# Causation

- The **epidemiological** approach to causation
- The **causal inference** approach (intro)

Block 1.2

# The epidemiological approach to causation

One of the main goal of (analytical) epidemiology is to learn about what **causes** and **prevents** diseases.

How epidemiologists determine causative and preventive factors involves a process known as **causal inference**. This process is particularly complex in **observational** studies.

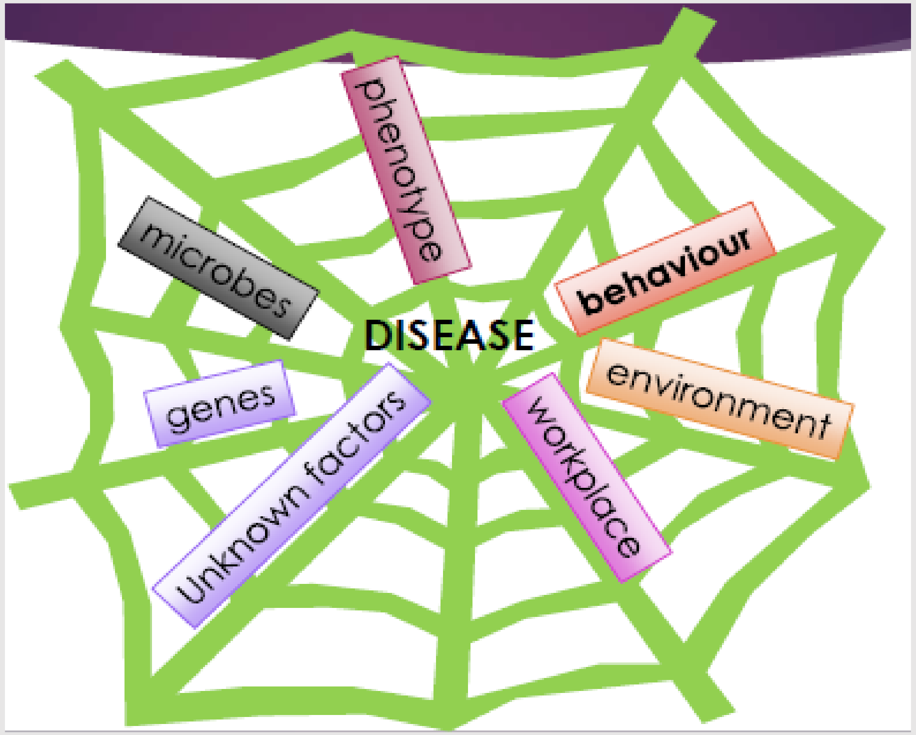
EVERYBODY WHO WENT TO THE MOON HAS EATEN CHICKEN!
GOOD GRIEF. CHICKEN MAKES YOU GO TO THE MOON!

# The epidemiological approach to causation

Epidemiological principles stand on **two** basic assumptions:

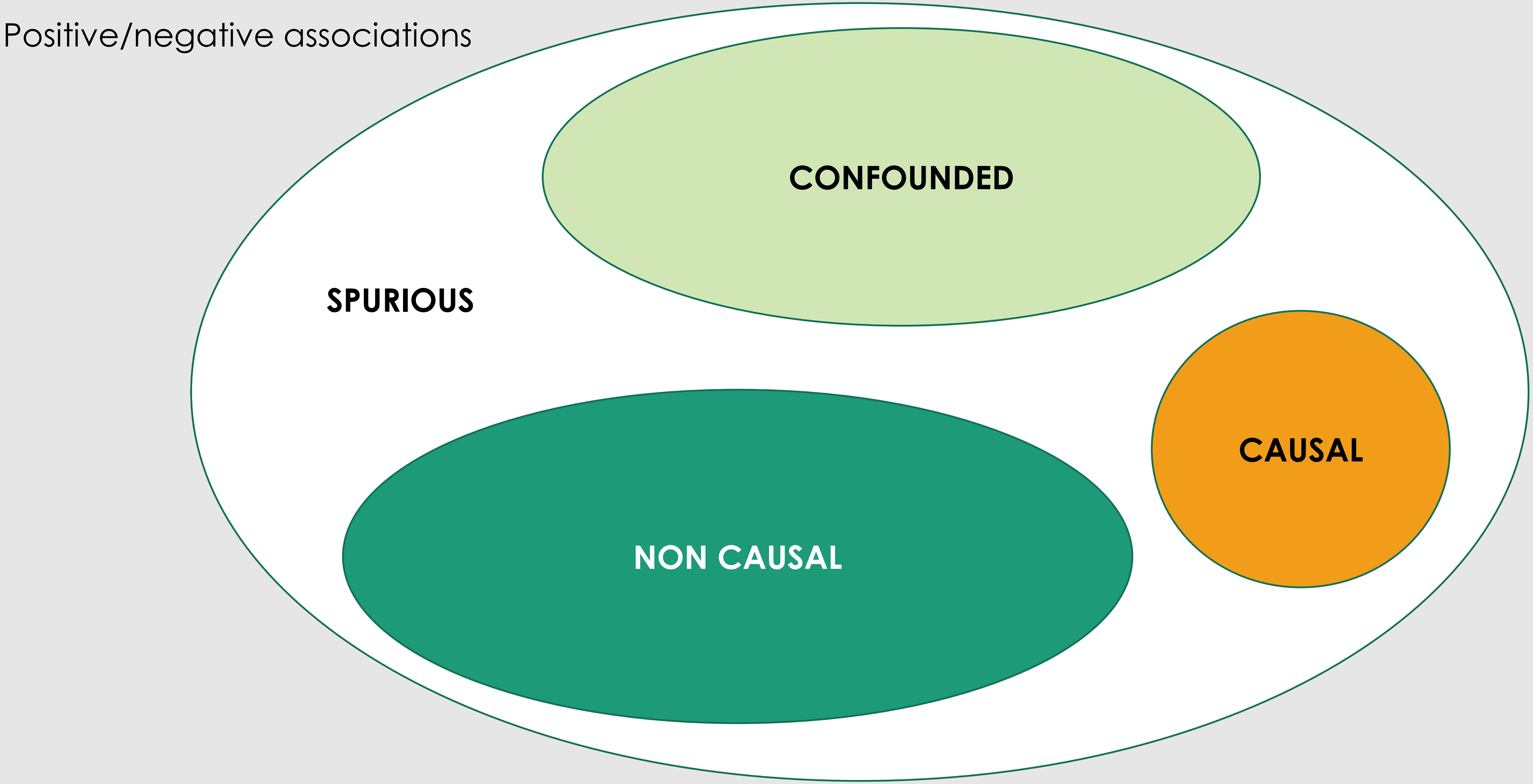
- Human disease does not occur *(completely…)* at random

- The disease and its cause - as well as preventive factors - can be identified by a thorough **investigation** of population.



Block 1.2

## From association to causation



Positive/negative associations

SPURIOUS

CONFOUNDED

NON CAUSAL

CAUSAL

# The epidemiological approach to causation

## What is Association?

Simultaneous *occurrence* of two variables *more often* than would be expected **by chance**.

If two attributes, say A and B,  are found to co-exist ***more often*** *than an ordinary chance*.

Useful to consider **as a first step** the concept of (statistical) correlation

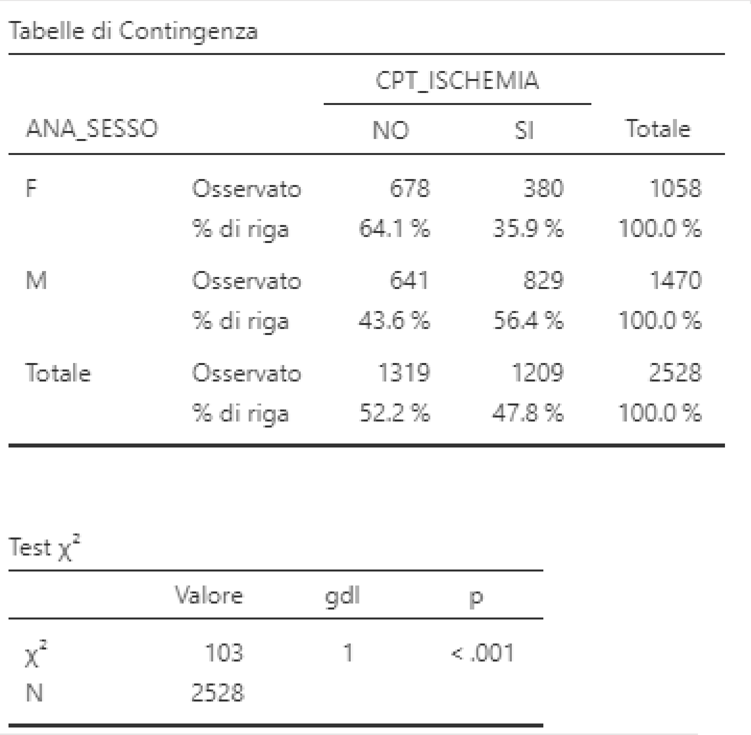
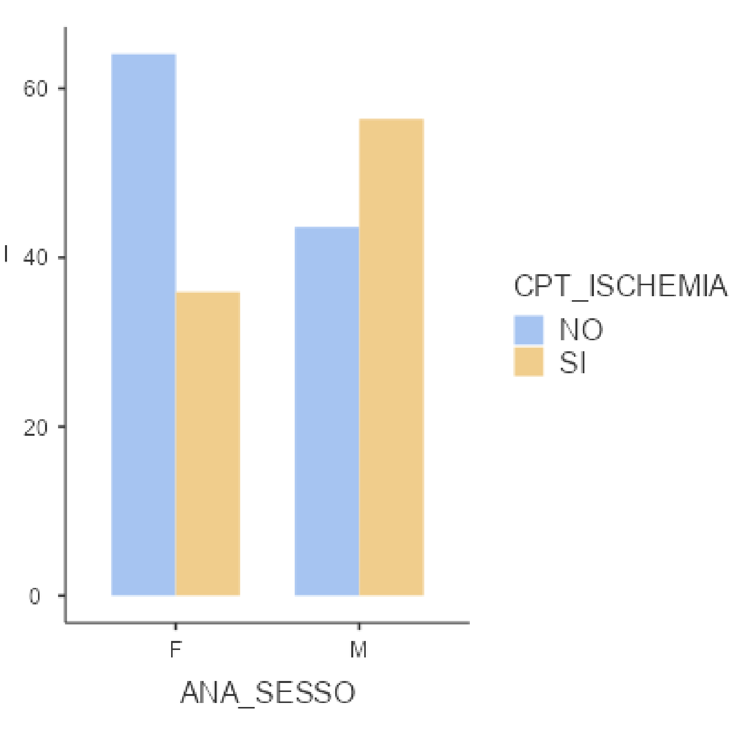


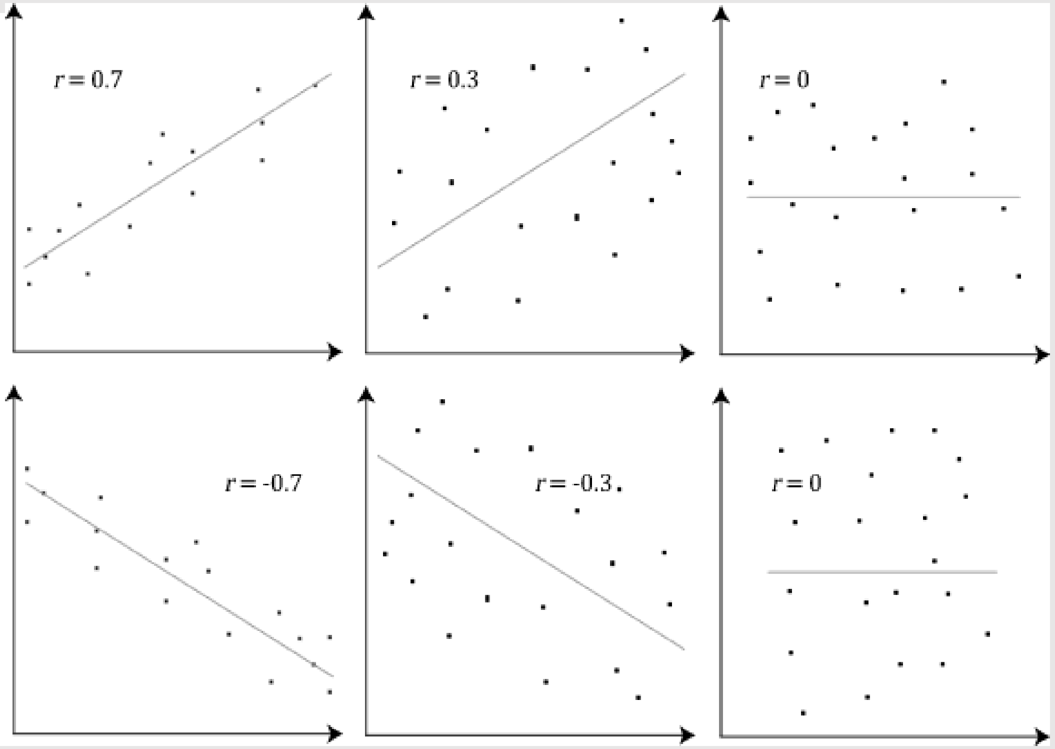
Tabelle di Contingenza

| ANA_SESSO | | CPT_ISCHEMIA NO | CPT_ISCHEMIA SI | Totale |
|---|---|---|---|---|
| F | Osservato | 678 | 380 | 1058 |
| | % di riga | 64.1 % | 35.9 % | 100.0 % |
| M | Osservato | 641 | 829 | 1470 |
| | % di riga | 43.6 % | 56.4 % | 100.0 % |
| Totale | Osservato | 1319 | 1209 | 2528 |
| | % di riga | 52.2 % | 47.8 % | 100.0 % |

