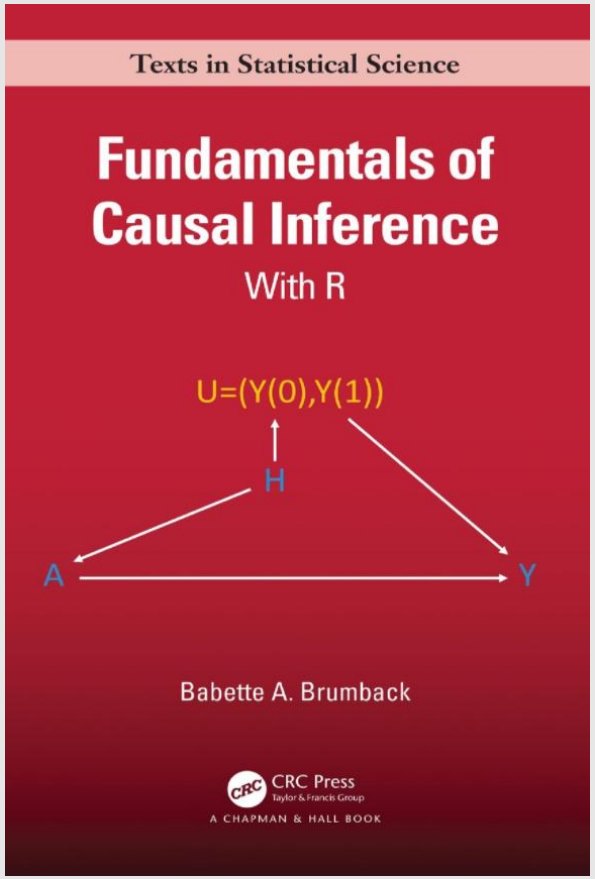
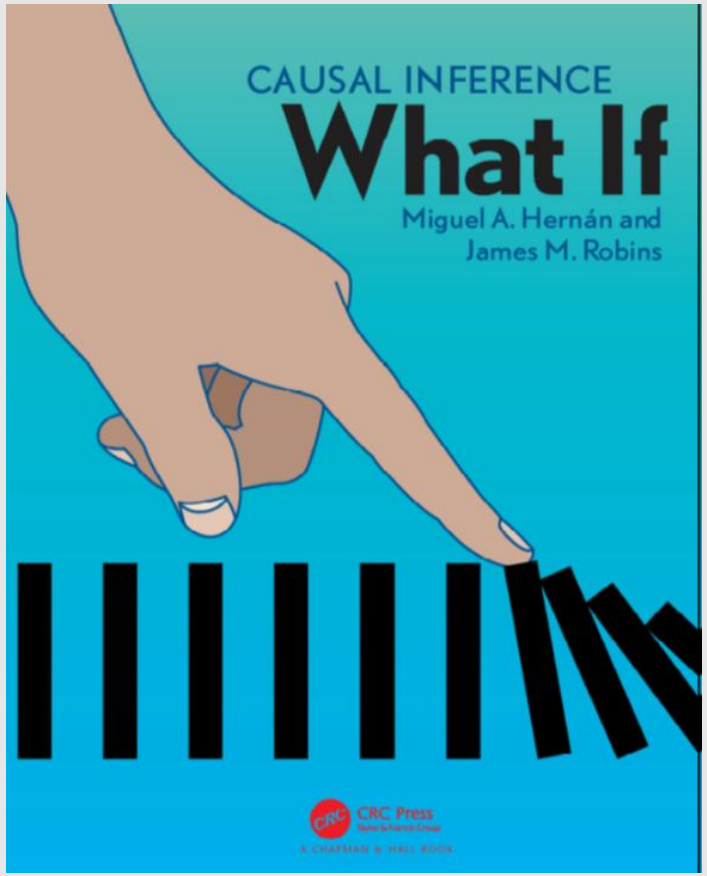
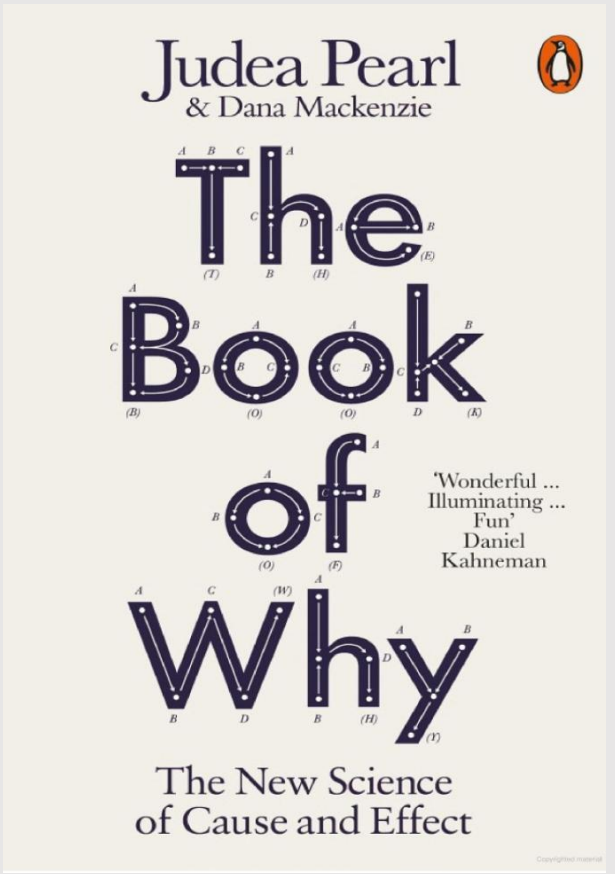
Causal Models in Epidemiological & Clinical Research (basic ideas…)







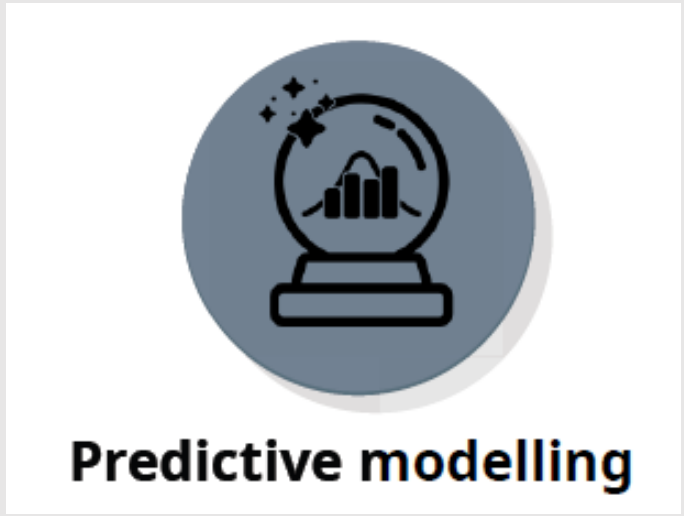
Learning without making assumptions is a myth

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

Different scientific aims

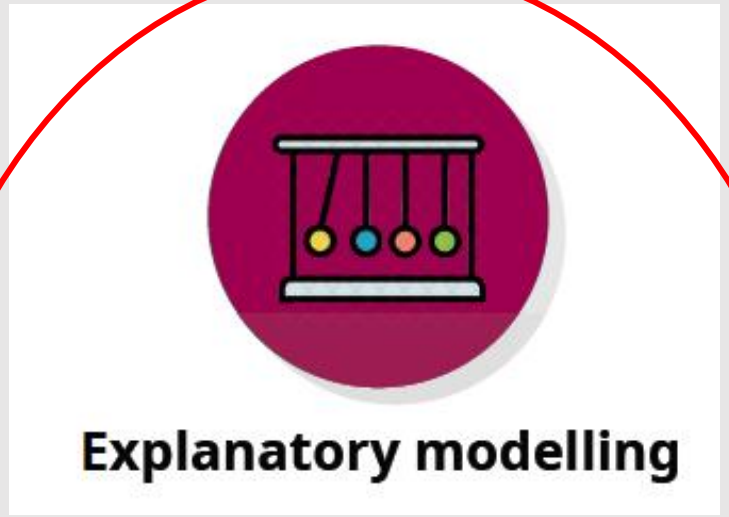
Describe the outcome of interest:
which factors *affect* it and how?

Estimate a *prevalence* in function of age and sex



Accurate predictions of future observations.
No concern about causality and confounding (association)

Risk of developing CVD in the next x years



Testing and comparing existing causal theories.

Effect of LDL on CVD risk

Policy recommendations (statins)

Independently from the primary scientific aim, a **basic statistical tool** is the REGRESSION model approach.

The key difference is in **the scope of** the REGRESSION model according to the main aim...

The **Descriptive** Aim (The "What")

Regression as a tool for **parsimonious** summarization.

Describing the Population: How regression describes the *average* individual.

Trend Identification: Using splines or polynomials to describe non-linear patterns

Data Reduction: Using regression to simplify complex datasets into understandable trends.

The Predictive Aim (The "Who")

Focus on accuracy and generalizability of predictions.

The **Bias-Variance** Tradeoff: Why "more variables" isn't always better.

Feature Selection: we don't care about the specific role of a variable (if it is a "confounder" or "independent predictor"..) only if *it adds signal*.

Validation: The necessity of Cross-Validation and External Validation.

The Causal Aim (The "Why")

Focus on unbiased estimation of an **effect** of an intervention/exposure.

Counterfactual Framework: Potential Outcomes.

Directed Acyclic Graphs (DAGs): How to choose variables based on theory, not (just) p-values…

Identifiability: assumptions of ***exchangeability, positivity, and consistency.***

Explanatory/Causal Models (Observational Studies)

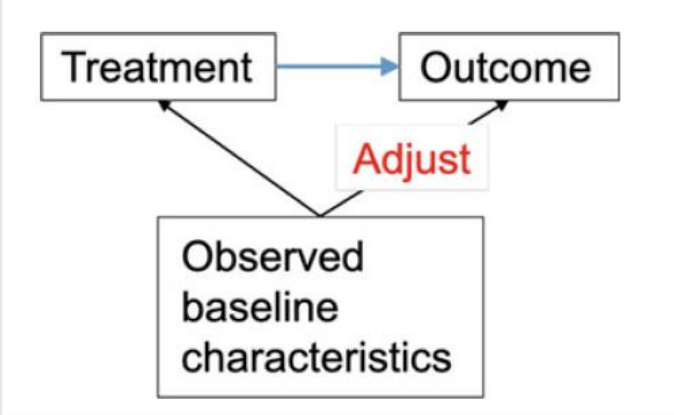
Confounding is one of the major concern in epidemiological analyses of observational studies, where we aim to estimate **causal** effects.

When treatments/exposures are compared, groups are often quite different **because of a lack** of randomization.

For example, subjects with specific characteristics ***are more likely*** to receive a certain treatment than other subjects (*confounding by indication*).

If these characteristics also affect the outcome, a direct comparison of treatments is **biased** and may merely reflect the lack of initial comparability.

