Introduction to Causal Machine Learning

For healthcare and life sciences

Javier González

September 20, 2020

Microsoft Research Cambridge

"I checked it very thoroughly, said the computer, and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Douglas Adams, The Hitchhiker's Guide to the Galaxy (1979)

Introduction

Causality and the empiricists

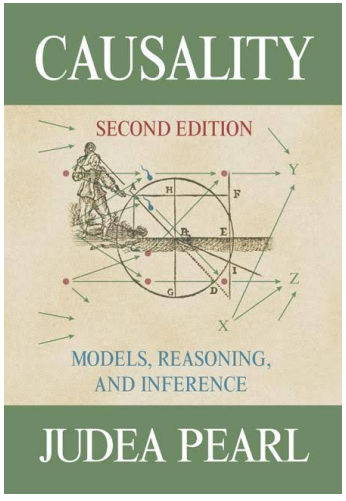
Hume

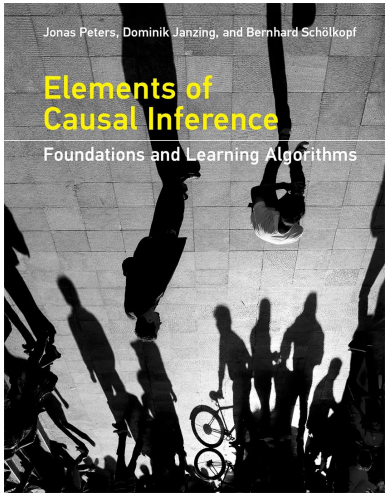
Kant

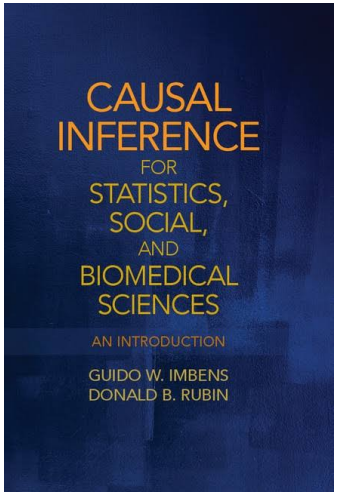
Hume (1711-1776): *"Causality is a rooted belief. We see cause and effect in constant conjunction so we believe that A causes B".*

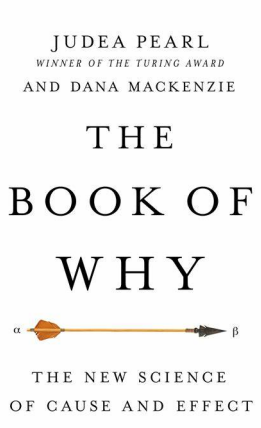
Kant (1724–1804) *"Everything that happens, that is, begins to be, presupposes something upon which it follows by rule".*

Causality in modern science









Goals of this tutorial

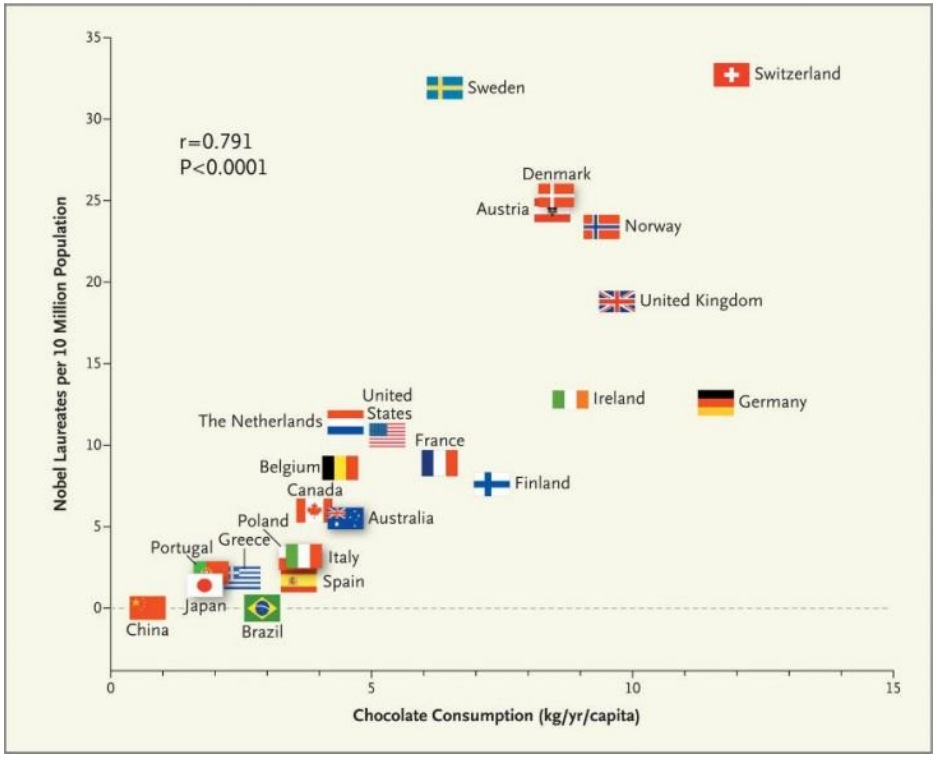
1. Introduce Pearl's framework of causality.
2. Introduce Rubin's Potential outcomes (PO) framework of causality.
3. Use both frameworks to understand identification, RCT, confounding, propensity scores and instrumental variables.
4. Study some examples in healthcare and biosciences.