Test $\chi^2$

| | Valore | gdl | p |
|---|---|---|---|
| $\chi^2$ | 103 | 1 | < .001 |
| N | 2528 | | |





Block 1.2

# Chocolate Consumption, Cognitive Function, and Nobel Laureates

N Engl J Med, 2012 Oct 18;367(16)

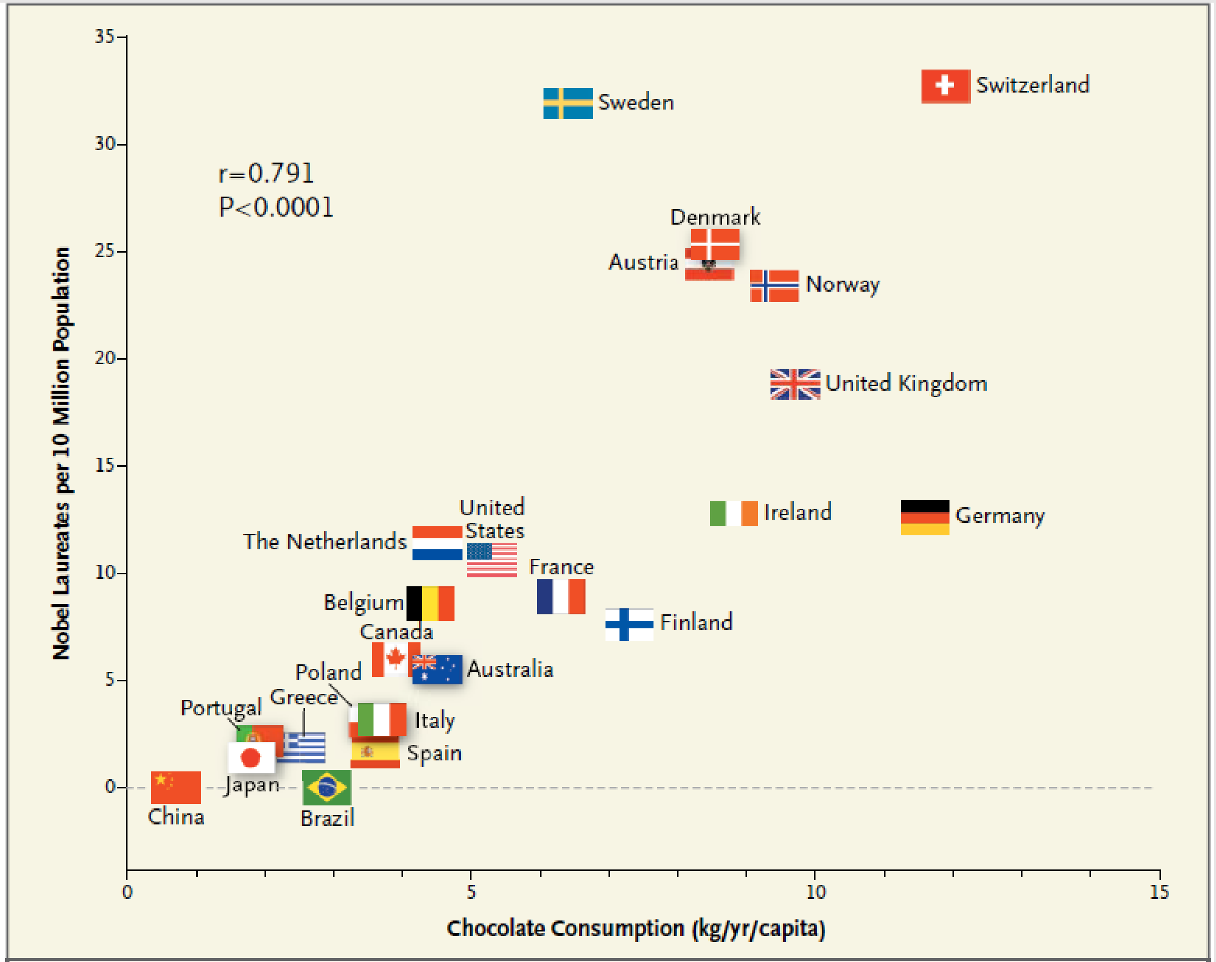
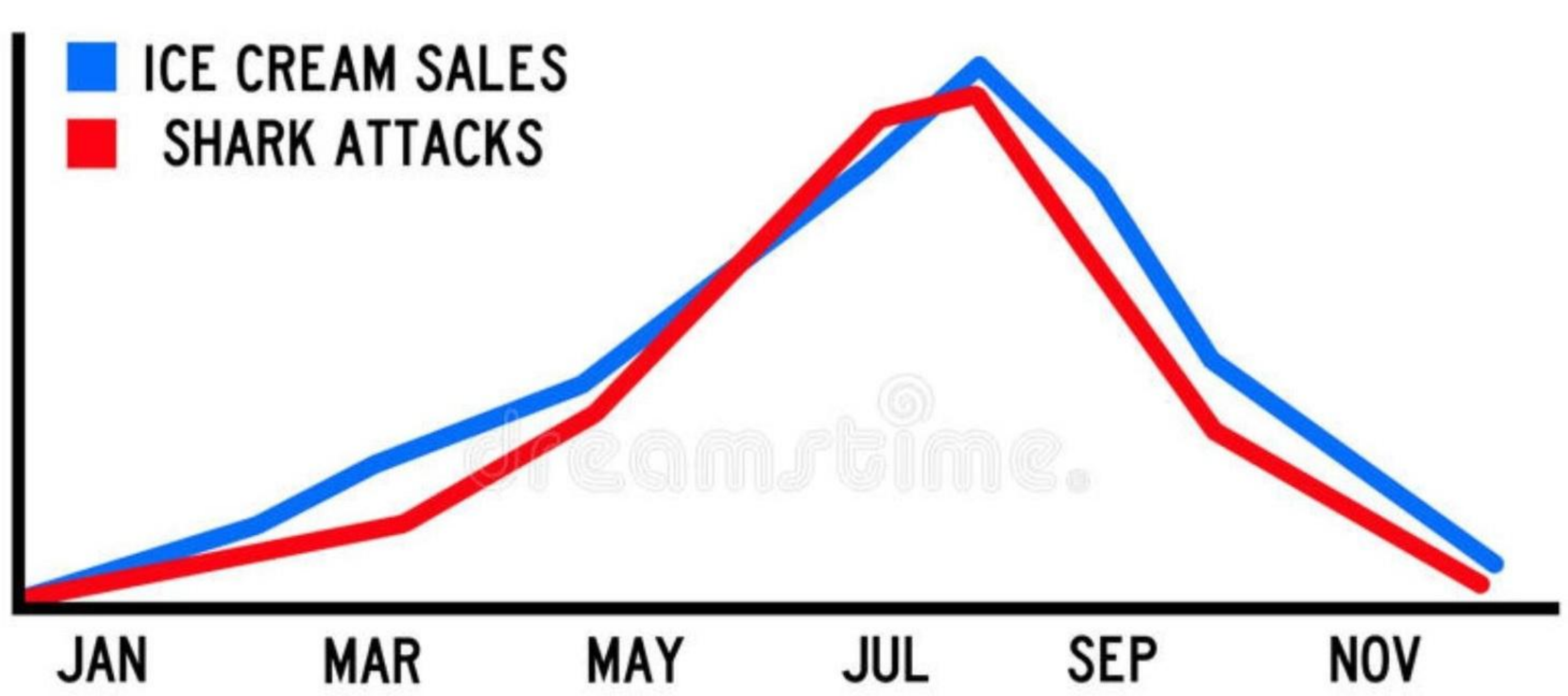


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

**Block 1.2**



They both increase during summer months

**Block 1.2**

## Association can be…

<u>**Spurious**</u> **:** not real, artificial, fortuitous, false, non-causal associations *due to chance*

An observed association between a disease and suspected factor *may not be real*

The ringing of alarm clocks **AND** the rising of the sun

Cock's crow causes the sun to rise (?!)

Neonatal mortality higher in those who were born in a hospital rather than at home.
Is home delivery ***better*** for newborn's health ?

↓

**high risk** deliveries higher in the hospital than at home
(***selection bias…***)

Block 1.2

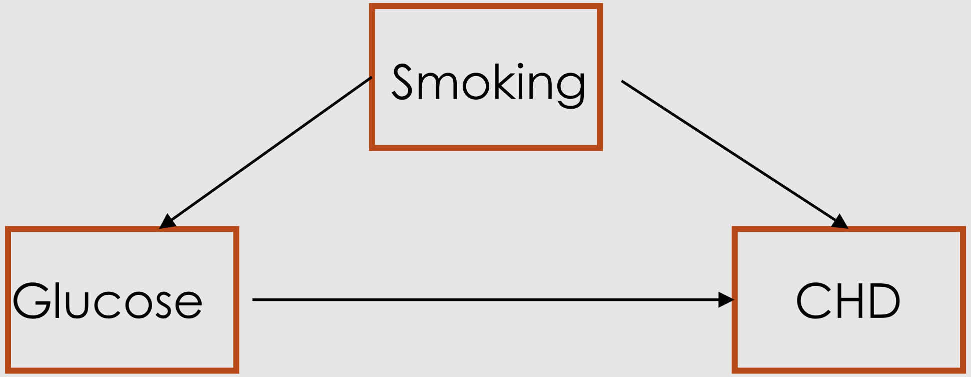
## Indirect/Confounded Association:

It is a **statistical** association between a characteristics and a disease due to the presence of ***another factor*** i.e. a common factor (**confounding** variable).

So the association is due to the presence of another factor which is **common** to both.

- Glucose and CHD (*Coronary Heart Disease*) [confounding factor could be cigarette smoking]

(smoking increases the consumption of coffee & amount of sugar consumed **and** increase the risk of CHD!)