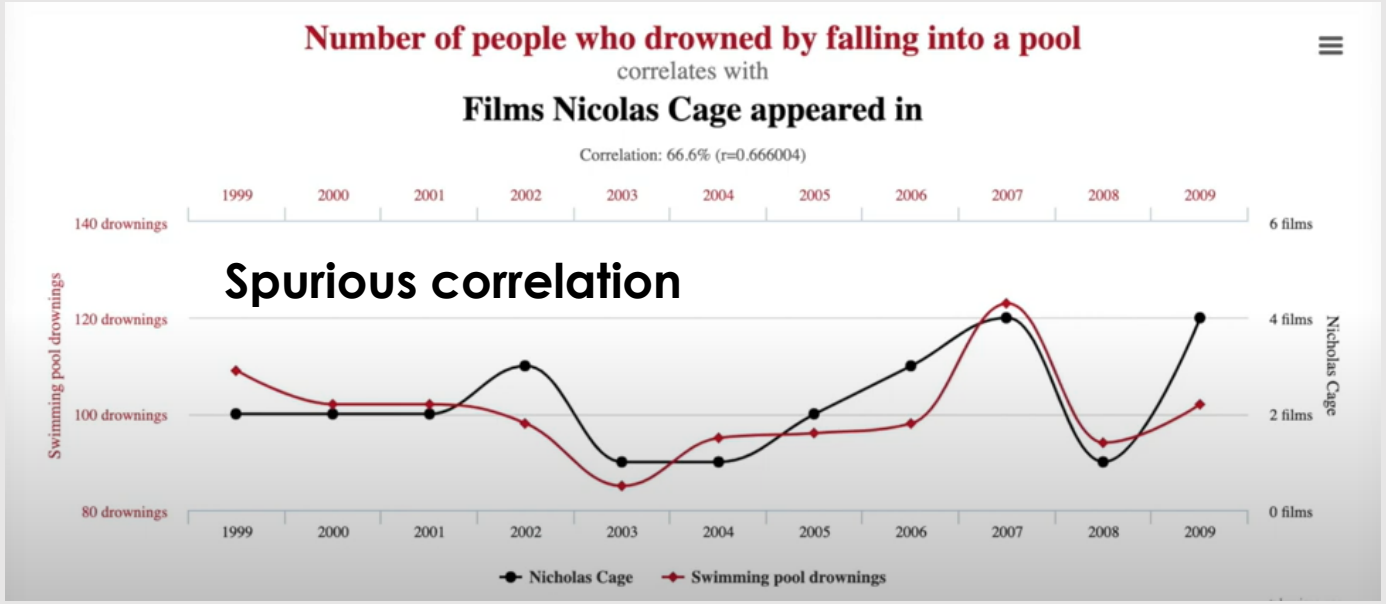
Often, randomization is not possible, and **observational studies are the only possible design**. Dealing with confounding is an essential step in such analyses.



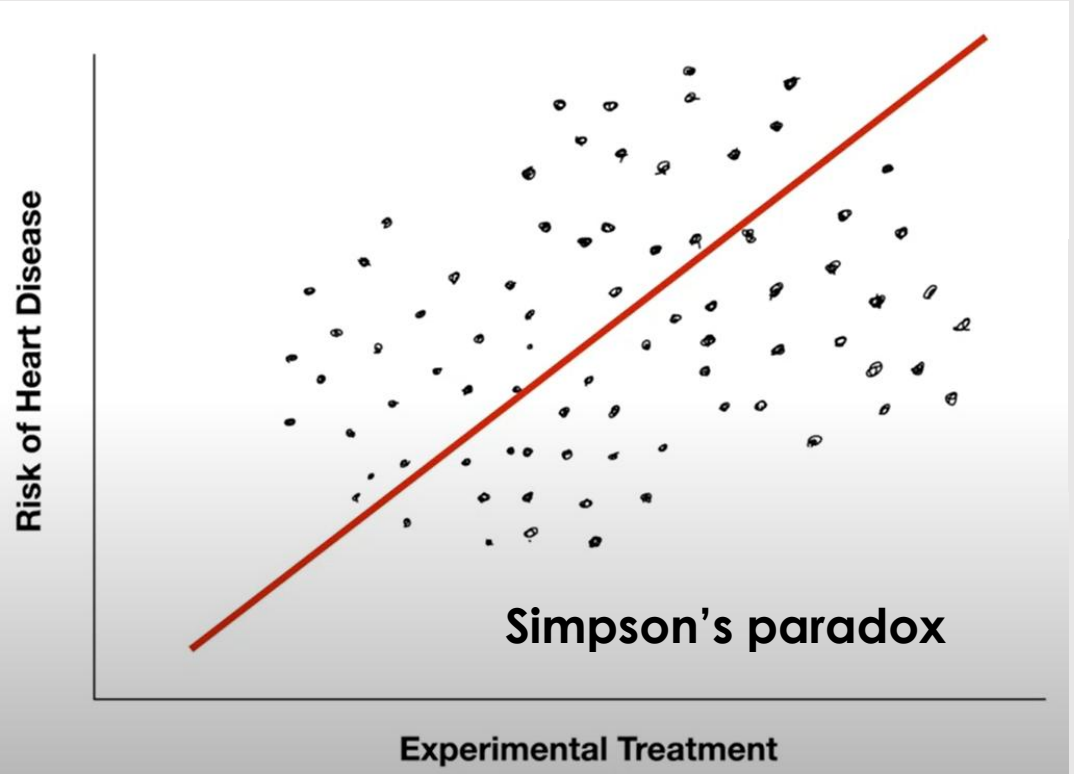
what is the treatment effect if baseline characteristics were *similar* between treatment groups?

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

Block 3.4

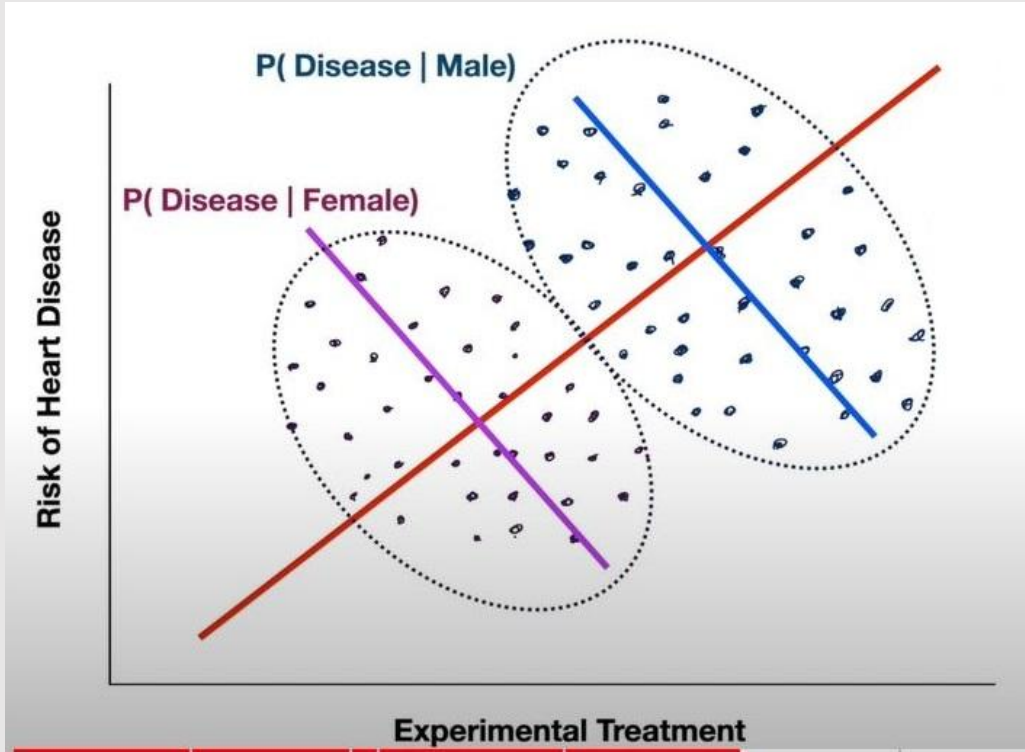


Spurious correlation



Simpson's paradox

Treatment is good for a man, good for a woman but bad for a person



In this context, **causal inference** methods must bridge a gap between goals and means.

Researchers seek **causation**, but the data, on their own, only communicate **associations**.

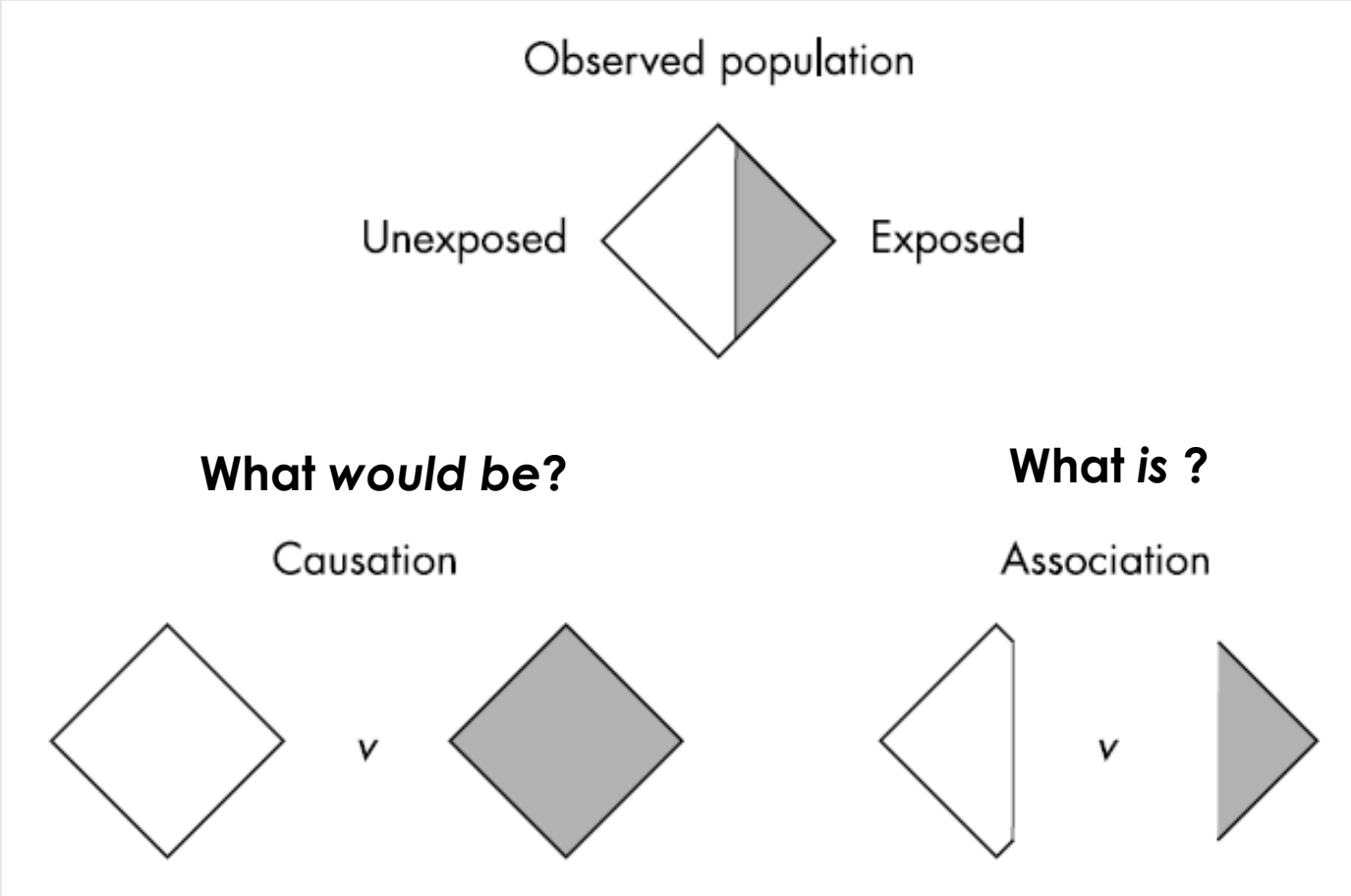
Associations usually consist of a **mixture** of causal and non-causal (spurious) components.

Therefore, a first step of ***identification analysis*** should determine whether, and under which conditions, it is possible to strip an observed association of all its spurious components.

Identification analysis requires causal **assumptions** about how the data were generated.

The sum of these causal assumptions is called a **causal model**, which must describe both *how the world works* (how observed and unobserved variables take their values) and *how the data were collected* (**which** variables and variable values are recorded/which **study design**).

REMIND: 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)

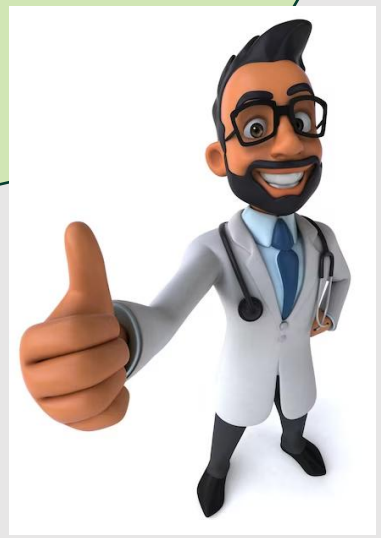
An **association** compares some outcome in two ***DIFFERENT*** groups …

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

How we **select** variables in the ***causal*** model ?

Depends on the scientific aim!

