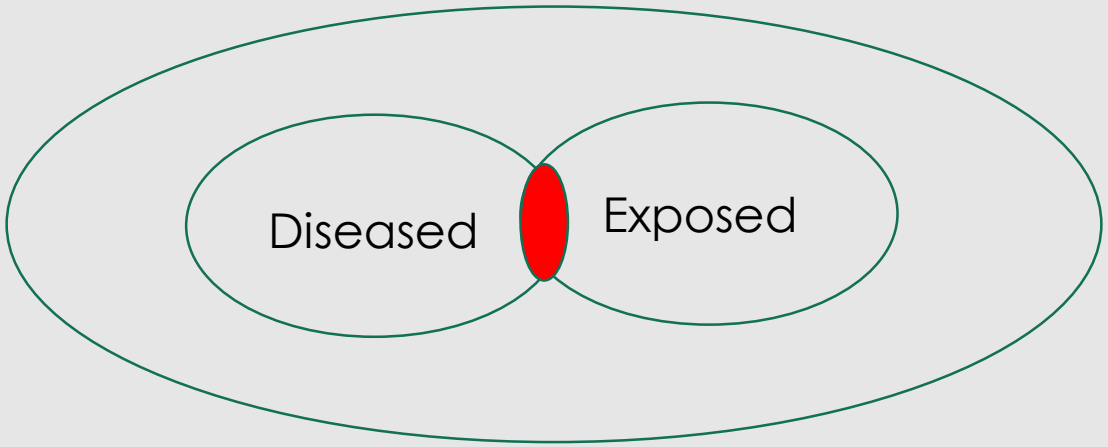
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

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?

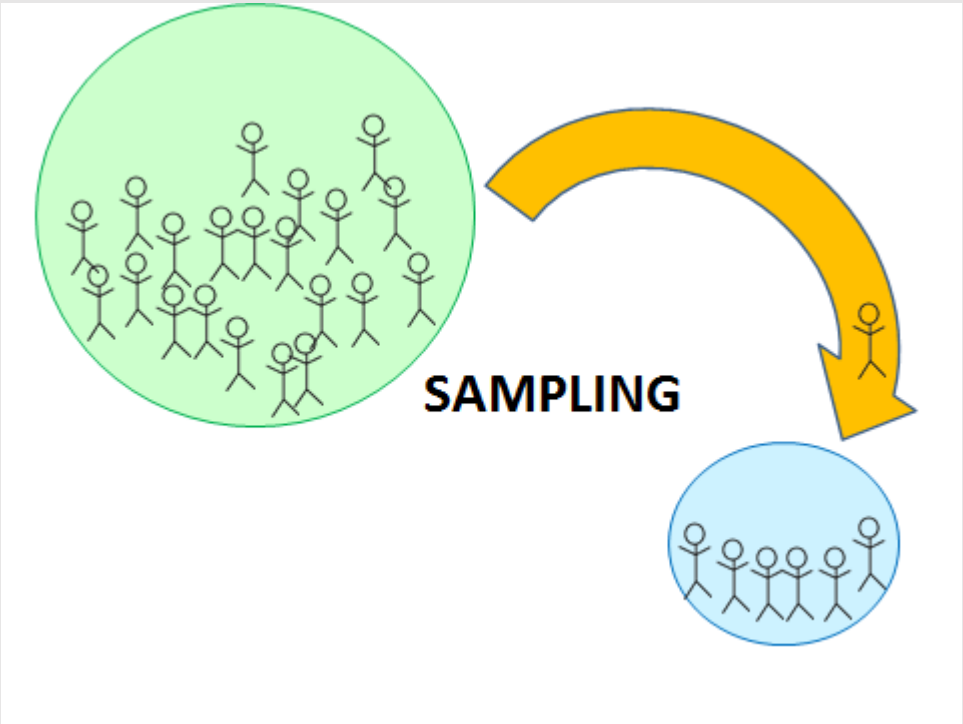
**3** forms of *sampling schemes* are most commonly used in observational studies



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 (**cohort study [primary/secondary data source]**).