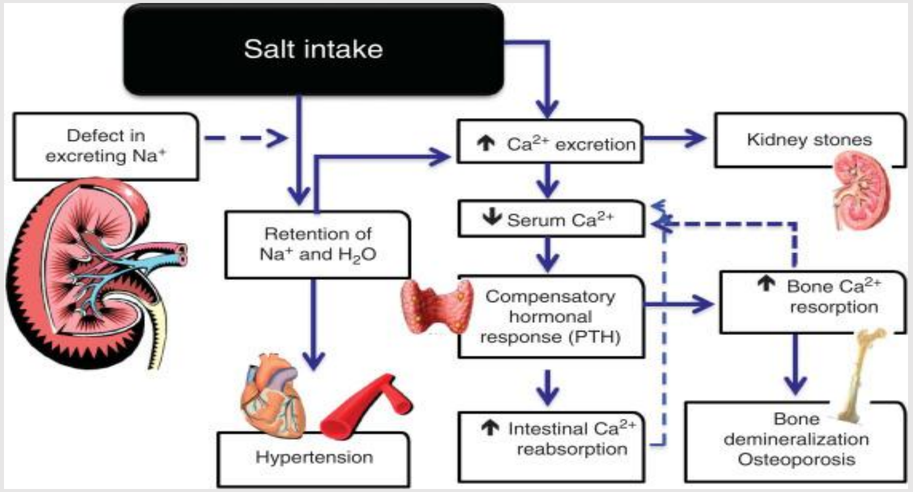
**Block 1.2**

**Non-causal:** *Non-directional* (***true***) association between two variables

Ex: alcohol use and smoking

**Finally…causal**:

- a change in the **independent** variable must cause ***change*** in **dependent** variable
- **time** and **direction** (salt intake and hypertension)
- the association between the two attributes is not through a third attribute
- when the disease is present, the factor must also be present

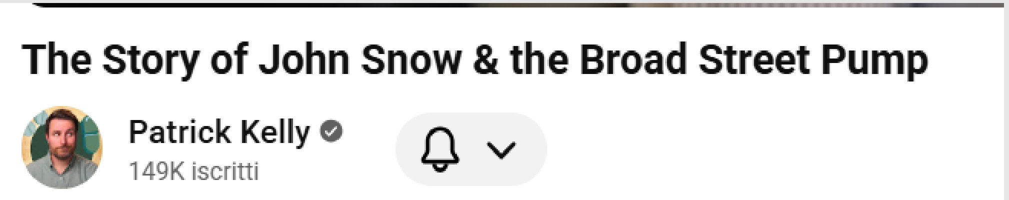


Salt intake
Defect in excreting $Na^+$
↑ $Ca^{2+}$ excretion
Kidney stones
Retention of $Na^+$ and $H_2O$
↓ Serum $Ca^{2+}$
Compensatory hormonal response (PTH)
↑ Bone $Ca^{2+}$ resorption
Hypertension
↑ Intestinal $Ca^{2+}$ reabsorption
Bone demineralization Osteoporosis

The American Heart Association recommends ≤ 2.5 mg of salt a day [ideal limit ≤ 1.5 mg per day], especially for those with high blood pressure.

# So…how to establish a causal relationship in [observational] epidemiology?

Statistical methods help but **prior** knowledge is required …



**The Story of John Snow & the Broad Street Pump**

Patrick Kelly ✓
149K iscritti

https://youtu.be/bALs7kNpNSM?si=LmtUElxKrElYI3w0

Epidemiologists use some ***criteria***:

- Temporal association
- Strength of association [-> *effect size*]
- Dose-response relationship
- Biological plausibility
- Alternate Explanations
- Effect of cessation of exposure
- Consistency of association [reproducibility]
- Specificity of association

## 1. Temporal association:

- The causal attribute must **precede** the disease or unfavorable outcome
- Exposure to the factor must have occurred **before** the disease developed
- **Length of interval** between exposure and disease is very important
- Its more obvious in *acute* disease than in *chronic* disease

Cause must precede the effect.

Drinking contaminated water → occurrence of diarrhea

[In many chronic cases, because of *insidious onset* and ignorance of precise ***induction period***, it gets hard to establish a temporal sequence as which comes first -the suspected agent or disease].

## 2. Strength of the association [*effect size*]:

- Relationship between cause and outcome could be strong or weak.
- With the ***increasing*** level of exposure to the risk factor there should be an ***increase*** *in the incidence* of the disease.
- Strong associations *are more likely to be causal* than weak.
- Weaker associations are more likely to be explained by undetected **bias/confounders**.
- But weaker association does not rule out causation!

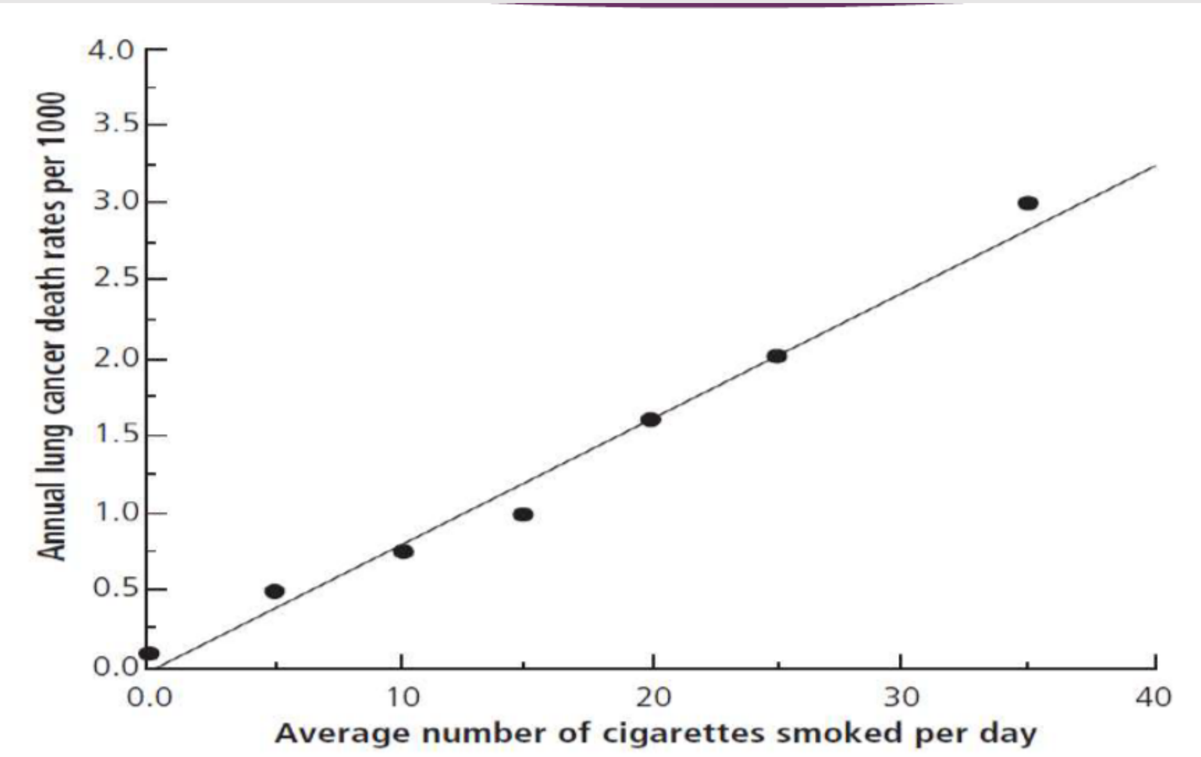
**Strength** of association can be **quantified** by (statistical) estimate of **risk**

[odds ratios, relative risk, attributable risk… etc…]

## 3. Dose-Response Relationship
(*The Biological Gradient*)

- As the dose of exposure increases, so does the risk of disease
- If a dose-response relationship is present, there is strong evidence for a causal relationship.
- However: the absence of a dose-response relationship does not necessarily rule out a causal relationship [think to *binary* exposures].
- In some cases in which a ***threshold*** may exist, no disease may develop *up to a certain level* of exposure; above this level, disease may develop [non-linearity…]

Death rates from lung cancer (per 1000) by number of cigarettes smoked, British male doctors, 1951 –1961



## 4. Biological Plausibility

The association must be **consistent** with the other knowledge (mechanism of action, evidence from animal experiments …etc…).

Sometimes the lack of plausibility may simply be due to the **lack of sufficient knowledge** regarding the pathogenesis of a disease.

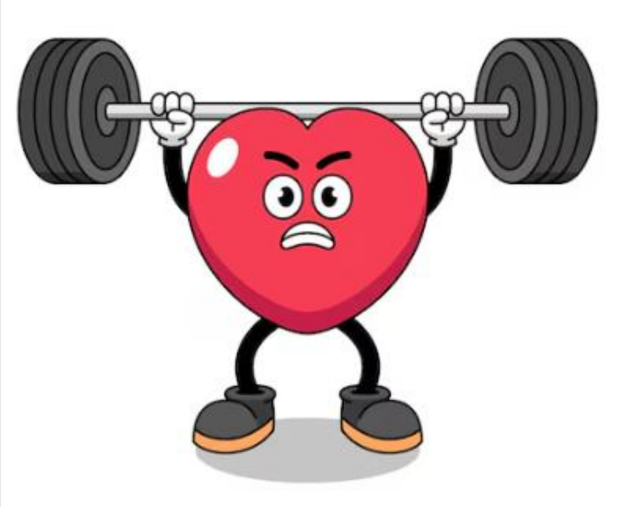
It is not so often based on logic or data but only on prior beliefs.

It is difficult to demonstrate where also **confounder** exhibits a biological gradient …

Risk Factor: ***Body Mass Index***

Outcome: ***Heart Disease***

Confounder: ***amount*** of ***fatty foods*** in the diet

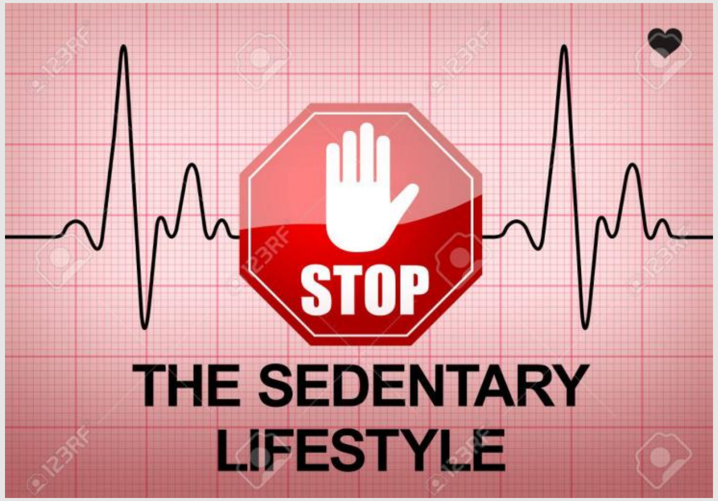


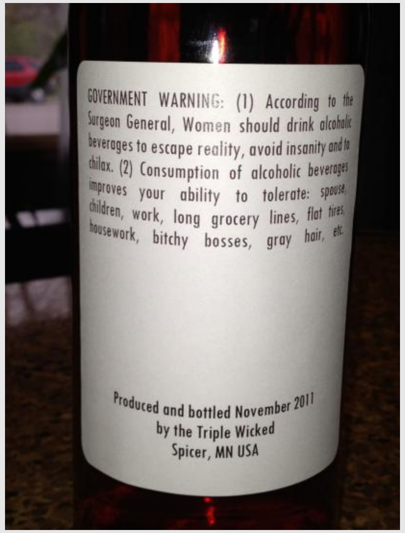
## 5. Consideration of alternate explanations:

In judging whether a reported association is causal, the extent to which the investigators have taken **other possible explanations** into account and the extent to which they have **ruled out** such explanations are important considerations.

## 6. Cessation of exposure:

If a factor is a cause of a disease, we would expect the risk of the disease to **decline** when exposure to the factor is reduced or eliminated…(basis of **public health policy actions**)





GOVERNMENT WARNING: (1) According to the Surgeon General, Women should drink alcoholic beverages to escape reality, avoid insanity and to chilax. (2) Consumption of alcoholic beverages improves your ability to tolerate: spouse, children, work, long grocery lines, flat tires, housework, bitchy bosses, gray hair, etc.

Produced and bottled November 2011
by the Triple Wicked
Spicer, MN USA

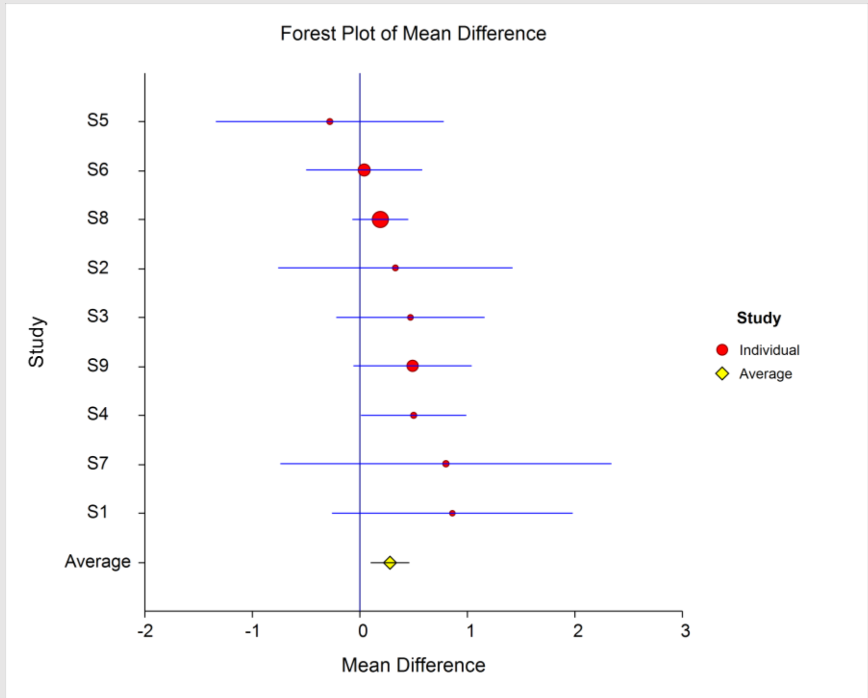
## 7. *Consistency* of the association:

Consistency is the occurrence of the association **at some other time and place repeatedly** unless there is a clear reason to expect different results.