Theory-driven

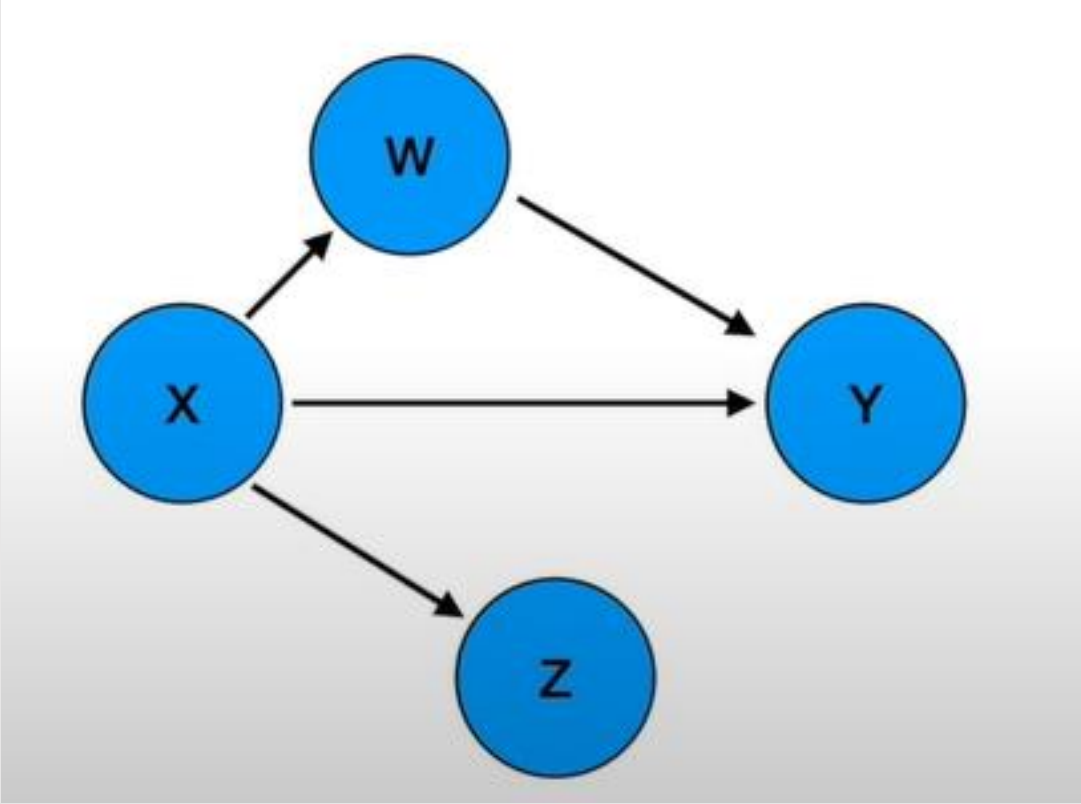


Data-driven

Explanatory modelling

Causality

X causes Y if *when all confounders are adjusted*, an *intervention* in X **results in a change** in Y, but intervention in Y **does not change** X



A causal model could be represented using a **DAG** (**D**irected **A**cyclic **G**raph)

Each arrow represents a causal influence

The graph is:

- **Directed** : each connection between two variables consists of an arrow
- **Acyclic** : no reverse* cycles

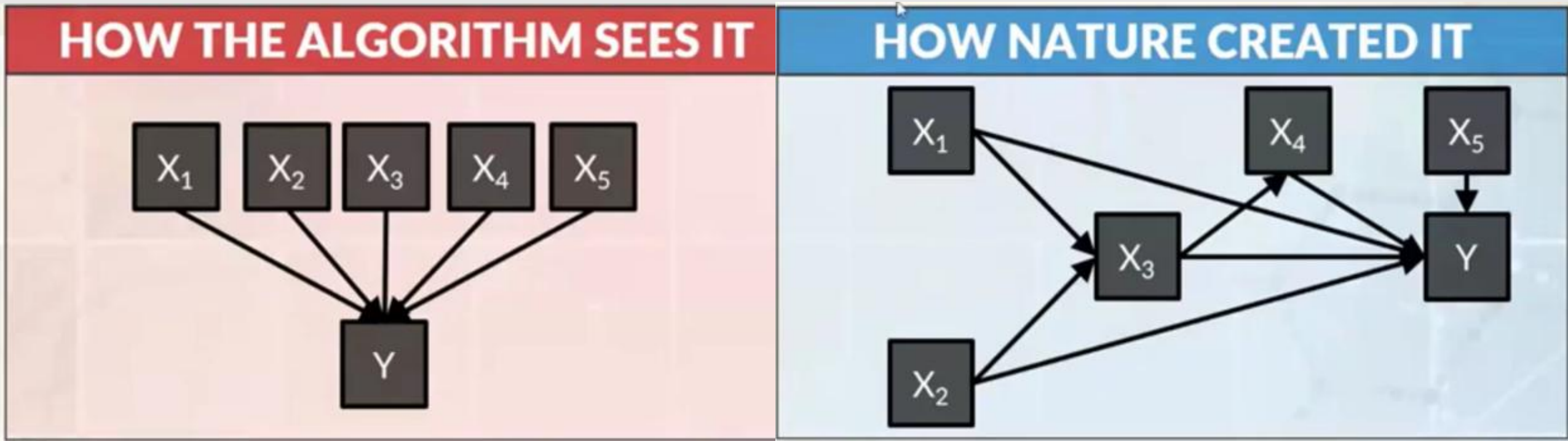
*A variable can't cause itself; however time varying processes can be depicted adding one *realization* of each variable per time unit

Knowledge of the data-generating mechanism has to be provided by **external theory** and **understanding**

→ i.e. a **causal** model

No software/algorithm can (currently) ***understand*** this

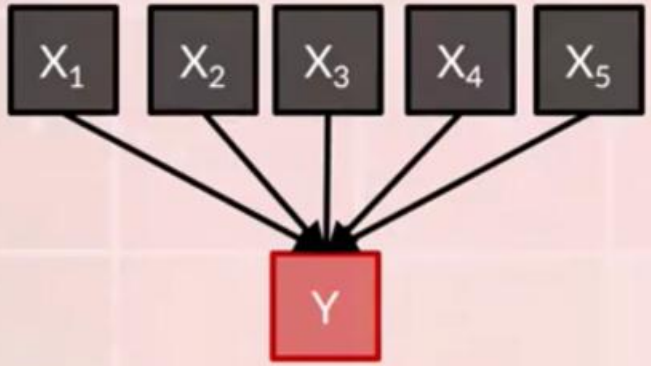
Prediction models can not be in general *causally interpreted* – however transparent they are …



Predictive vs causal modelling

PREDICTIVE MODEL

- **Outcome**-focused

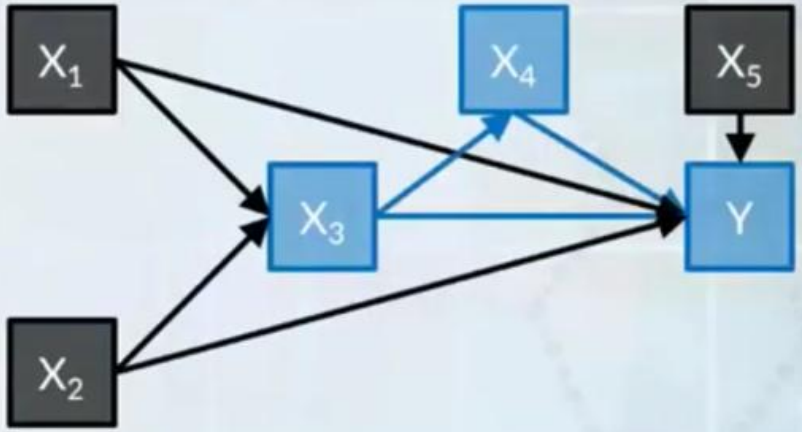


X_i is ____ Y

'correlated with'
'a predictor of'
'associated with'

CAUSAL MODEL

- **Effect**-focused



The ____ of X_2 on Y is...

'total causal effect'
'direct causal effect'

Predictive vs causal modelling

Predictive Model

Aim: Predict values of outcome (p)

Maximise: performance measures (R^2, calibration, AUC…)

Covariates selection focused on:

- Balancing precision & parsimony
- Availability of variables
- Maximising joint information

Coefficients: *associations*

Automation: Favoured

Causal Model

Aim: Estimate a causal effect

Maximise: accuracy of the effect estimate

Covariates selection focused on:

- External knowledge & judgement
- Role of variables
- Minimizing confounding

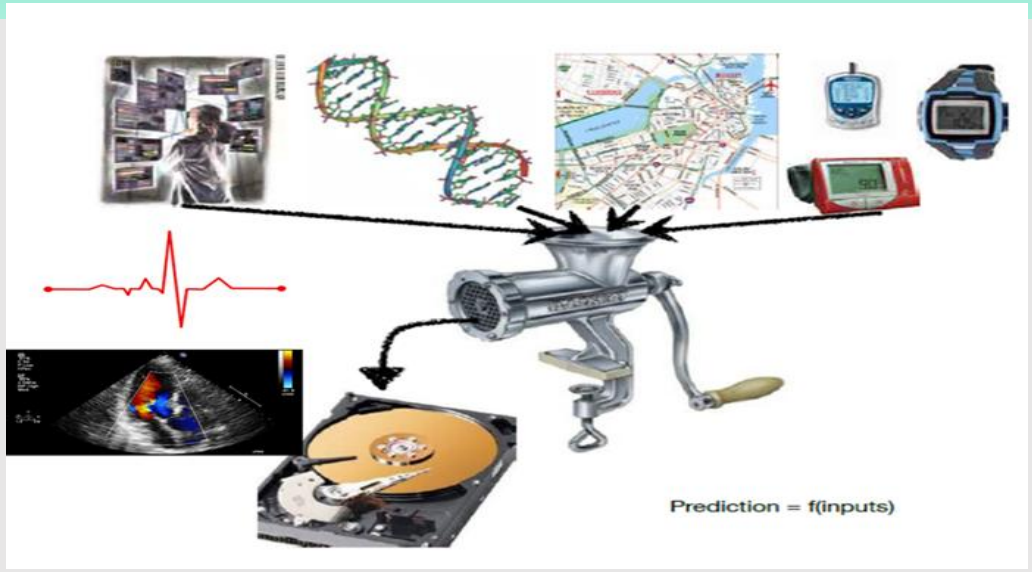
Coefficients: *Interpretable in the causal sense*

Automation: Not possible

Two *extremes* in regression models

Prognostic/Predictive modelling

- We are trying to find **predictors** of some outcome
- It is their **joint value** as predictors that is important
- We simply want the **most predictive** model
- We compare entire models to judge which is best



Causal/explanatory analysis

- The putative causal factor **must be** in the model
- Other factors are in the model because **help us understand** the causal factor (they are of no interest in themselves)
- We focus on the **estimation** of the putative causal effect (at a population level)

Smoking & Lung Cancer
A Tale of Two Statisticians

Works in public health (**explanatory**):

- I wish to establish whether it is causal
- If so I can warn smokers to quit and this will benefit their health [**intervention**]
- It is important for me to rule out possible confounding factors