Game: questions about questions

- Decide whether the following statements are **True** or **False**.
- Write down ANY information you are using answer the question.
- If you are not sure of the answer describe what you would do to be more certain.

Work in groups and discuss!

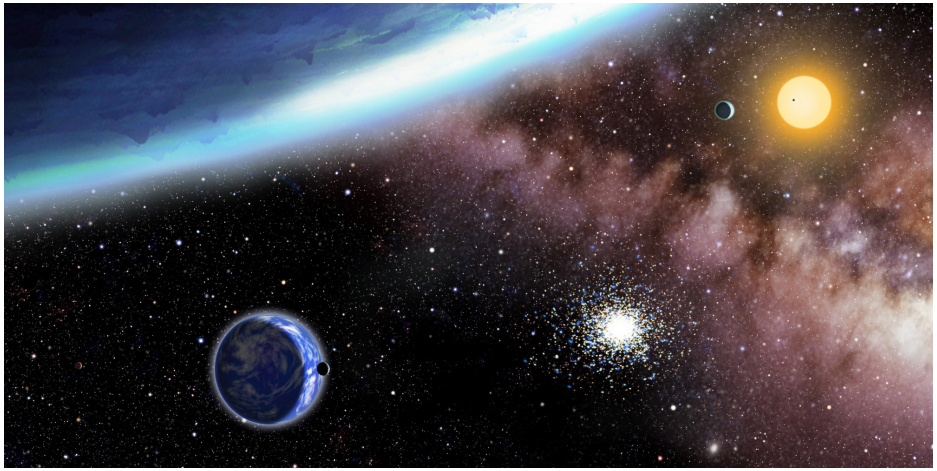
Statement 1



I can build a model able to predict the number Nobel Laureates per capita using chocolate consumption.

True or **False**?

Statement 2



$$GravityForce = \frac{Mass_1 \times Mass_2}{distance^2}$$

The mass of two planets has a direct effect in the how they are attracted to each other.

True or **False**?