If a relationship is causal, the findings should be **consistent** with other data*. Lack of consistency however does not rule out a causal association**.

*Repeated observation of an association in ***different*** populations under ***different*** circumstances.

***Statistical tool: metanalysis**



** **different** populations?
different methods ?
…careful evaluation of all aspects of **study design** !

## 8. Specificity of the association:

(the **weakest** of the criteria, should probably be eliminated…)

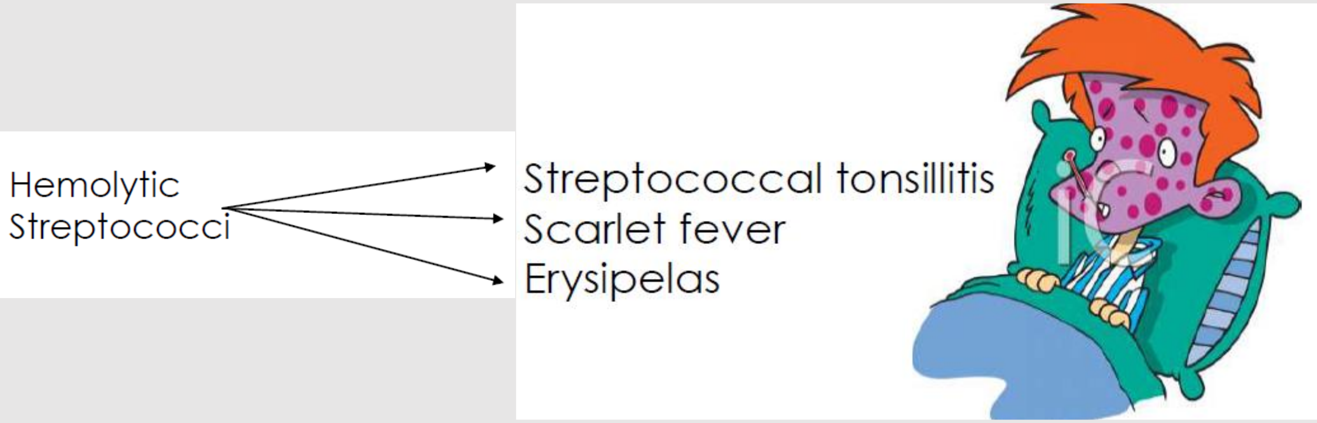
Specificity implies a **one to one** relationship between the cause and effect.

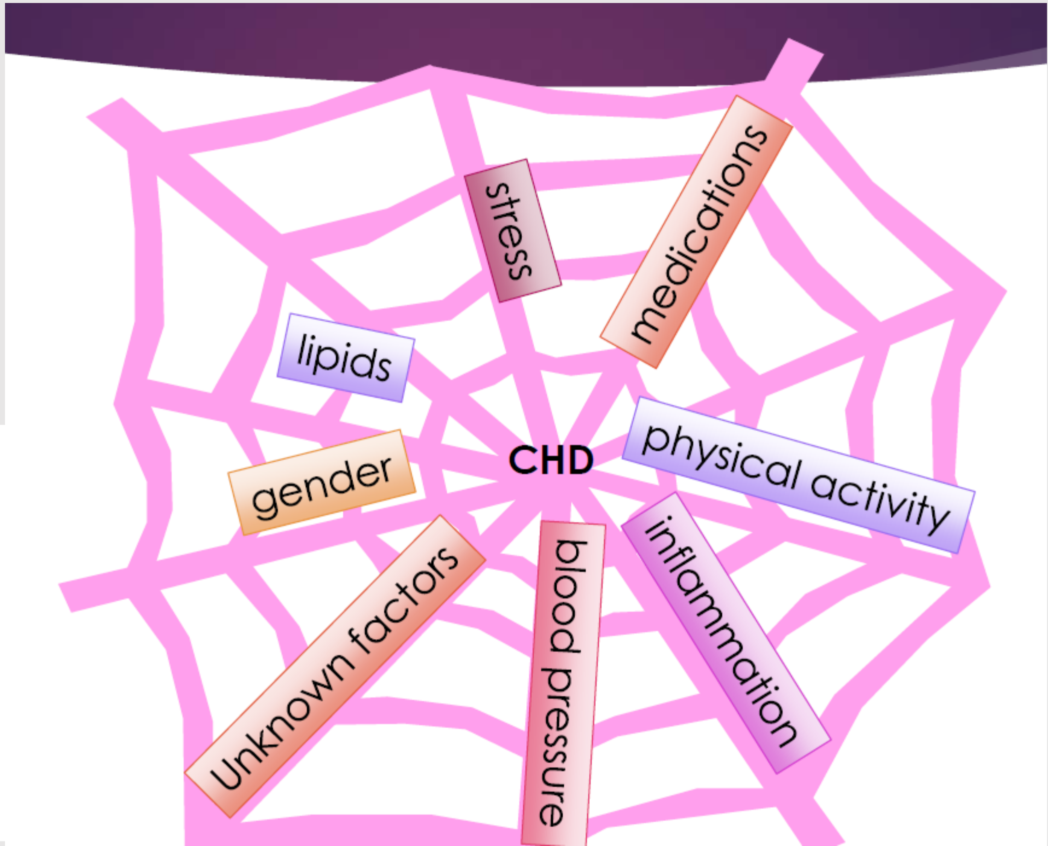
It’s the most difficult to occur for 2 reasons:

- Single cause or factor can give rise to more than 1 disease
- Most diseases are due to multiple factors.

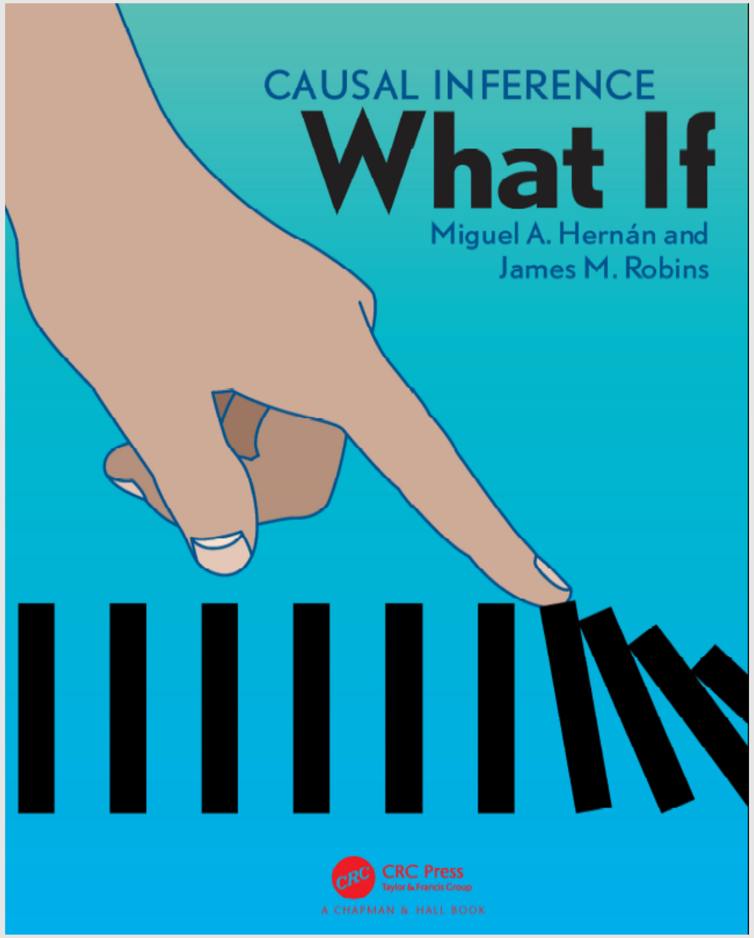
Ex: Smoking is associated with many diseases.

- Not everyone who smokes develops cancer
- Not everyone who develops cancer has smoked


Hemolytic Streptococci → Streptococcal tonsillitis, Scarlet fever, Erysipelas



# The Statistical point of view on causality…(intro)



**Causal inference** is a complex scientific task that relies on triangulating evidence from multiple sources and on the application of a variety of methodological approaches.

We (as statisticians) remain *agnostic* about metaphysical concepts like causality and cause.

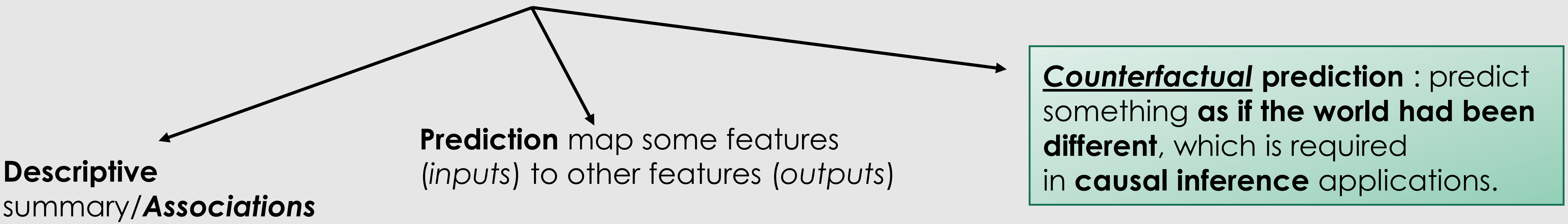
We rather focus on the identification and **estimation** of causal effects in populations, that is, **numerical quantities** that measure **changes in the distribution** of an outcome under **different interventions/exposures**.

**[*EXPLANATORY*** models framework]

https://miguelhernan.org/whatifbook

Block 1.2

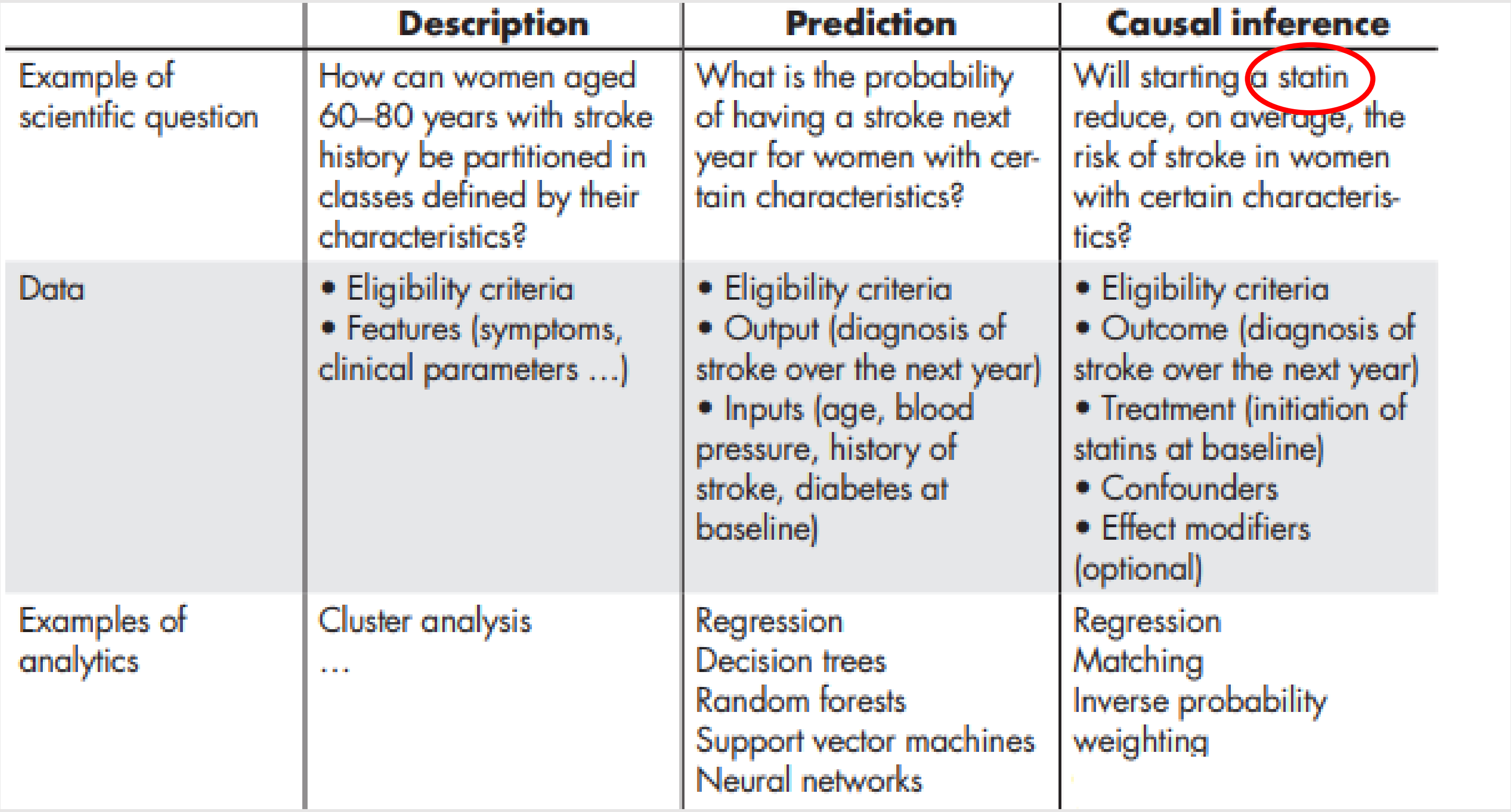
# Again : a classification of Data Science Tasks…