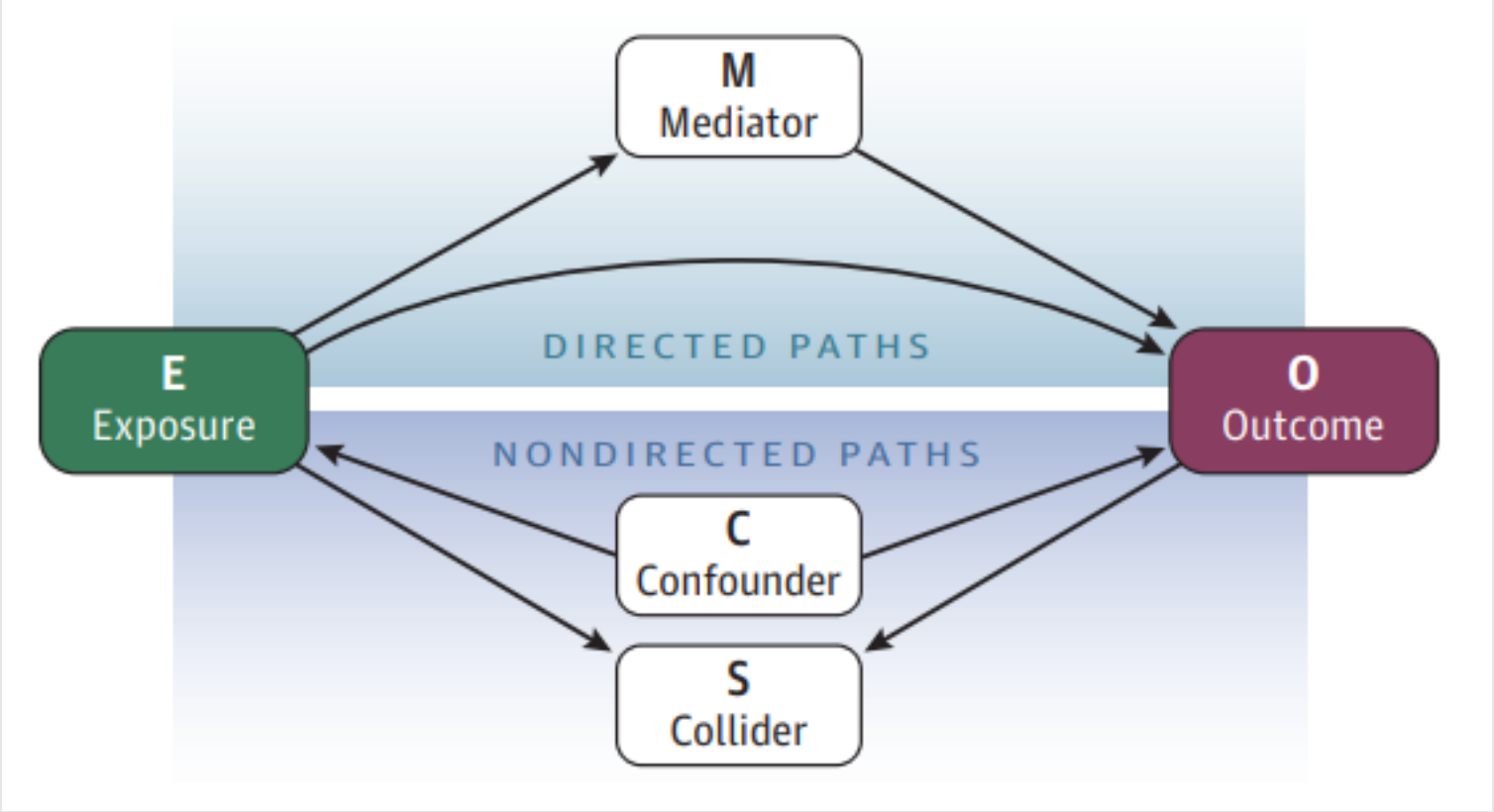
Works in life insurance (**predictive**):

- I don't care if it is causal or not
- The data show that smokers are much more likely to get lung cancer
- That's enough for me to take account of it in setting the premiums

…so again: how to *represent* the role of covariates ?

DAGs are visual representations of qualitative causal assumptions

They encode researchers' expert knowledge and beliefs about *how the world works*



Do you remember? Chocolate Consumption, Cognitive Function, and Nobel Laureates….

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

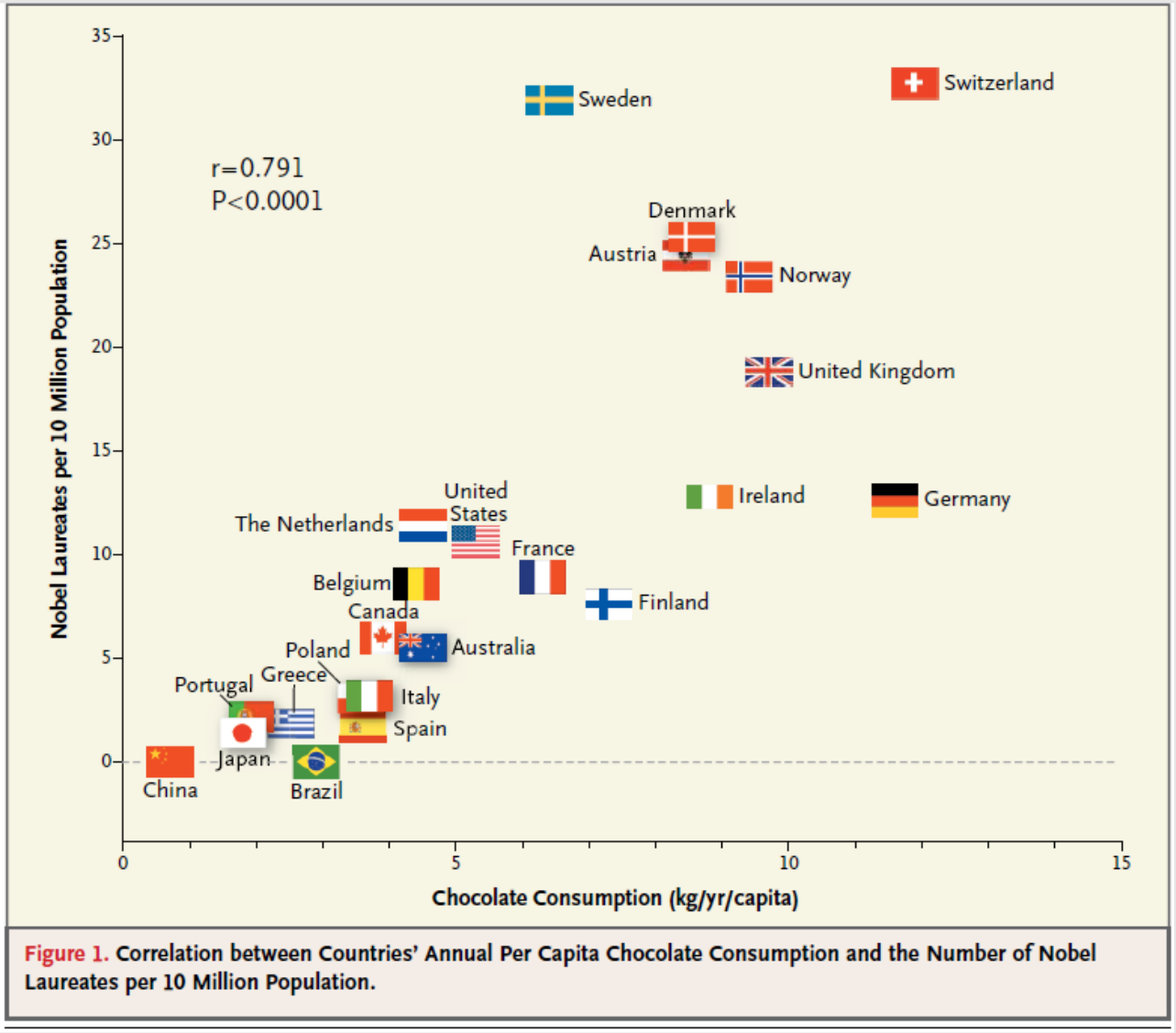
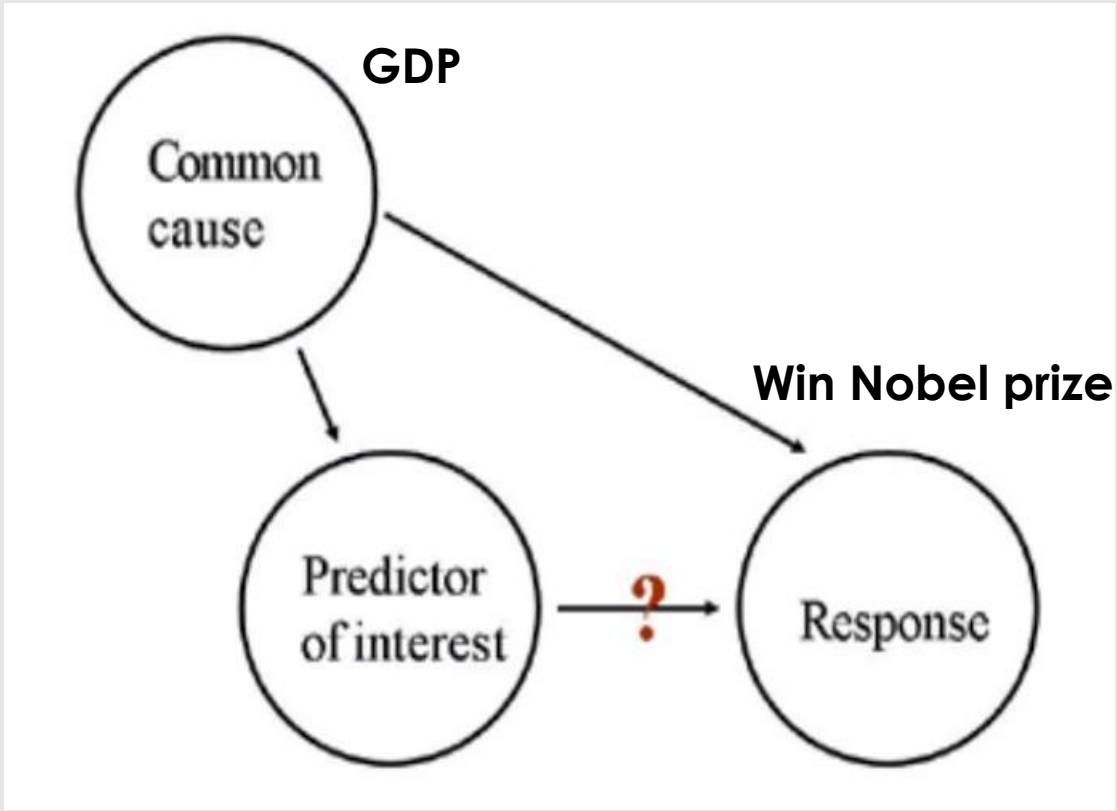
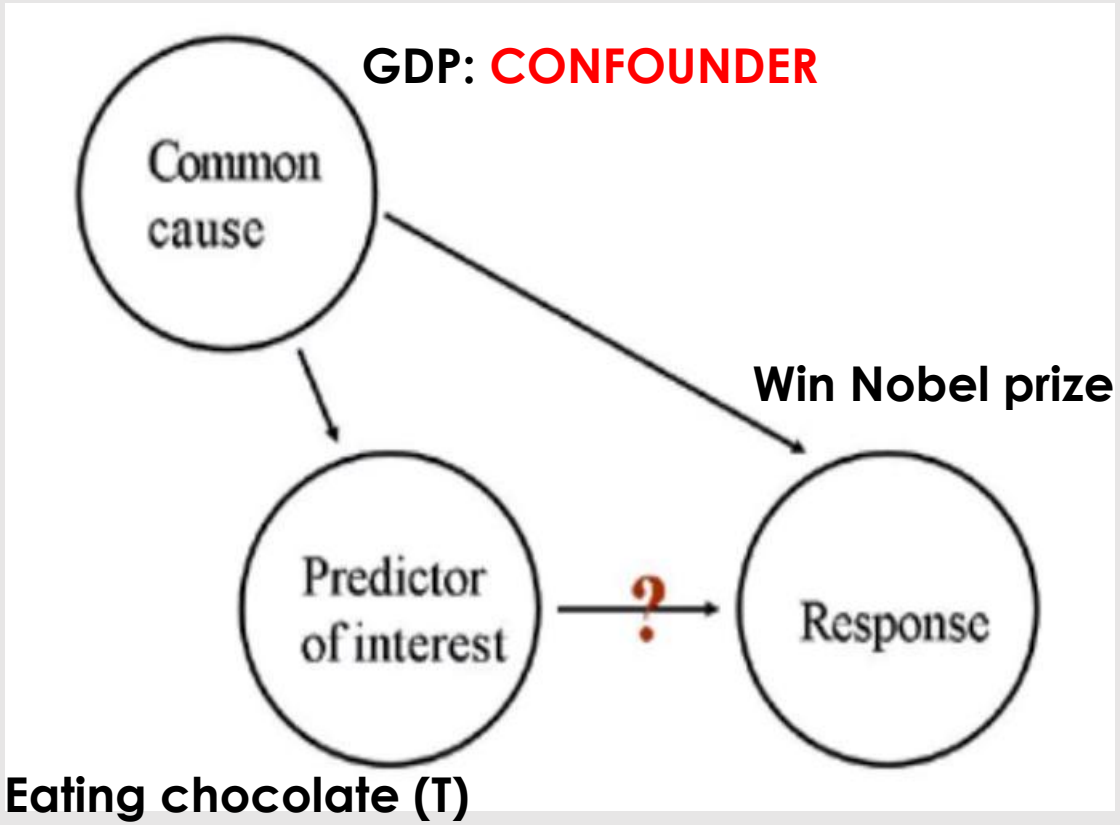