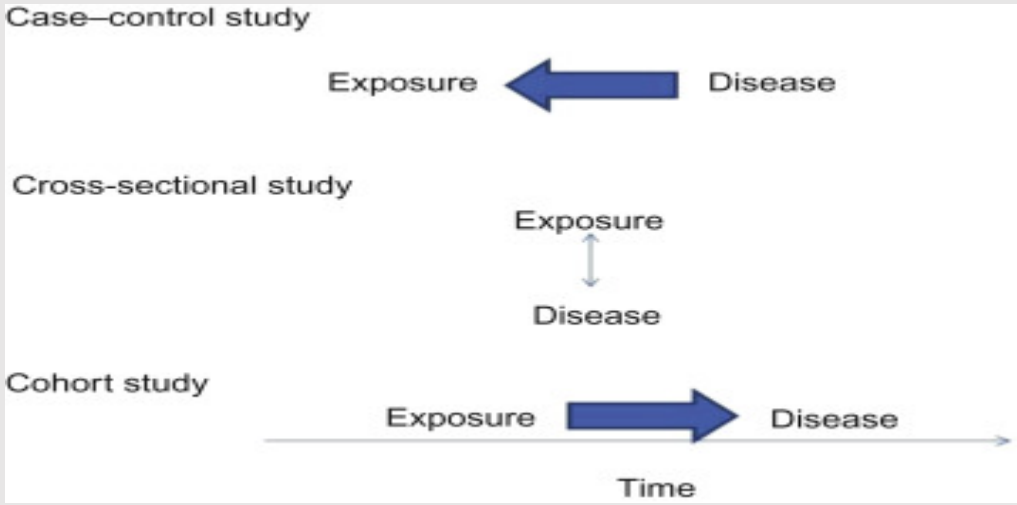
- In a **retrospective** study, *measurement of [past] exposure* occurs ***after*** an individual's disease status has been already determined (there is no follow up, like in **case-control** studies).

A specific type of population-based study is the **cross-sectional** study : here measurement of D and E **always *coincides* with sampling timing** (simultaneous measure of D and E, no follow up).

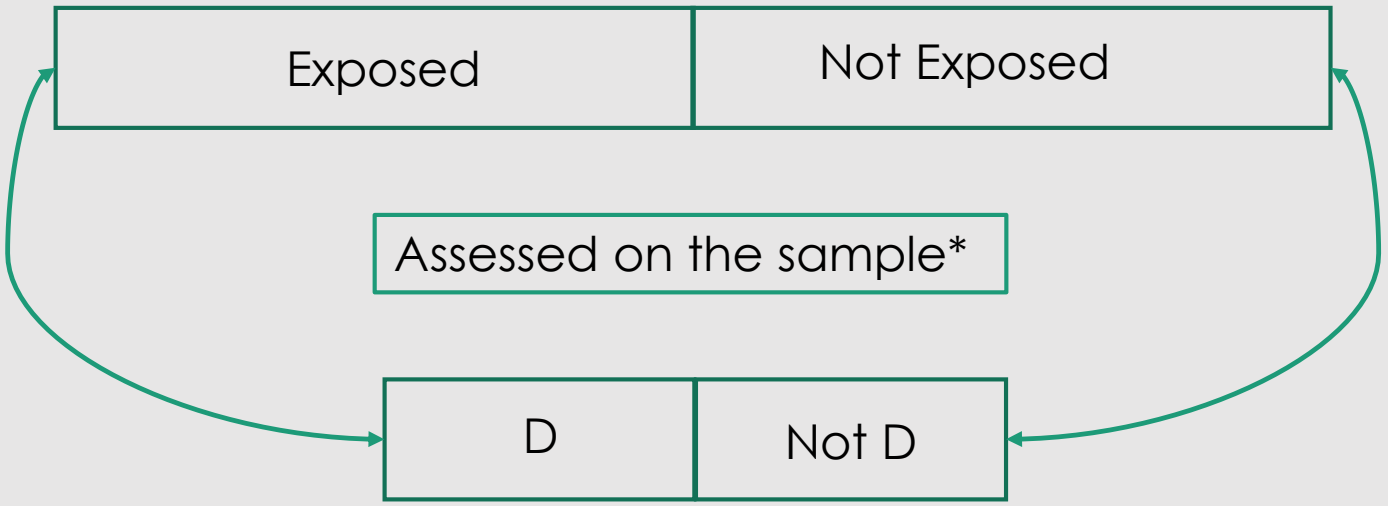
Whether a study is prospective or retrospective is not relevant to the development of *statistical* properties.

Prospective vs Retrospective may have instead considerable influence on the **quality** and **validity** of exposure measurement and on the ascertainment of **causal** relationship.

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



# Population-based studies [*Cohort if prospective]




Population-based studies ideally measure exposures, confounders and outcome times of **all** population members.

However, they are often **very expensive** in terms of time and resources.

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

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

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 [Block 1] but 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:

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

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

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

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

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

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

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

$$\widehat{AR} = \frac{(14/200 - 7/141)}{14/200} = 0.29$$

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

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 a cohort design (but related to *causality* assumptions...)