Statement 3

MENU nature
International journal of science

Scientific Correspondence | Published: 13 May 1999

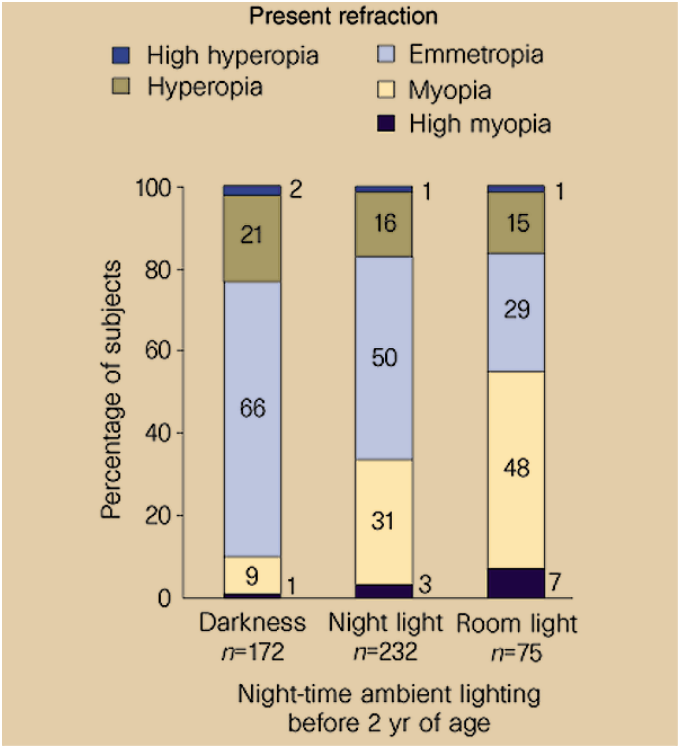
Myopia and ambient lighting at night

Graham E. Quinn, Chai H. Shin, Maureen G. Maguire & Richard A. Stone

Nature **399**, 113–114 (13 May 1999) | Download Citation

Abstract

Myopia, or short-sightedness, occurs when the image of distant objects, focused by the cornea and lens, falls in front of the retina. It commonly arises from excessive postnatal eye growth, particularly in the vitreous cavity. Its prevalence is increasing and now reaches 70–90% in some Asian populations[1,2]. As well as requiring optical correction, myopia is a leading risk factor for acquired blindness in adults because it predisposes individuals to retinal detachment, retinal degeneration and glaucoma. It typically develops in the early school years but can manifest into early adulthood[2]. Its aetiology is poorly understood but may involve genetic and environmental factors[1,2], such as viewing close objects, although how this stimulates eye growth is not known[3]. We have looked at the effects of light exposure on vision, and find a strong association between myopia and night-time ambient light exposure during sleep in children before they reach two years of age.



Sleeping with the light on has a correlation with having myopia.

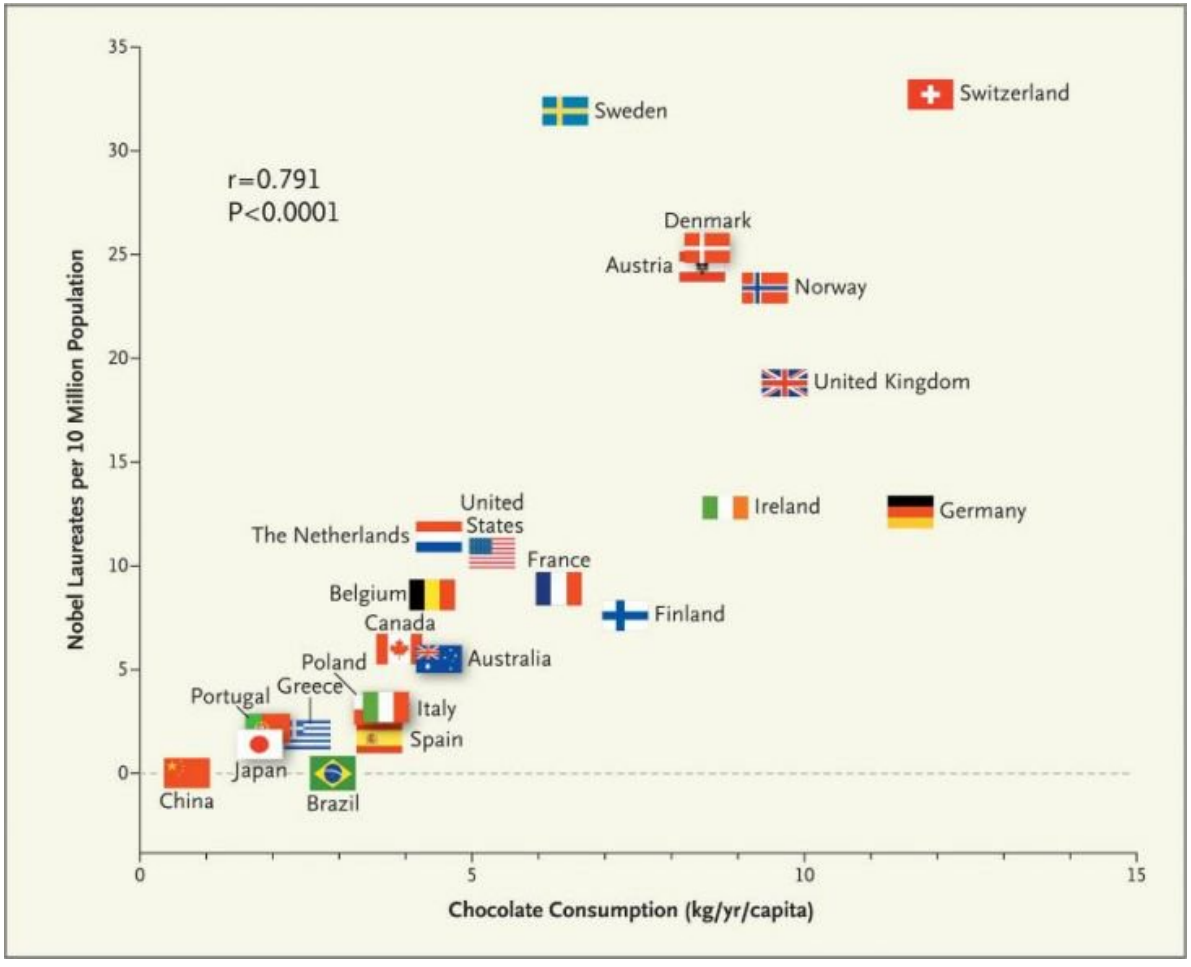
True or **False**?