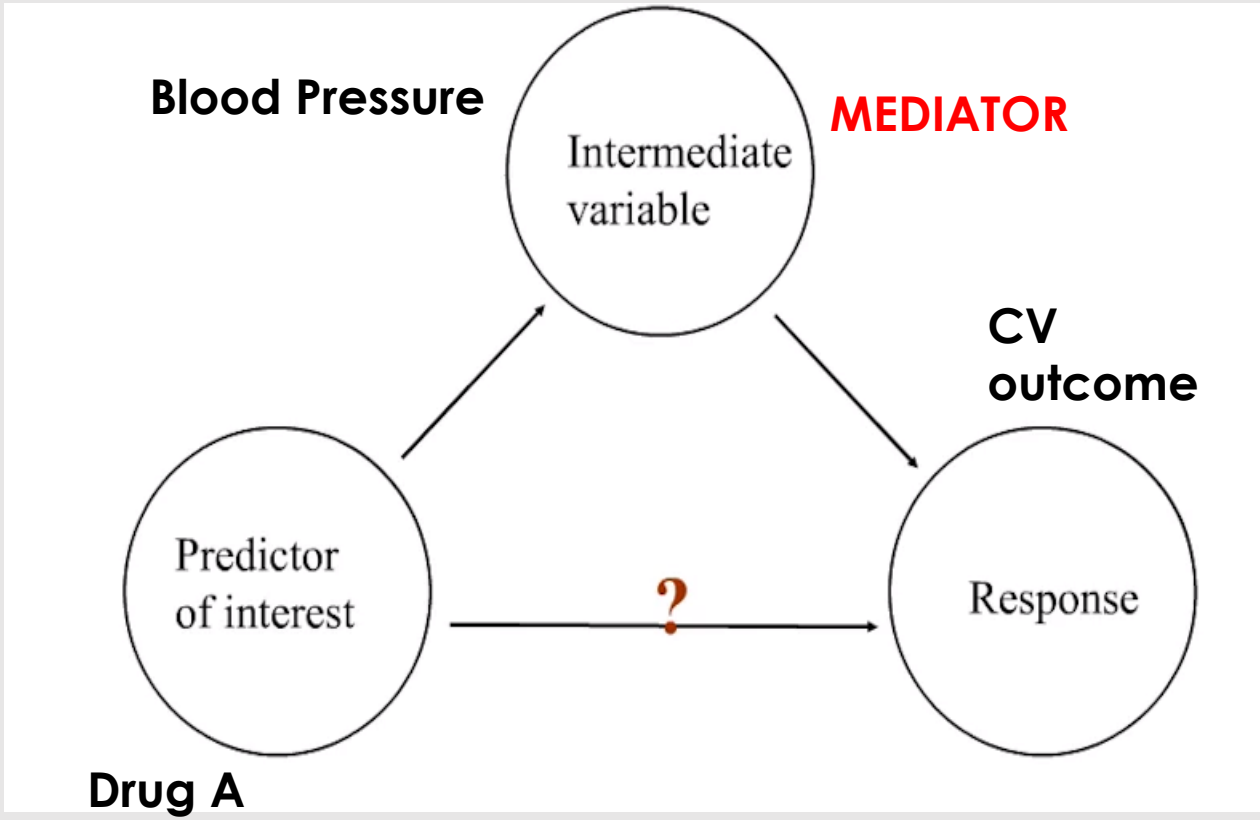
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.



Eating chocolate (T)

Examples





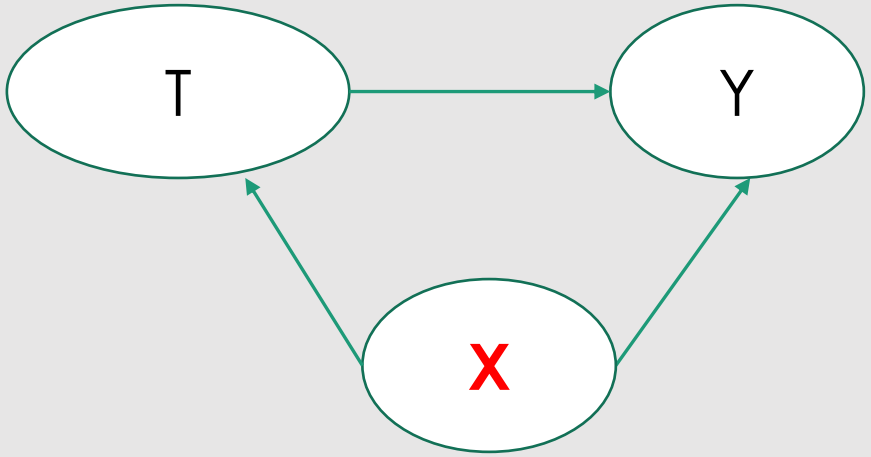
First: **Define relationships** between variables […a priori knowledge]

***Intermediate/Mediator vs Confounder: confounder IS NOT on the *causal pathway* between predictor and outcome**

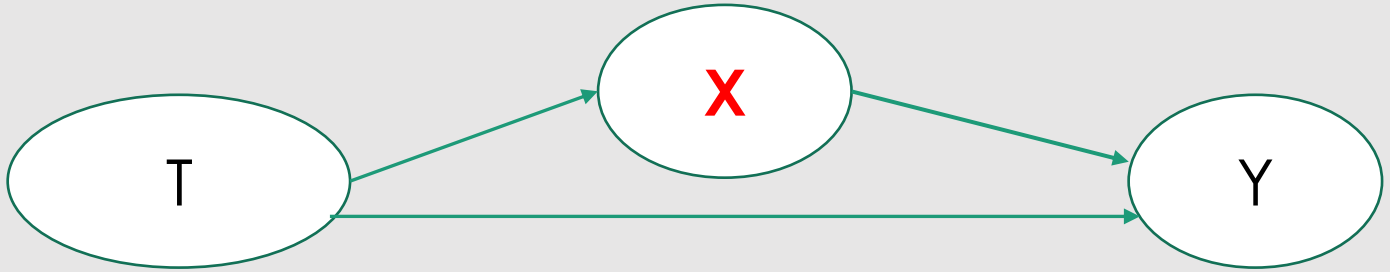
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Examples of different DAGs

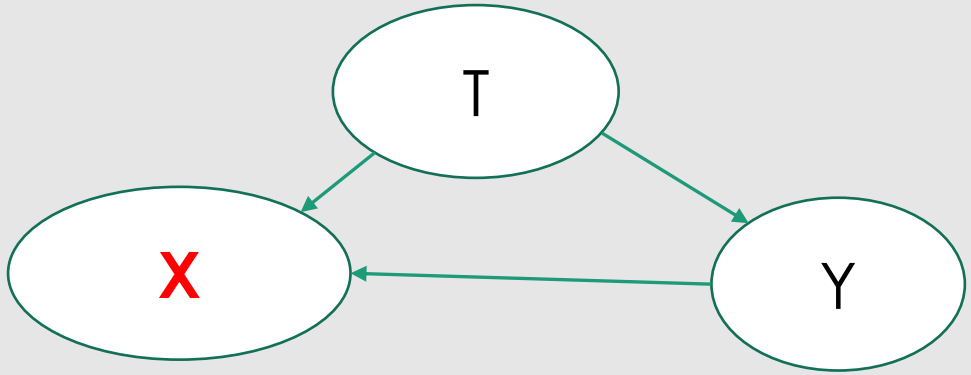
1. Confounder



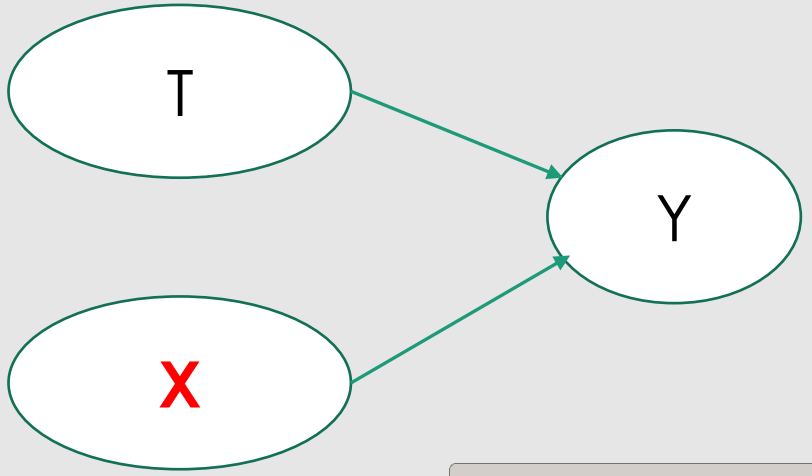
2. Mediator/Intermediate



3. Collider



4. Independent Predictor



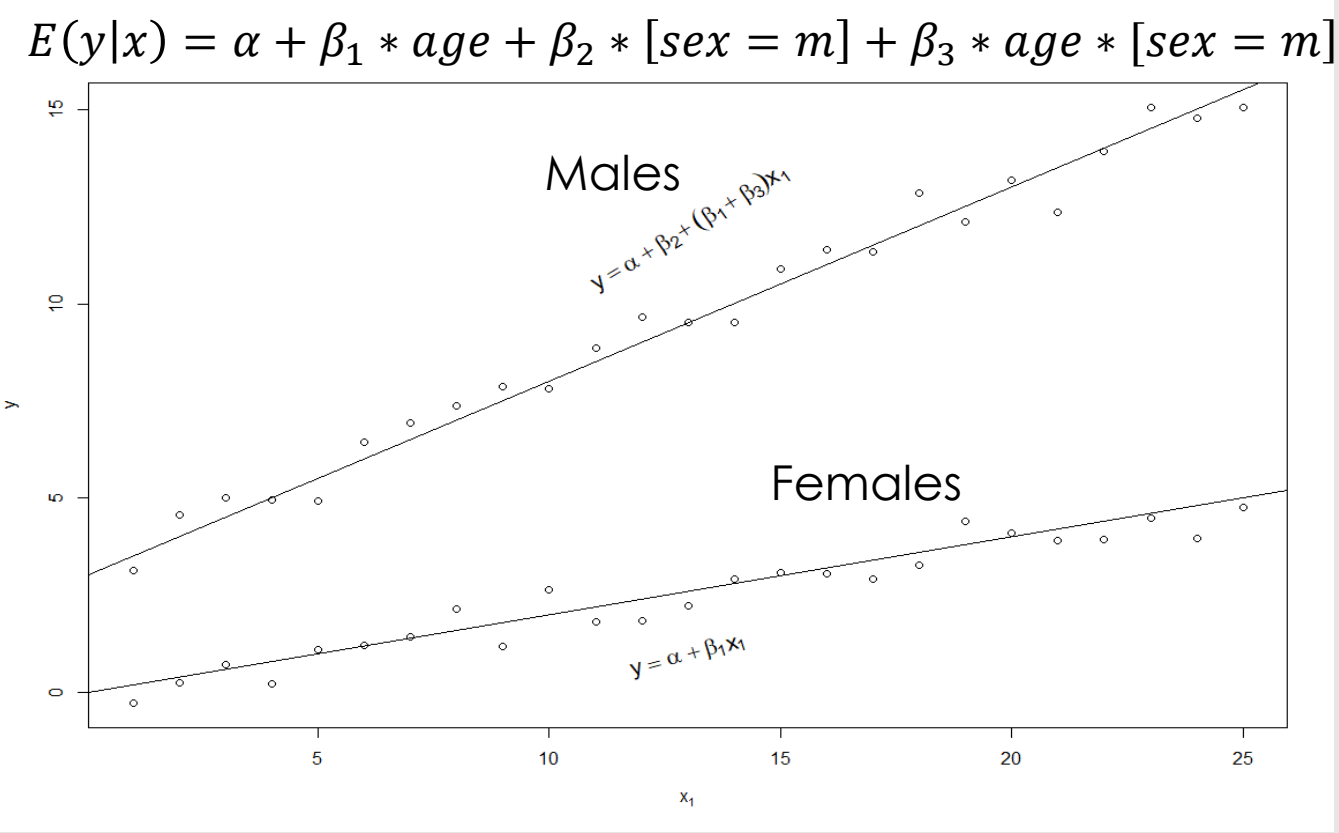
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