**Descriptive** summary/***Associations***

**Prediction** map some features (*inputs*) to other features (*outputs*)

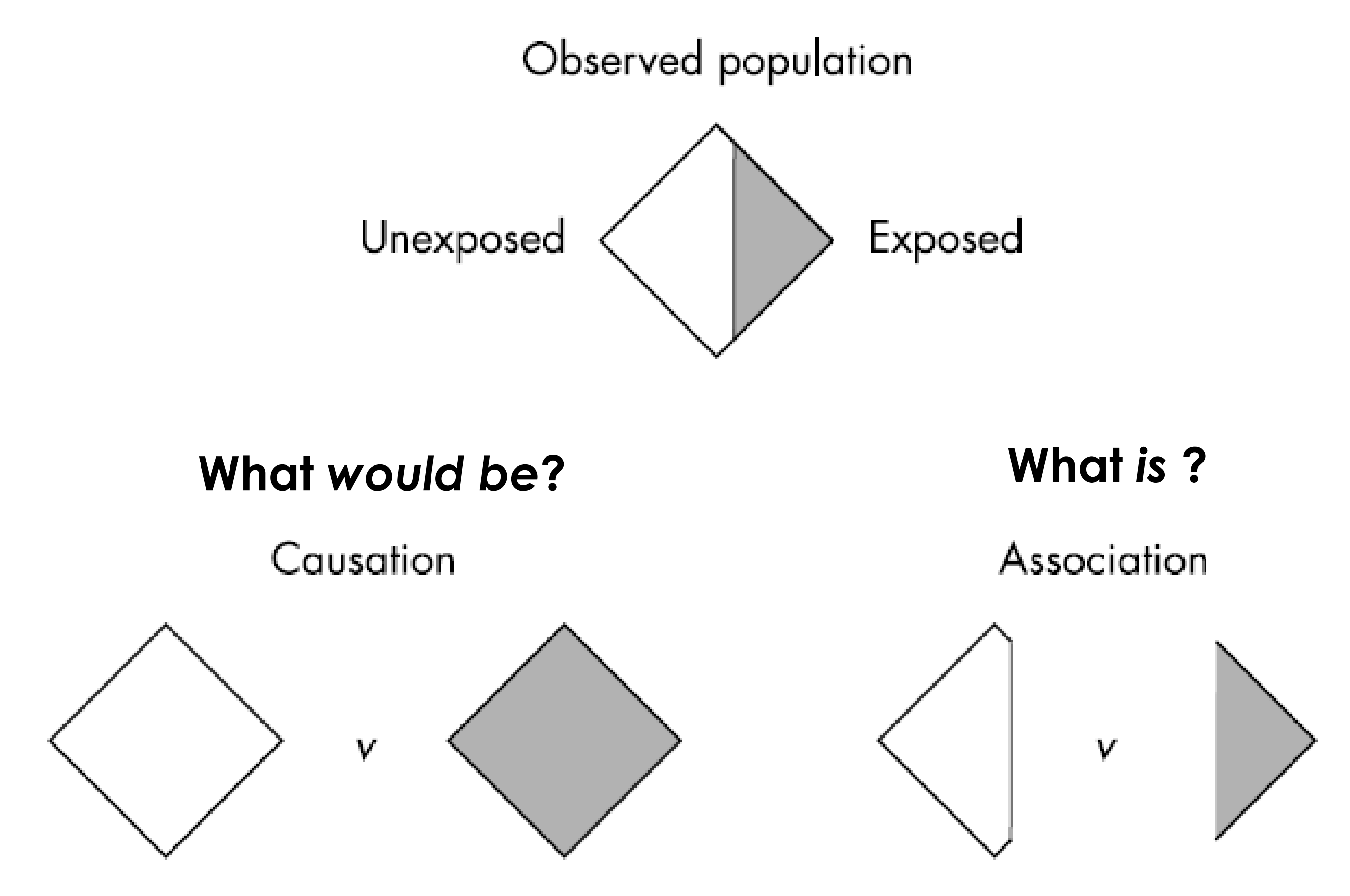
***Counterfactual*** **prediction** : predict something **as if the world had been different**, which is required in **causal inference** applications.



| | Description | Prediction | Causal inference |
|---|---|---|---|
| Example of scientific question | How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics? | What is the probability of having a stroke next year for women with certain characteristics? | Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics? |
| Data | • Eligibility criteria<br>• Features (symptoms, clinical parameters …) | • Eligibility criteria<br>• Output (diagnosis of stroke over the next year)<br>• Inputs (age, blood pressure, history of stroke, diabetes at baseline) | • Eligibility criteria<br>• Outcome (diagnosis of stroke over the next year)<br>• Treatment (initiation of statins at baseline)<br>• Confounders<br>• Effect modifiers (optional) |
| Examples of analytics | Cluster analysis<br>… | Regression<br>Decision trees<br>Random forests<br>Support vector machines<br>Neural networks | Regression<br>Matching<br>Inverse probability weighting |

https://www.hsph.harvard.edu/wp-content/uploads/sites/1268/2019/04/hernan_chance19.pdf

# What is a **causal** effect?



Comparison of **potential outcomes** for **THE SAME** well defined population:

$Y_1$ Potential outcome if treated/exposed

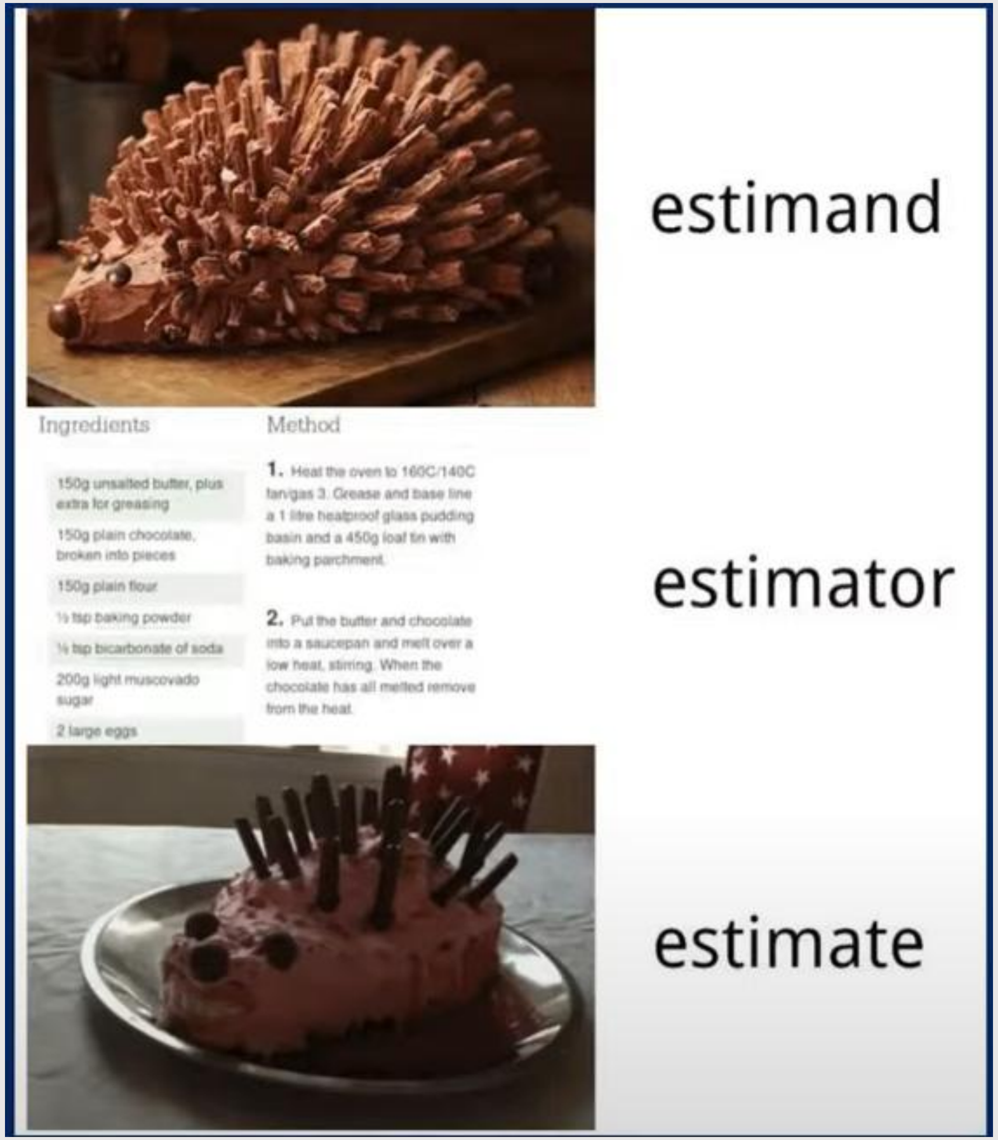
$Y_0$ Potential outcome if control (**not** treated/exposed)

An **association** compares some outcome in two ***DIFFERENT*** groups …

Figure from Hernan: https://pubmed.ncbi.nlm.nih.gov/15026432/

Block 1.2

# *The estimand* vs the *estimator*



Ingredients

- 150g unsalted butter, plus extra for greasing
- 150g plain chocolate, broken into pieces
- 150g plain flour
- ½ tsp baking powder
- ½ tsp bicarbonate of soda
- 200g light muscovado sugar
- 2 large eggs

Method

1. Heat the oven to 160C/140C fan/gas 3. Grease and base line a 1 litre heatproof glass pudding basin and a 450g loaf tin with baking parchment.

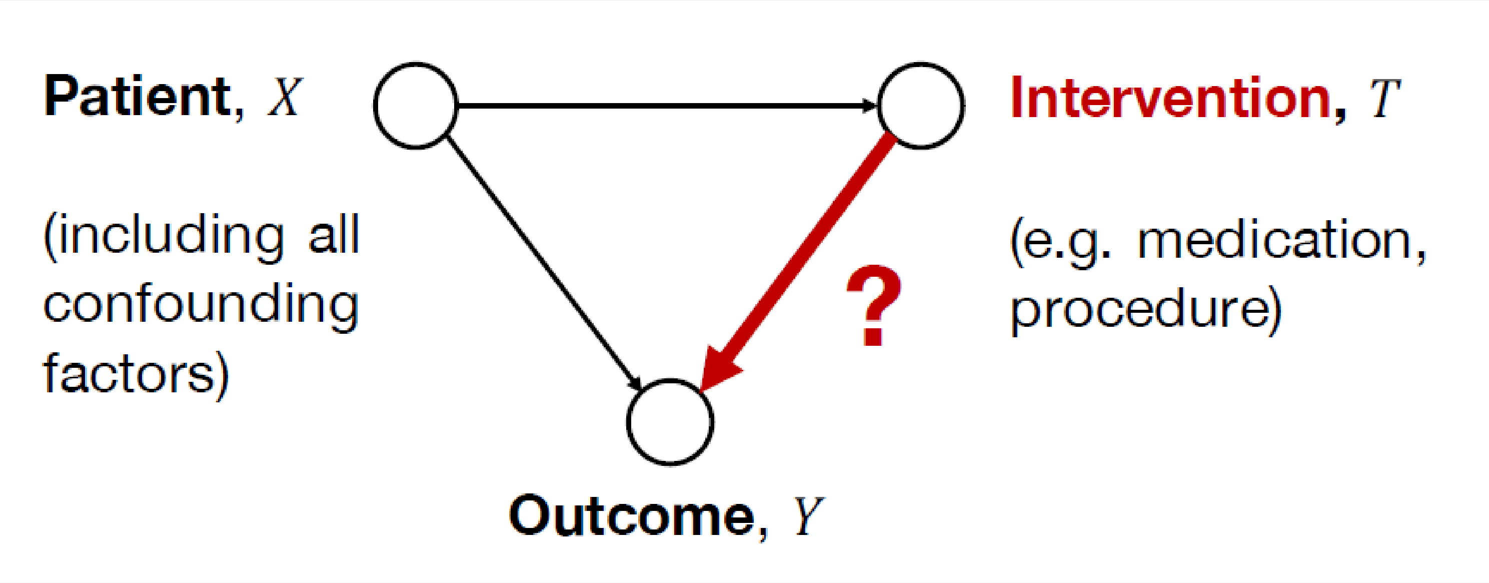
2. Put the butter and chocolate into a saucepan and melt over a low heat, stirring. When the chocolate has all melted remove from the heat.