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

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

This division is important in determining the amount of information that a cohort study yields 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 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

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

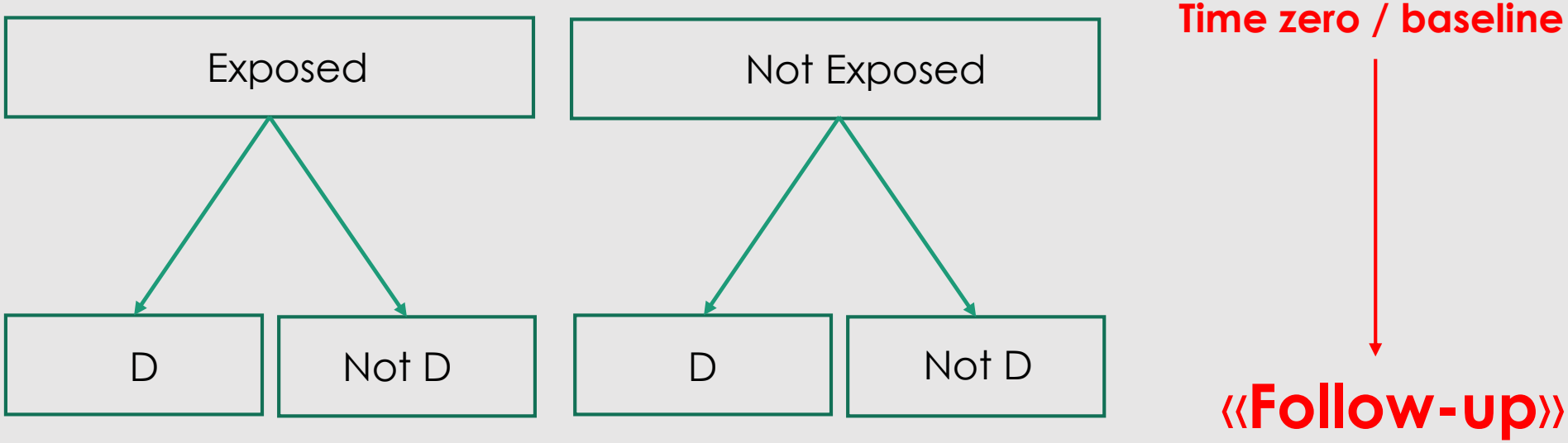
$$\widehat{OR} = \frac{(12/100):(88/100)}{(5/100):(95/100)} = 2.59$$

**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 |

# 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**).

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

# 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** ($\overline{D}$).

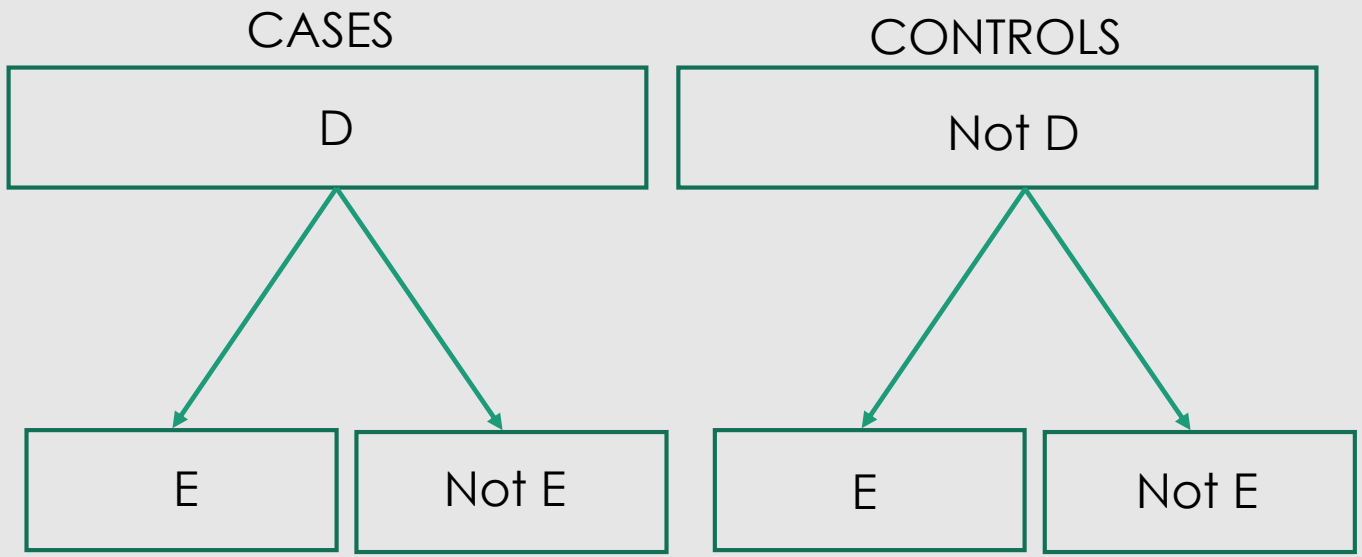
1. Identify **two subgroups** of the population on the basis of the presence or absence of D.