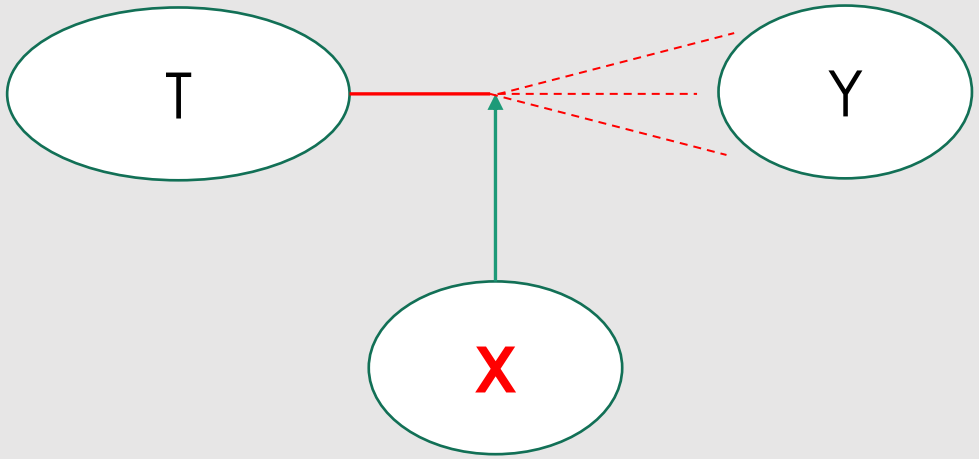
When a treatment has different effect in subgroups defined by X, there is **effect modification (moderators)** :

$$E(Y_1 - Y_0)|X = x_i) \neq E(Y_1 - Y_0|X = x_j)$$

! Could not be *represented* in a DAG !

5. Effect modifier





Building a DAG is a team-work…

Not this but….

This !

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

Basically there are **two types of effects that are of interest** in causal inference:

Average Causal Effect of a treatment or intervention on an outcome *across the entire population (or on the treated)* **ATE/ATT**

MARGINAL EFFECT

- Standardization (g-formula)
- Matching
- IPTW (Inverse Probability of Treatment Weigths)

Causal effect of a treatment **within specific subgroups** or conditions defined by certain covariates (**CATE**)

CONDITIONAL EFFECT

- Stratification
- Regression coefficients

(Linear) Regression model for the outcome (no interaction)

Suppose a **true** model has a treatment T and a confounder x for outcome y :

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + e_i$$

β_1 : conditional **and** marginal treatment effect (ATE = CATE)

In fact, if we want here to calculate the marginal effect we should **average (standardize)** over values of X:

$$\widehat{ATE} = \beta_0 + \beta_1 + \beta_2 E(X) - \beta_0 - \beta_2 E(X) = \beta_1$$

The conditional effect is equal to the marginal effect when there is **no interaction** between the treatment variable **and any other** covariate included in the model.

(Linear) Regression model if we *ignore* a confounder

Suppose a **true** model has a treatment T and a confounder x for outcome y :

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + e_i$$

If x *is related* to the treatment, we could write:

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

If we *ignore* the confounder x, we would fit the model:

$y_i = \beta_0^* + \beta_1^* T_i + e_i^*$ $\quad$ β_1^* : estimated effect if x is omitted

If we *come back* to the true model:

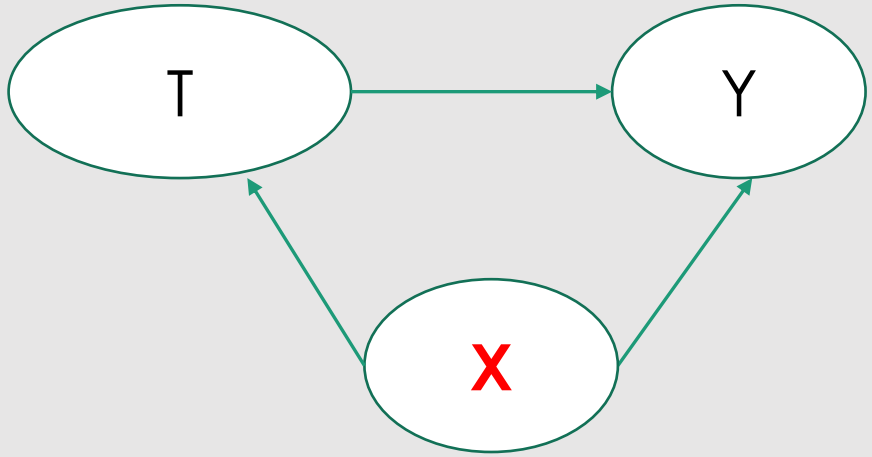
$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + e_i = \beta_0 + \beta_1 T_i + \beta_2(\gamma_0 + \gamma_1 T_i + \nu_i) + e_i$$

$$= \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)T_i + \beta_2\nu_i + e_i$$

$$\Rightarrow \quad \beta_1^* = \beta_1 + \beta_2\gamma_1$$

Estimation *without* x is correct only if $\beta_2\gamma_1$=0

- **β_2=0 x not associated with y**
- **γ_1=0 x not associated with T**



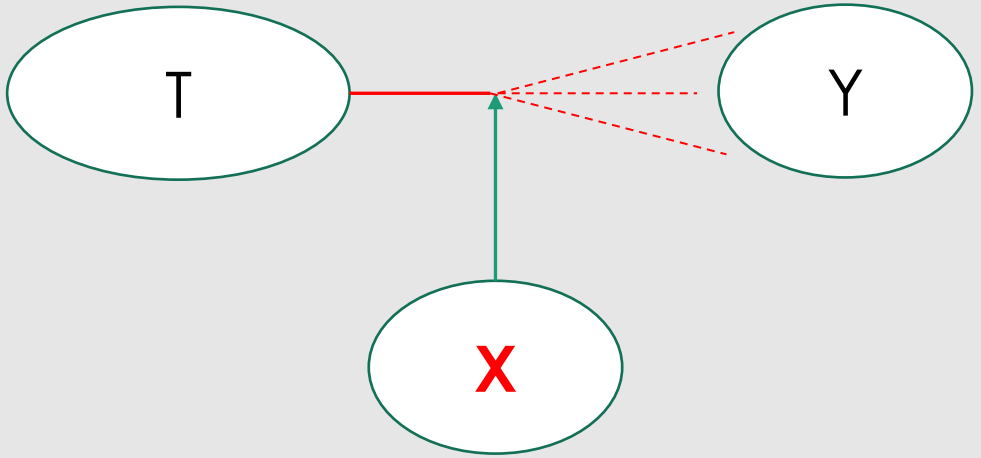
(Linear) Regression model for the outcome (with interaction)

Suppose a **true** model has a binary treatment T and a covariate (continuous) X that has an interaction with the treatment (is an **effect modifier**):

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \beta_3 x_i * T_i + e_i$$

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}[\beta_0 + \beta_1 + \beta_2 x_i + \beta_3 x_i - \beta_0 - \beta_2 x_i]$$

$$= \frac{1}{N}\sum_{i=1}^{N}[\beta_1 + \beta_3 x_i] = \beta_1 + \beta_3 E(X)$$



Here the conditional effect **is not equal to** the marginal effect since there is an **interaction** between the treatment variable and a covariate included in the model.

* In **logistic regression** $\exp(\beta)$ could be only **CATE** (with/without interactions)

General basic rules

1. **Confounder**: include (adjust/stratify)

2. **Mediator**: exclude

(for *total effect* estimation, then there are more advanced topic: disentagle direct and indirect effects…)

3. **Collider**: **to be discussed with experts: in general *exclude*** (but depend on the path…)

4. **Independent predictor**: include/exclude (in relation to sample size/precision)

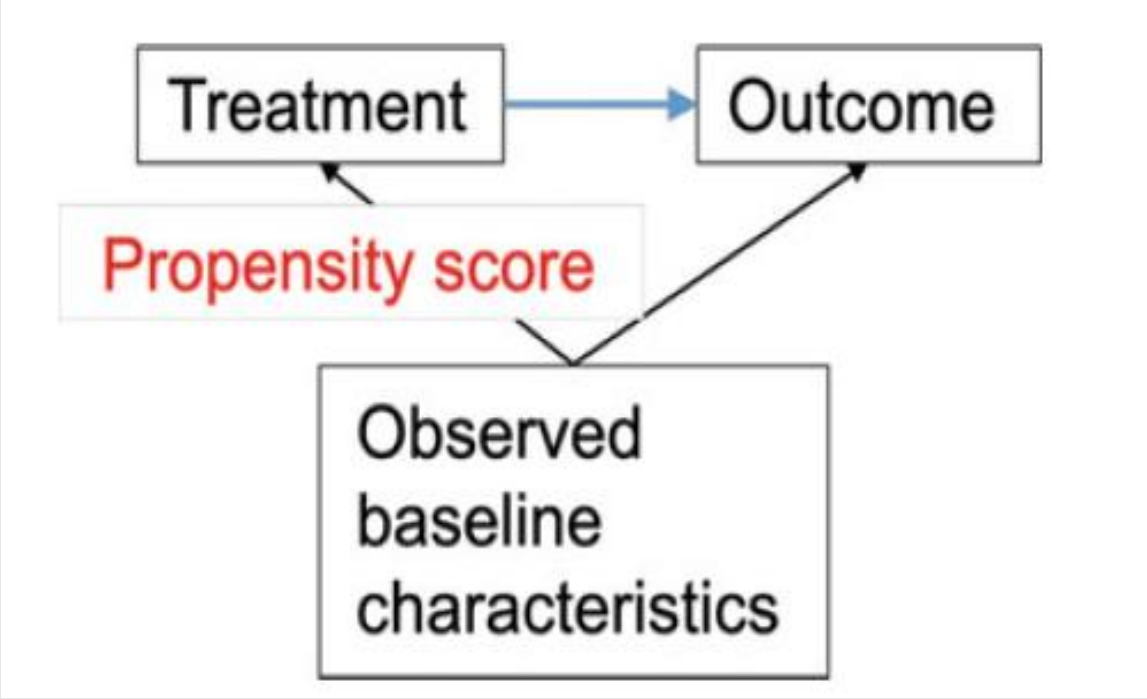
5. **Effect modifier**: include (interaction)

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

Regression models are used to **estimate treatment effects** ***trying to adjust*** **for (measured) confounders** between treatment groups.

This approach relies on the underlying assumption that the ***specified*** model is correct. Another issue is that there is *no warning* if there is no **overlap** (**positivity violation**) between the treated and controls.

In other words, models could be estimated over regions with no or little data.