It is confusing…in practice we often use *observed* data from two **different** groups to **estimate** a *causal* effect*…

But it is important to distinguish the ***estimand*** – the thing we want to learn about – from the ***estimator*** – how we learn about it …

* «easy process» only in **randomized** studies…

# Typical context [in observational studies]


Patient, $X$
(including all confounding factors)
Intervention, $T$
(e.g. medication, procedure)
?
Outcome, $Y$

(High dimensional…)

For simplicity think to $T_i$ is a **binary** exposure/treatment and also to the outcome as a **binary** variable *:

$$T_i = \begin{cases} 0 & \text{Untreated} \\ 1 & \text{Treated} \end{cases} \qquad Y_i = \begin{cases} 1 & \text{Yes} \\ 0 & \text{No} \end{cases}$$

*generalizable to categorical/continuous…

Block 1.2

# 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**:

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

$$\frac{\sum_i ITE_i}{n}$$

***simple to estimate*** in RCT (***randomized*** control trials)

# Potential Outcomes Framework

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

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

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

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

Observed **factual** outcome:

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

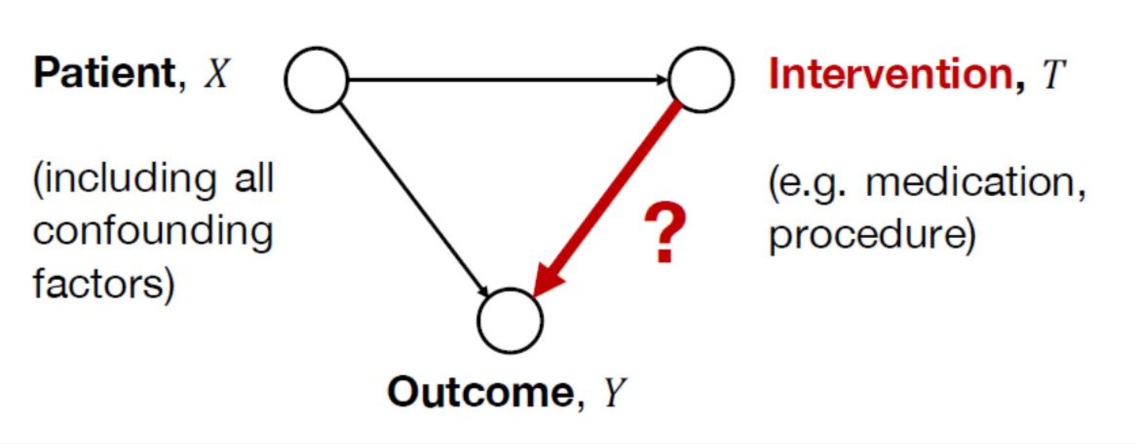
Unobserved **counterfactual** outcome:

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

The so-called **fundamental problem of causal inference** is that one can never **directly** observe causal effects, because we can never observe **both** potential outcomes ***for any individual*** (***at the same time***).

Block 1.2

# Potential Outcomes Framework


Patient, $X$
(including all confounding factors)
Intervention, $T$
(e.g. medication, procedure)
?
Outcome, $Y$

If we want to *take into account* some *characteristics* of subjects [indicated by X]:

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

**Conditional** Average Treatment Effect*

among individuals with the same *covariates* X

$$ATE = E_x[E[Y_1 - Y_0 | X = x]]$$

Average Treatment Effect

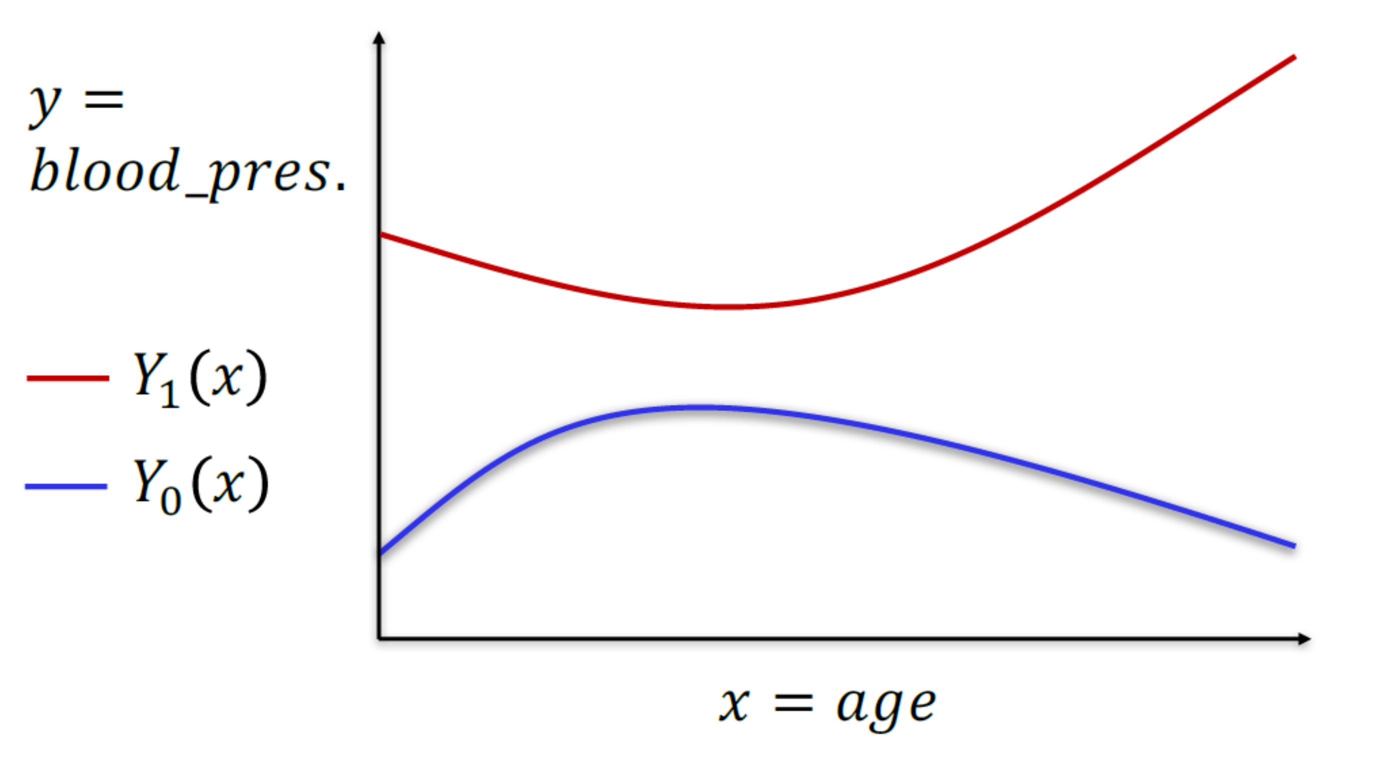
over a population *represented by the distribution* of X

* Effect is *everybody* is treated with respect to everybody is not treated **among people with X=x**

# The fundamental problem of causal inference

We only ever observe **one** of the [*two*] potential outcomes

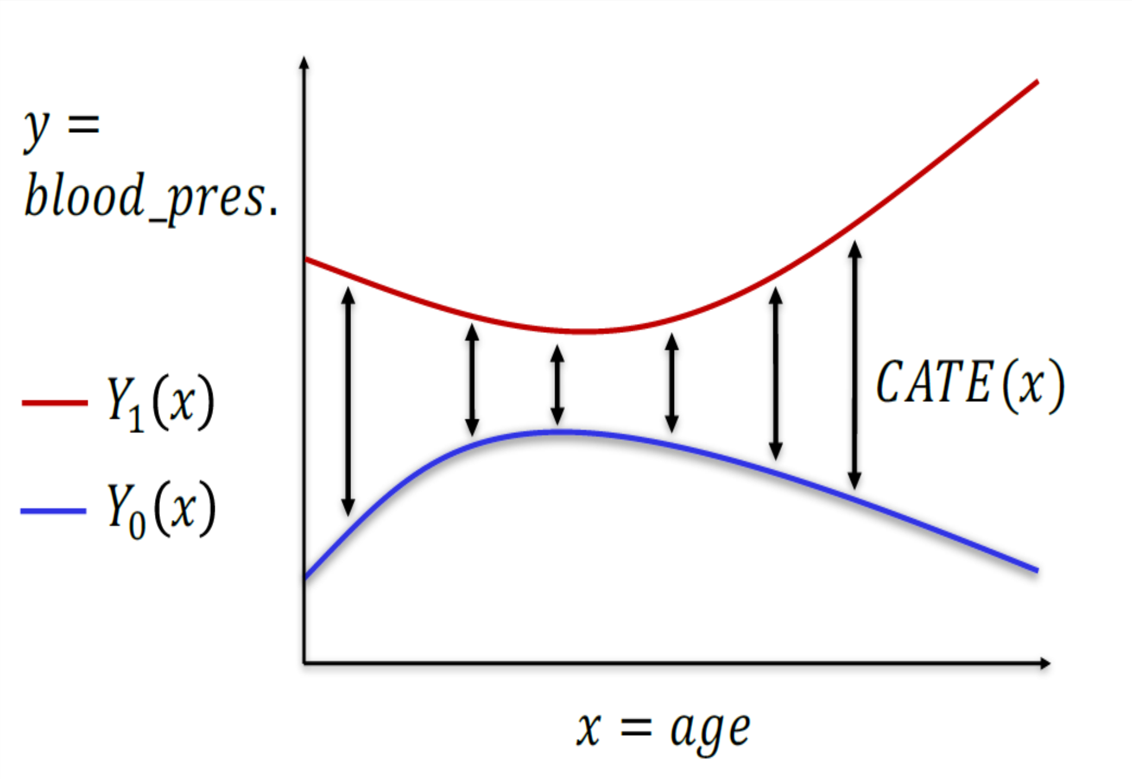
Example – Blood pressure and age:

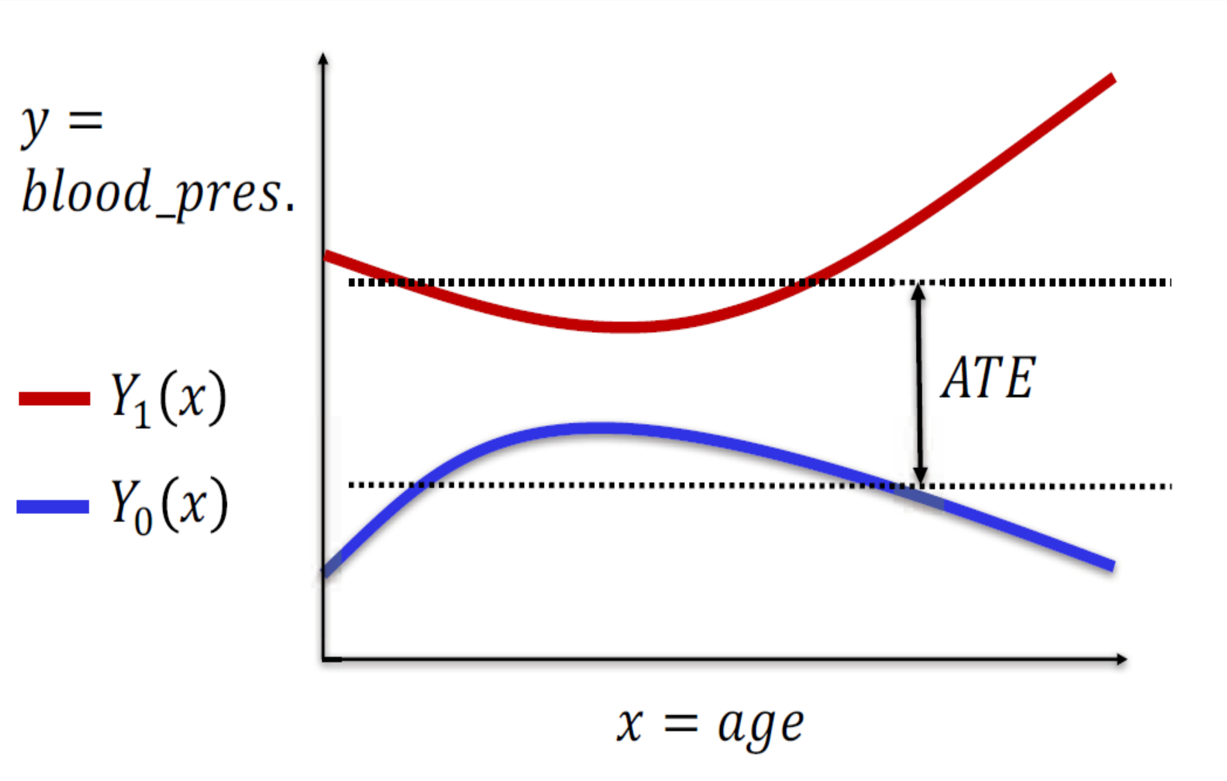


Suppose individuals are characterized by just one feature X: age.