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].

# 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.

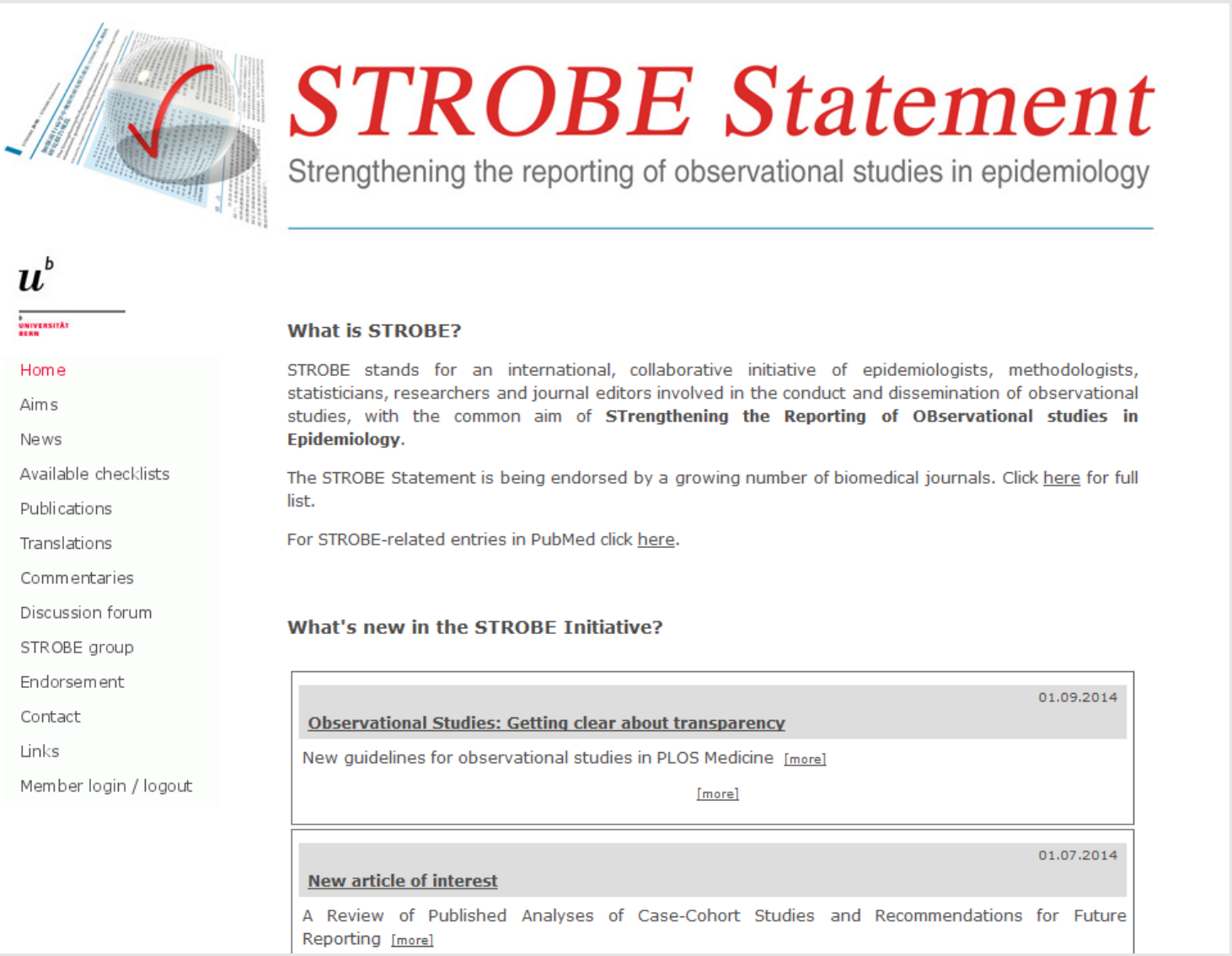
However, we can estimate the **Odds Ratio** for **E associated with D** and for symmetry the **reverse**:

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

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

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

# 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