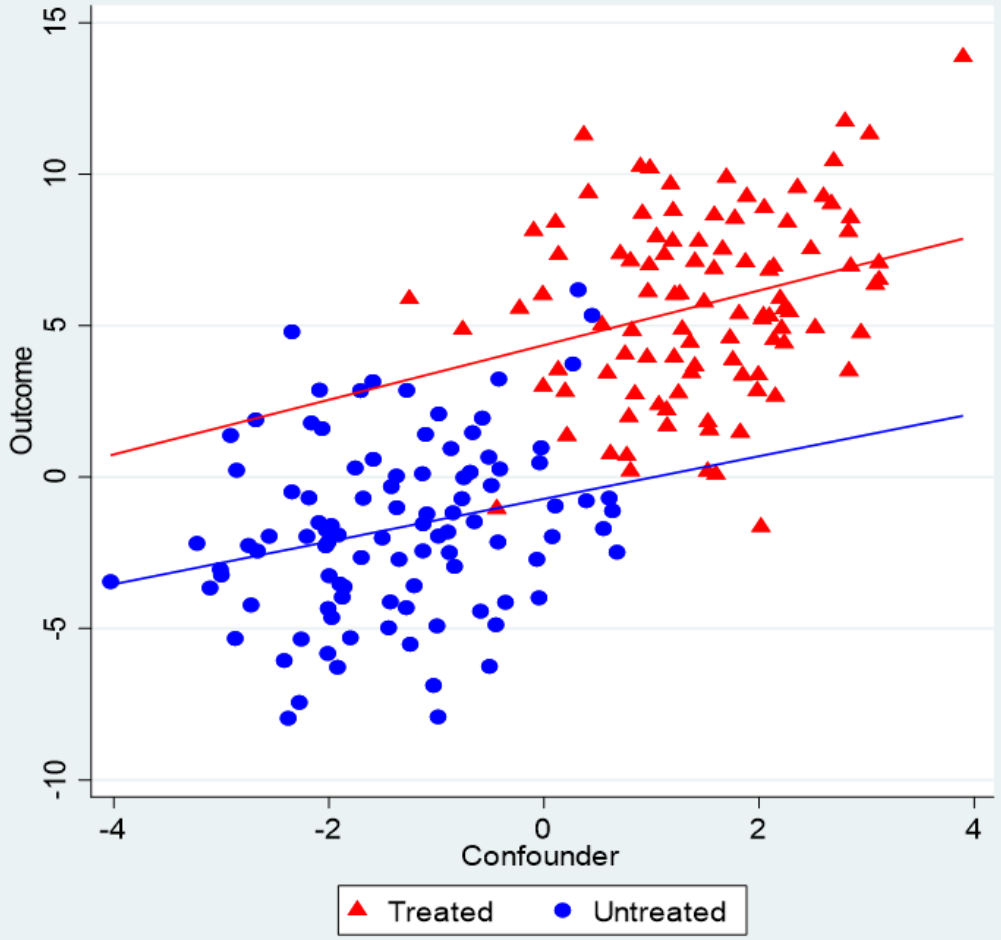


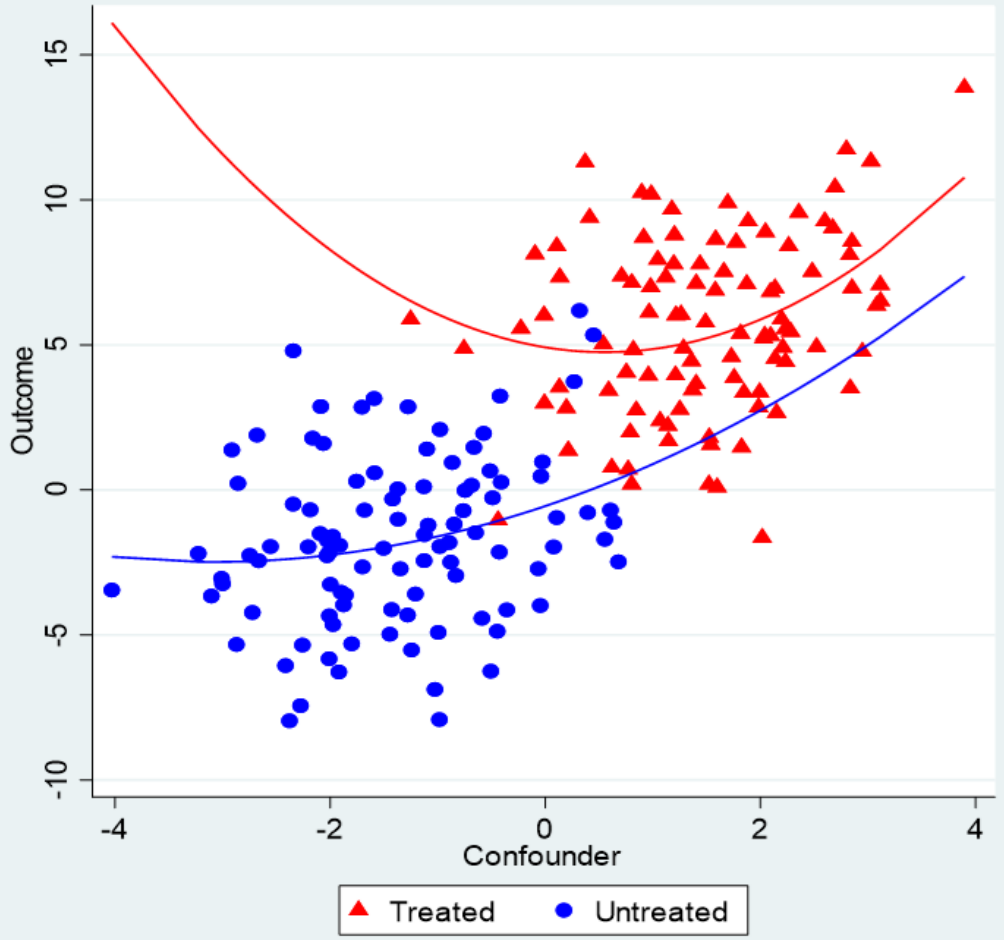
Adjustment with regression analysis is moreover **problematic** when the outcome is relatively **rare** or we have a **high-dimensional** confounders set.

An alternative is to use the **propensity score** which could be especially attractive in the setting of rare outcomes.

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

No overlap on the confounder range : dangerous extrapolation !





Propensity Score

The Propensity Score (**PS**) for each subject *i* (Rosembaum & Rubin, 1983) is defined as follows:

$$e_i = \Pr(T_i = 1 | \boldsymbol{X_i})$$

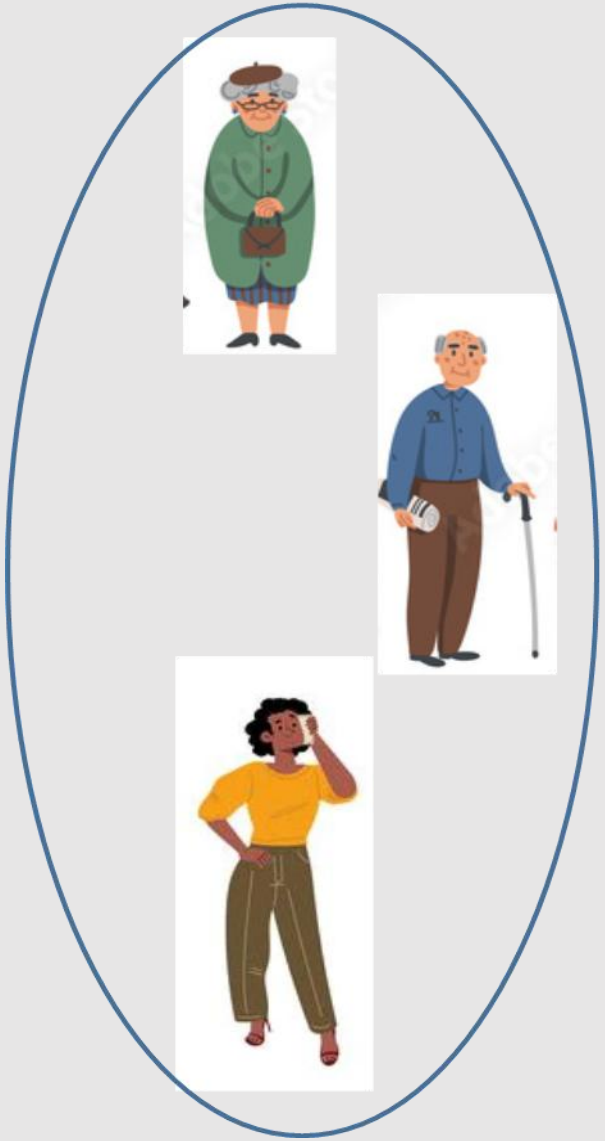
The **PS** is a **balancing score**, $e = b(X)$ such that:

$$X \perp T | \boldsymbol{b(X)}$$

Treated (T=1) and control (T=0) subjects with the same propensity score $e(x)$ have the same distribution of the observed covariates X

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

T=0 | **PS=p**



Conditional Exchangeability

On average, within groups of individual **with the same propensity score**, we expect the subjects to be **comparable** between treatment groups

$$Y_t \perp T|X \Rightarrow Y_t \perp T|PS$$

T=1 | **PS=p**

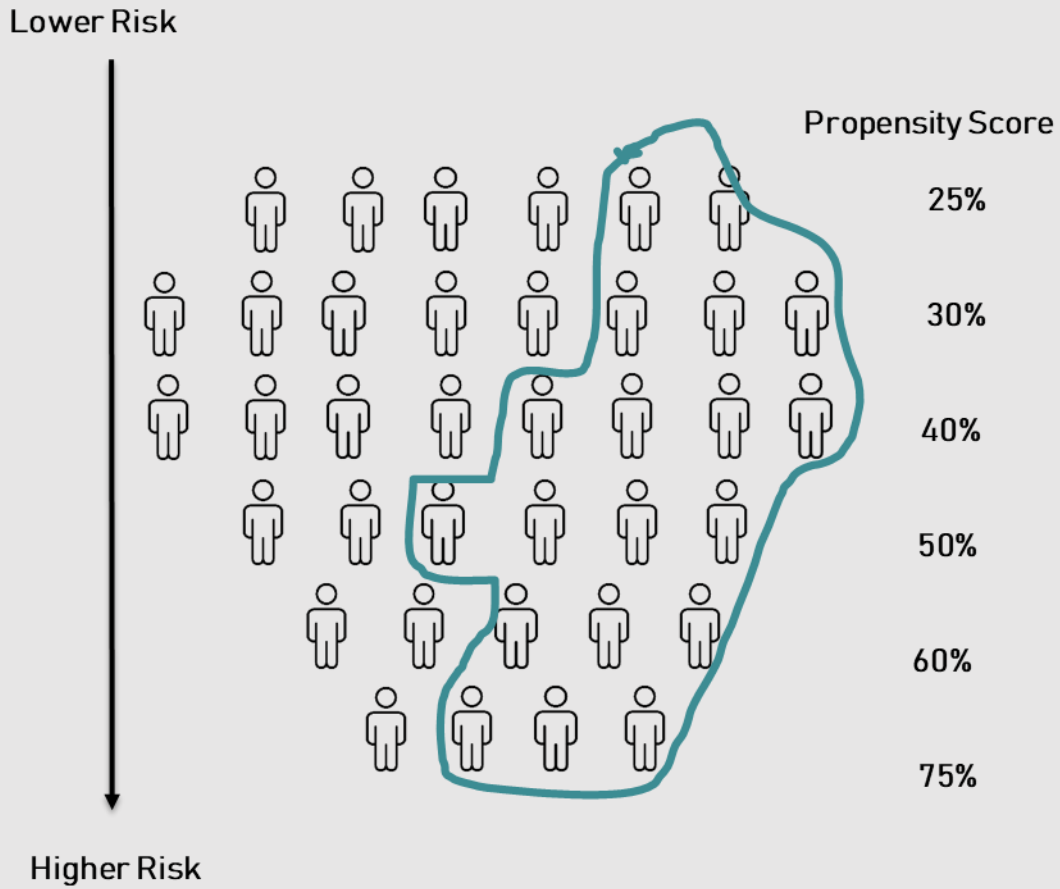
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