The two curves are the ***potential*** outcomes of what would happen to blood pressure (BP) under treatment zero, (**blue** curve), or treatment one, (**red** curve).

**Block 1.2**





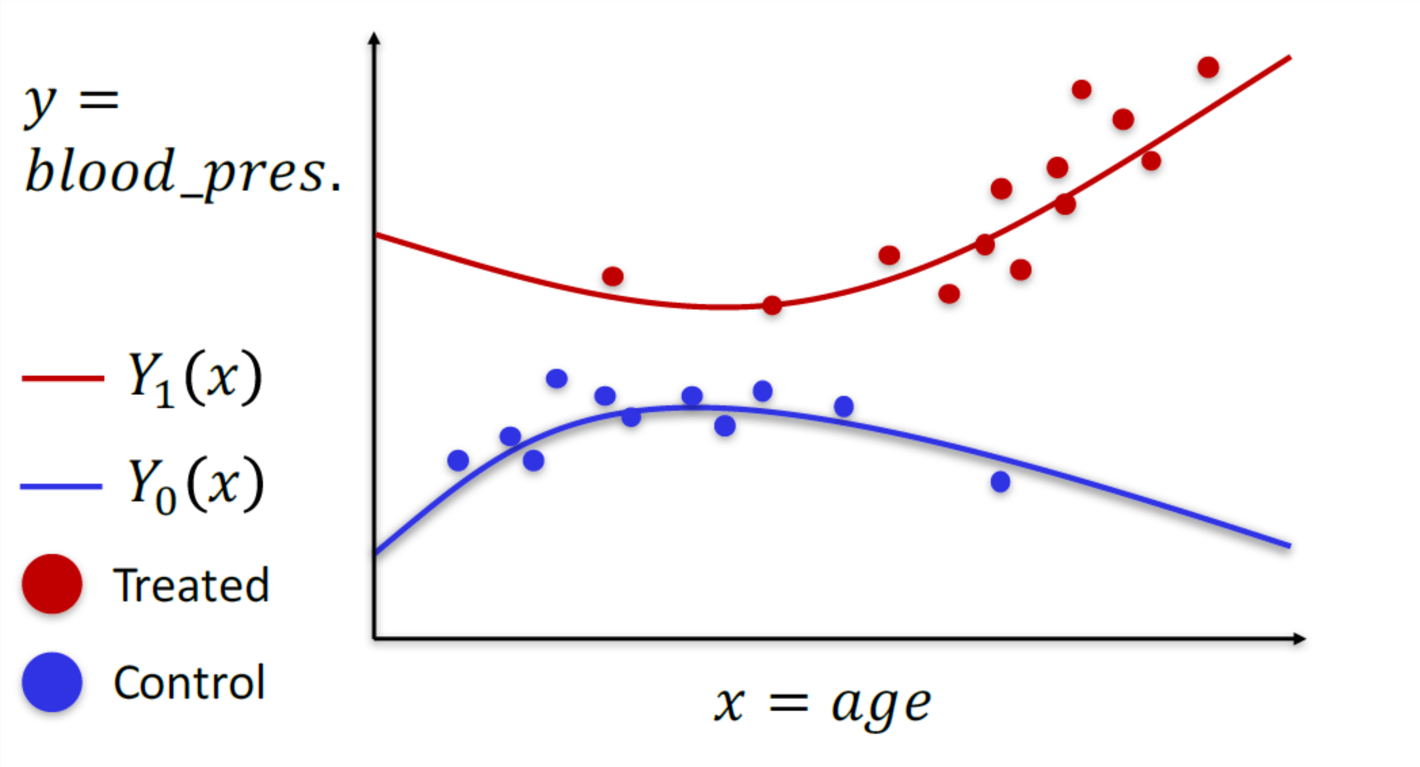
**Blue** BP is low for the young and for the elderly. For middle age, BP is in the higher range.

**Red** young people have much higher BP, and so do older people.

What about the ***difference*** for each subject at certain age ? That is the **CATE** effect

And what about the ***observed*** data ??

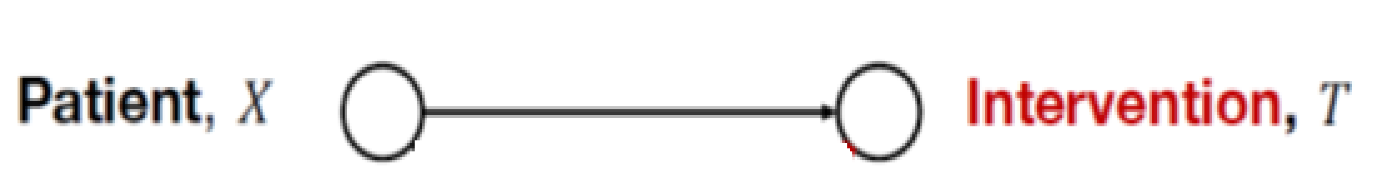
**Block 1.2**



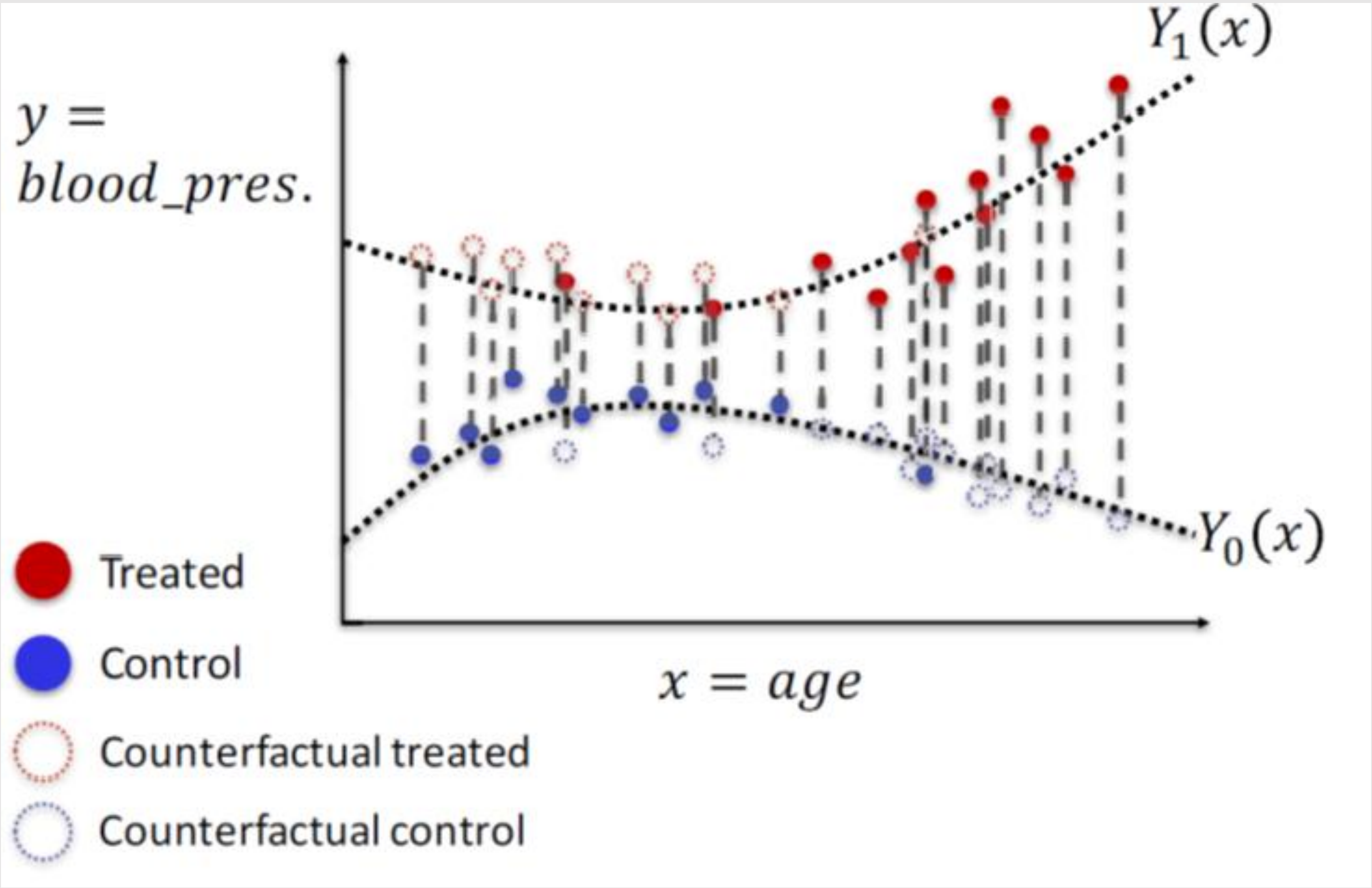
We **observe** data points that might be **unevenly** distributed [esp. in **observational studies**].

**Blue** treatment happens to be *given more* to young, and **red** *more to older* people.

Variety of reasons: access to medication, socioeconomic reasons, existing treatment guidelines…..



Block 1.2



For each subject, **what would have happened** if he/she had gotten ***the other*** treatment?

→ counterfactuals/potential outcomes.

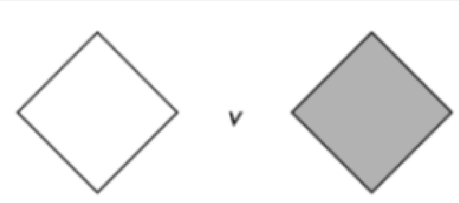
Dotted circles are the counterfactuals outcomes and the filled circles are the ***observed ones***.