We **can estimate the propensity score** from the data:

e.g. using logistic regression:

$$\log\left(\frac{e_i}{1-e_i}\right) = X_i\alpha$$

$$e_i = \frac{\exp(X_i\alpha)}{1+\exp(X_i\alpha)}$$

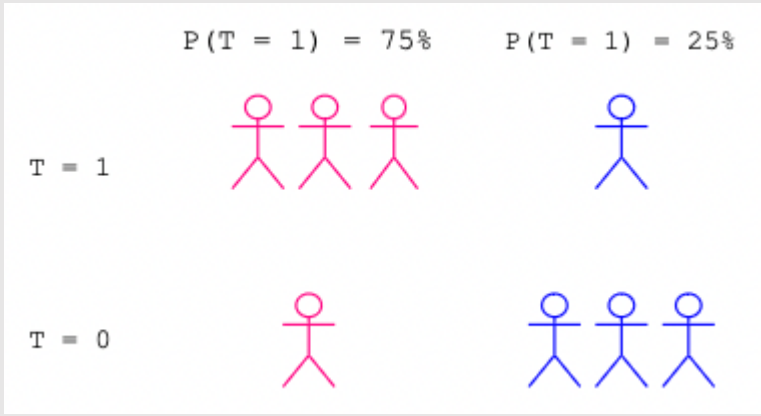


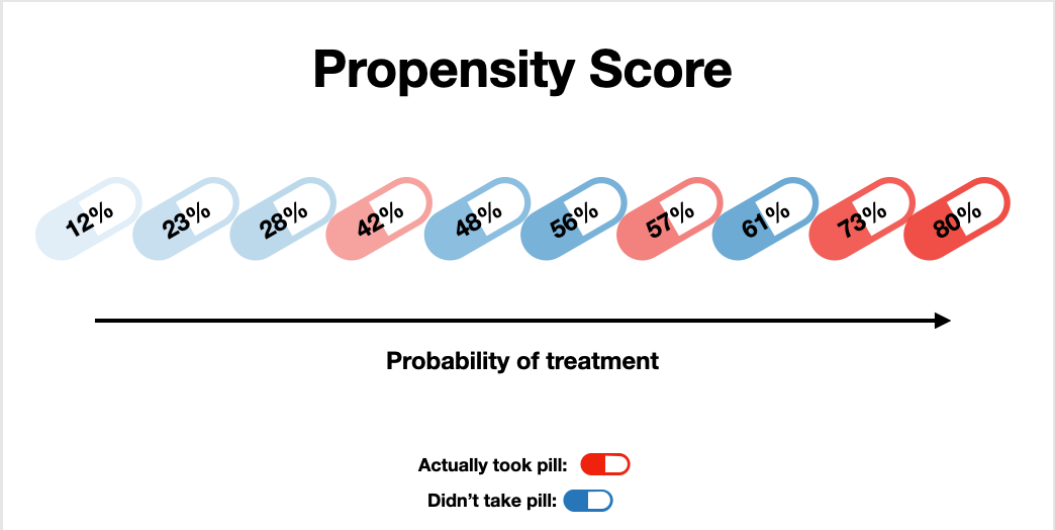
…Or using **machine learning** algorithms !
The important thing here is just to have a predicted probability to receive the treatment

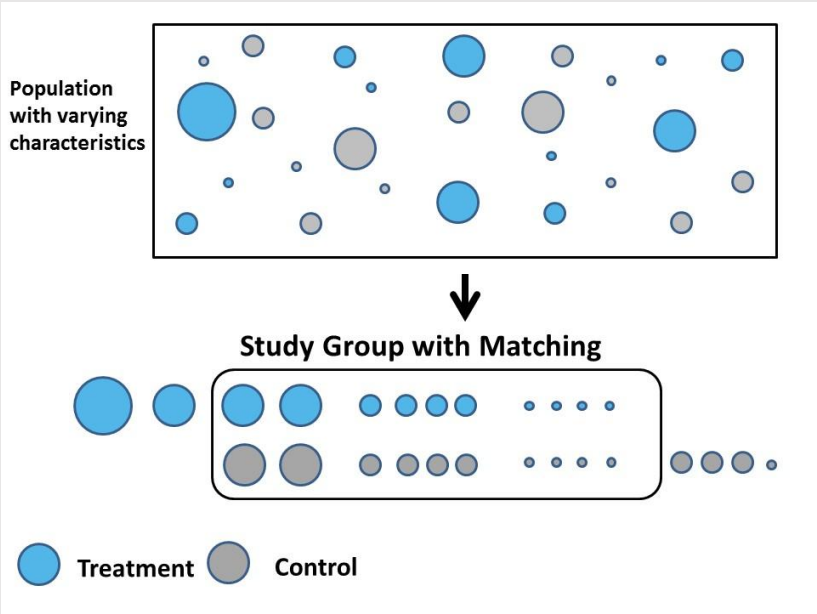
Construction of the *control* group:

Four *main* approaches:

1. Covariate Adjustment
2. Matching
3. Stratification
4. **I**nverse **P**robability of **T**reatment **W**eighting (IPTW)







Covariate Adjustment: use PS as a covariate !

Method

$$Y_i = T_i\beta + f(\widehat{e_i})$$

$\hat{\beta}$ is the *average*/conditional* treatment effect (*linear regression)

✓ Simple, efficient …

✗ Assumption of *some* functional relationship between the outcome and the PS

Diagnostic Tools

1. The distribution of the PS in the two groups **can help verifying the positivity assumption**

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

Matching

Method

Matched sets of **treated** with **untreated** subjects **who share a similar value of PS**

In the matched sample we can compare the outcome using **statistical methods** for **paired experiments**

Only the ATT effect could be estimated

✓ **Popular**

✓ **Intuitive**

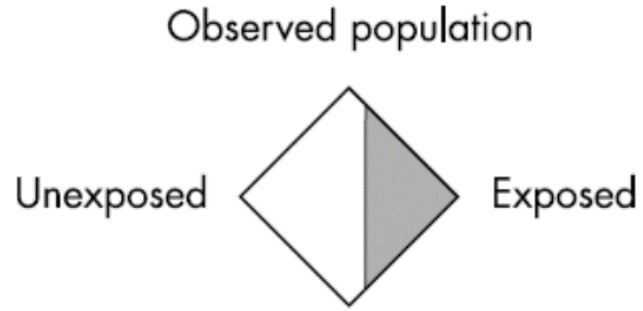
✗ **Many variants for the matching procedure**

✗ **It has been argued that appropriate SE are hard to obtain**

✗ **It discards a lot of data in a non-deterministic manner**

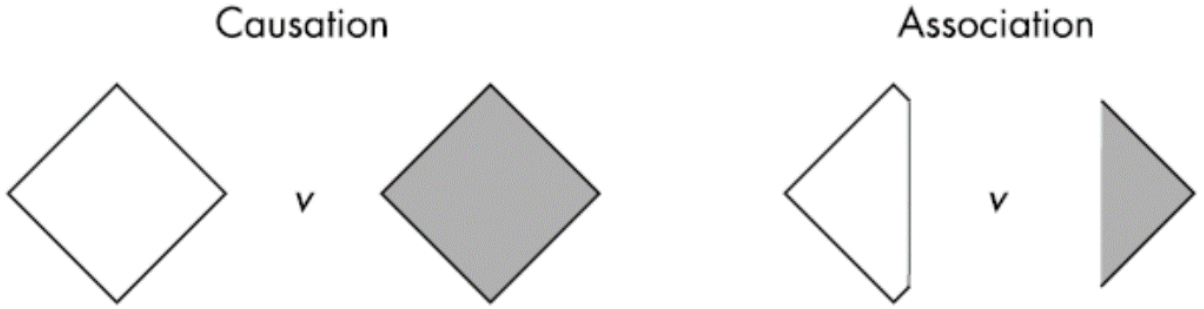
Diagnostic Tools

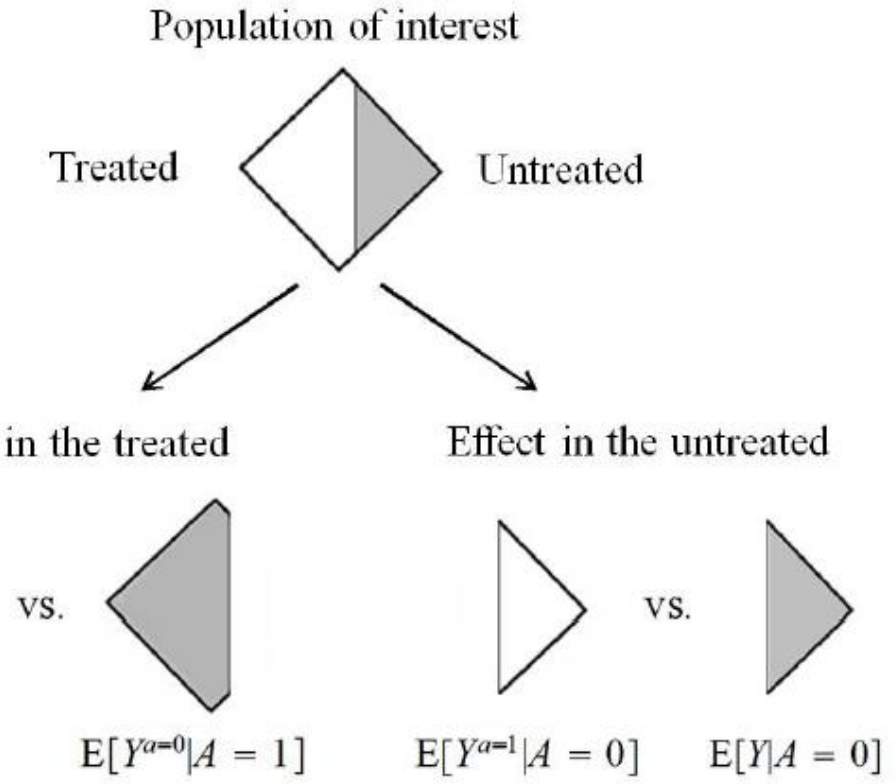
1. The distribution of the PS in the two groups **can help verifying the positivity assumption**
2. **Characteristics can be compared in the matched sample to assess achievement of balance**



$$ITE_i = Y_1(i) - Y_0(i)$$ **For all subjects**

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





For treated subjects

$$ITE_i^{T=1} = Y_1(i) - Y_0(i)$$

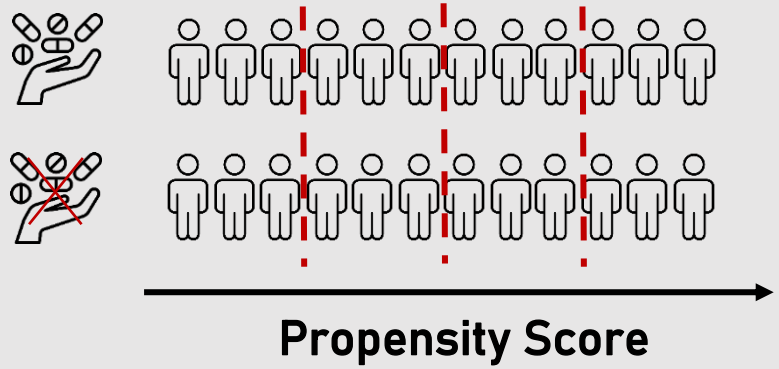
$$ATT = E(Y_1 - Y_0|T = 1)$$

Stratification

Method

We obtain exclusive subsets according to values of the propensity scores

In **each stratum we can** compare the outcome between the two treatment groups (**CATE**) and then eventually use a **weighted mean** to obtain the marginal treatment effect (**ATE**)



✓ **Simple**

✗ **It can leave residual confounding**

✗ **We need to be careful to have a sufficient sample size per strata/outcomes**

Diagnostic Tools

1. The distribution of the PS in the two groups **can help verifying the positivity assumption**
2. **Pre-treatment characteristics can be compared within strata**

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

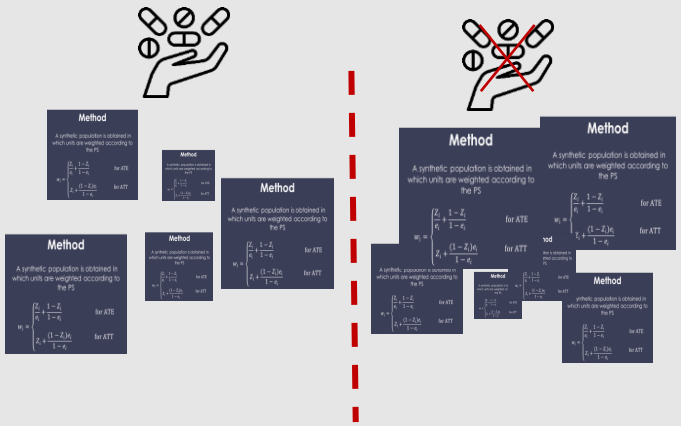
IPTW

Method

A **synthetic** population is obtained in which units are weighted according to the PS

$$w_i = \begin{cases} \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i} & \text{for ATE} \\ T_i + \frac{(1 - T_i)e_i}{1 - e_i} & \text{for ATT} \end{cases}$$

✗ **Extremes weights → imprecise estimates**



✓ **Mathematically appealing**

✓ **Extends to complex scenarios**

Diagnostic Tools

1. The distribution of the PS in the two groups **can help verifying the positivity assumption**
2. **Pre-treatment characteristics can be compared in the weighted dataset**

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