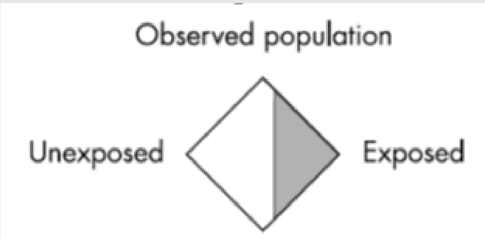
# ***Extreme*** scenario (with some numbers)

## ***Potential*** Outcomes

| Subject | Age | BLUE PILL | RED PILL |
|---|---|---|---|
| 1 | 25 | **90** | *85* |
| 2 | 30 | **95** | *90* |
| 3 | 35 | **98** | *90* |
| 4 | 40 | **100** | *96* |
| 5 | 41 | **105** | *98* |
| 6 | 42 | **110** | *111* |
| 7 | 43 | **108** | *107* |
| 8 | 44 | **110** | *111* |
| 9 | 45 | **108** | *107* |
| 10 | 46 | *100* | **120** |
| 11 | 47 | *110* | **130** |
| 12 | 48 | *120* | **135** |
| 13 | 49 | *100* | **140** |
| 14 | 50 | *105* | **115** |
| 15 | 55 | *110* | **120** |
| 16 | 60 | *108* | **120** |
| 17 | 65 | *110* | **140** |
| 18 | 70 | *108* | **150** |
| 19 | 75 | *110* | **160** |

$$ATE = \frac{\sum_i ITE_i}{n} = 12$$



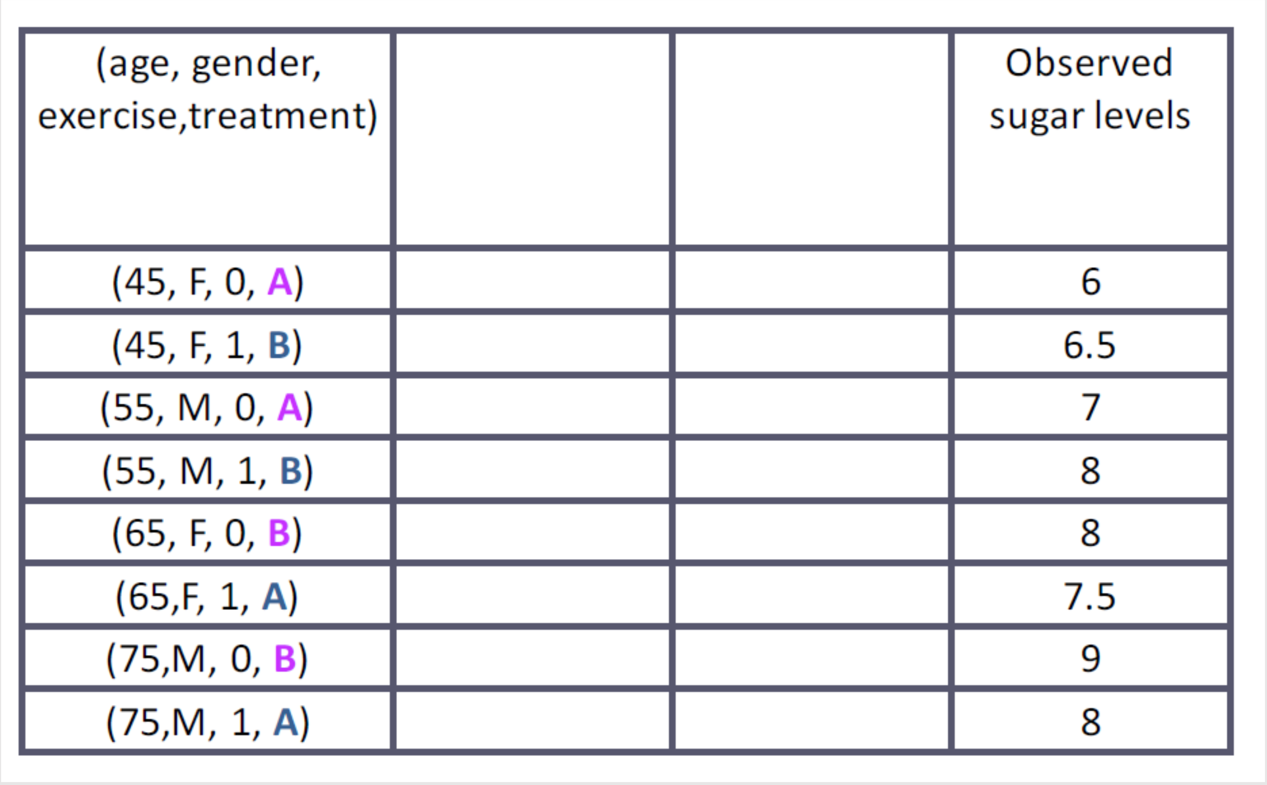


$$E(BP|red\ pill) - E(BP|blue\ pill) = 30$$

## ***Observed*** Outcomes

| Subject | Age | BLUE PILL | RED PILL |
|---|---|---|---|
| 1 | 25 | 90 | |
| 2 | 30 | 95 | |
| 3 | 35 | 98 | |
| 4 | 40 | 100 | |
| 5 | 41 | 105 | |
| 6 | 42 | 110 | |
| 7 | 43 | 108 | |
| 8 | 44 | 110 | |
| 9 | 45 | 108 | |
| 10 | 46 | | 120 |
| 11 | 47 | | 130 |
| 12 | 48 | | 135 |
| 13 | 49 | | 140 |
| 14 | 50 | | 115 |
| 15 | 55 | | 120 |
| 16 | 60 | | 120 |
| 17 | 65 | | 140 |
| 18 | 70 | | 150 |
| 19 | 75 | | 160 |

Treatment is considered ***effective*** if the difference in BP > 20 points…



| (age, gender, exercise,treatment) | | | Observed sugar levels |
|---|---|---|---|
| (45, F, 0, A) | | | 6 |
| (45, F, 1, B) | | | 6.5 |
| (55, M, 0, A) | | | 7 |
| (55, M, 1, B) | | | 8 |
| (65, F, 0, B) | | | 8 |
| (65,F, 1, A) | | | 7.5 |
| (75,M, 0, B) | | | 9 |
| (75,M, 1, A) | | | 8 |

mean(sugar|medication B)-mean(sugar|medication A) = ?

**7.875 -7.125 = 0.75**

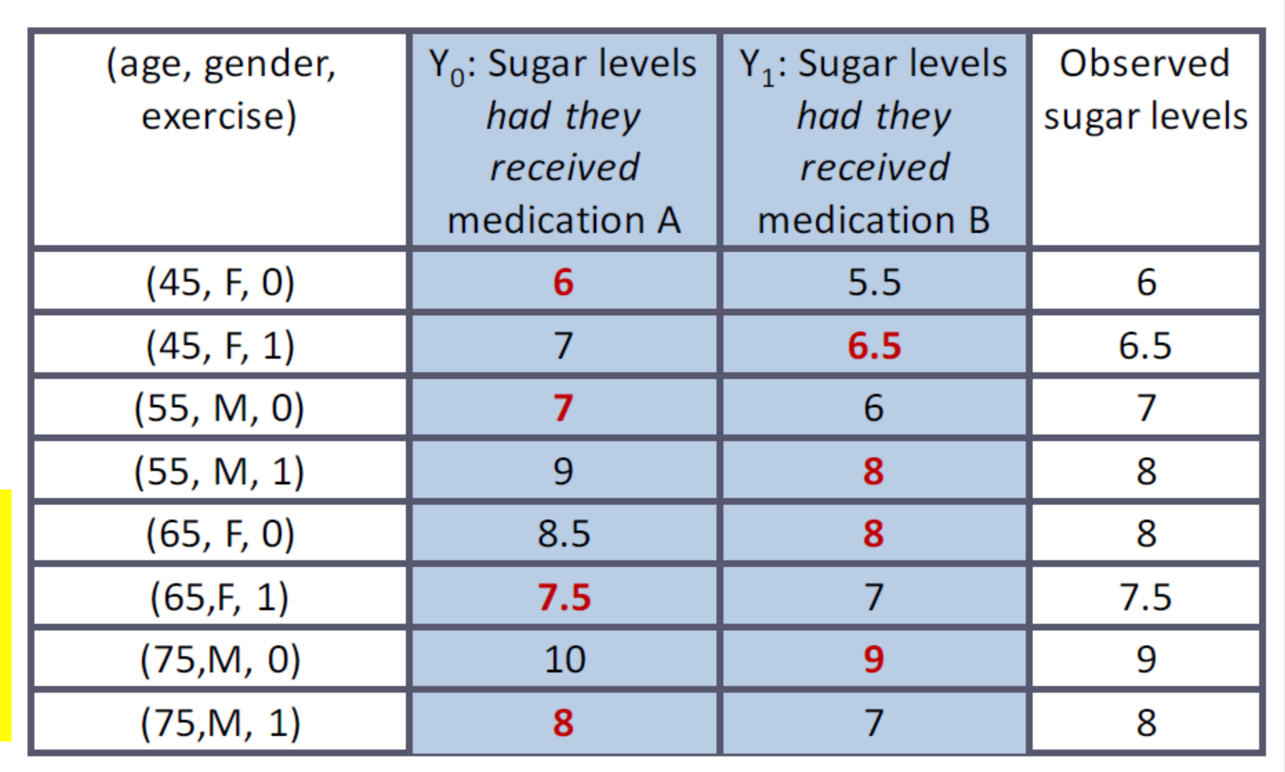
To **solve** the problem, we have to **make some <u>assumptions</u>**…(in **observational studies, in RCTs it is easier !**)

age, gender, whether they exercise regularly, what treatment they got, which is A or B.

Observed sugar glucose levels at the end of the treatment.

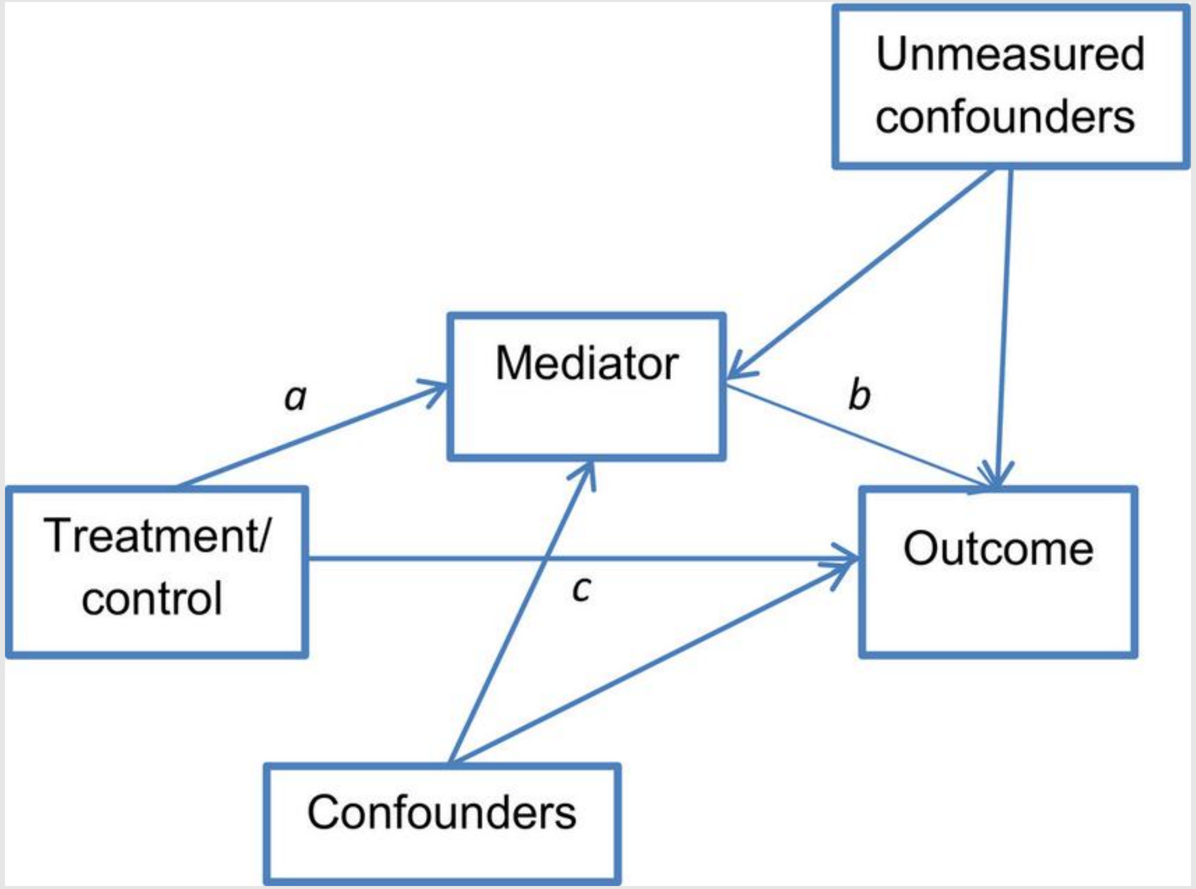
mean(sugar|*had they received B*) – mean(sugar|*had they received A*) = ?

**7.125 -7.875 = -0.75**



| (age, gender, exercise) | $Y_0$: Sugar levels *had they received* medication A | $Y_1$: Sugar levels *had they received* medication B | Observed sugar levels |
|---|---|---|---|
| (45, F, 0) | **6** | 5.5 | 6 |
| (45, F, 1) | 7 | **6.5** | 6.5 |
| (55, M, 0) | **7** | 6 | 7 |
| (55, M, 1) | 9 | **8** | 8 |
| (65, F, 0) | 8.5 | **8** | 8 |
| (65,F, 1) | **7.5** | 7 | 7.5 |
| (75,M, 0) | 10 | **9** | 9 |
| (75,M, 1) | **8** | 7 | 8 |

**Block 1.2**



Stay tuned !!... Something more in block 2 and 3…