Statement 4

Last summer I went for holidays for two weeks. If my friend Pablo hadn't watered my plants, today they would be all dead.

True or **False**?

Statement 5



Eating chocolate makes you smarter.

True or **False**?

Statement 6

Had Hilary Clinton been a man, she would have won the general election to Donald Trump.

True or **False**?

Types of questions

Statement	Question type
1. Predict Nobel with chocolate.	Association
2. Planets and gravity	Causation
3. Light at night and myopia.	Association
4. Pablo watering my plants.	Counterfactual
5. Chocolate makes you smarter.	Causation
6. Hilary and Trump	Counterfactual

Association questions:

Can be answered with $\mathbb{P}(X, Y)$

Causation and counterfactuals questions:

Can be answered with $\mathbb{P}(X, Y)$ **AND/OR** a causal mechanism.

Machine Learning

Classic Machine Learning:

$$Data + Model \xrightarrow{\text{Compute}} Prediction \xrightarrow{\text{Reward}} Decision$$

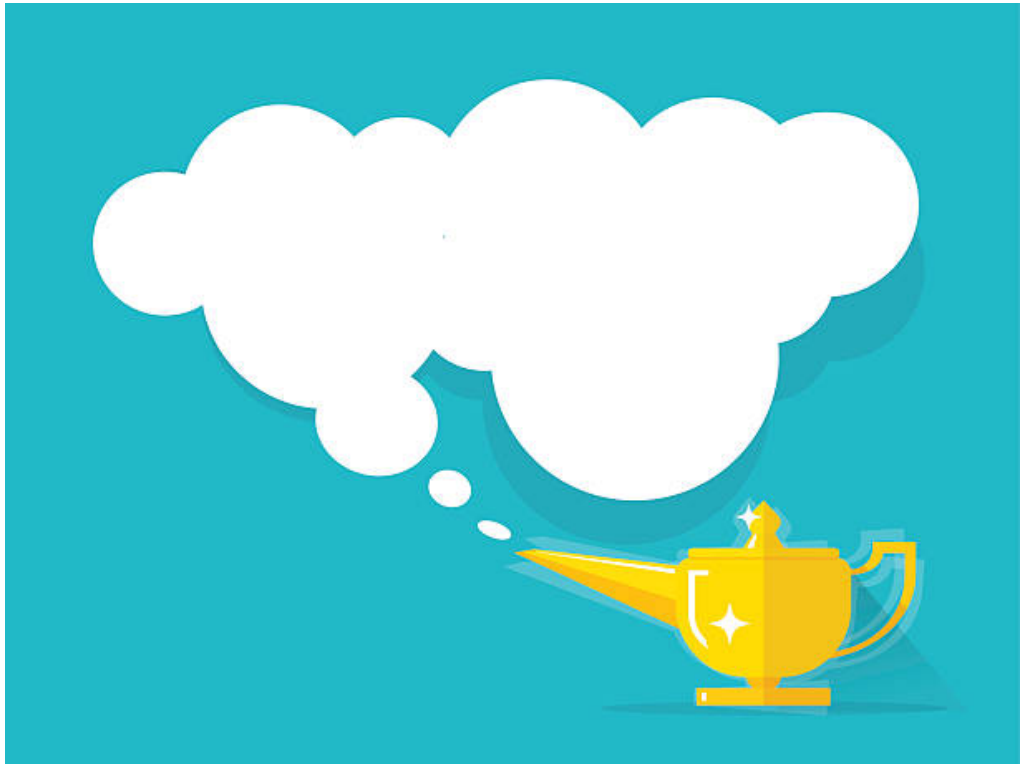
Causal Machine Learning:

$$Data + Model + Causal\ Mechanism \xrightarrow{\text{Compute}} Prediction \xrightarrow{\text{Reward}} Decision$$

Classic Machine Learning

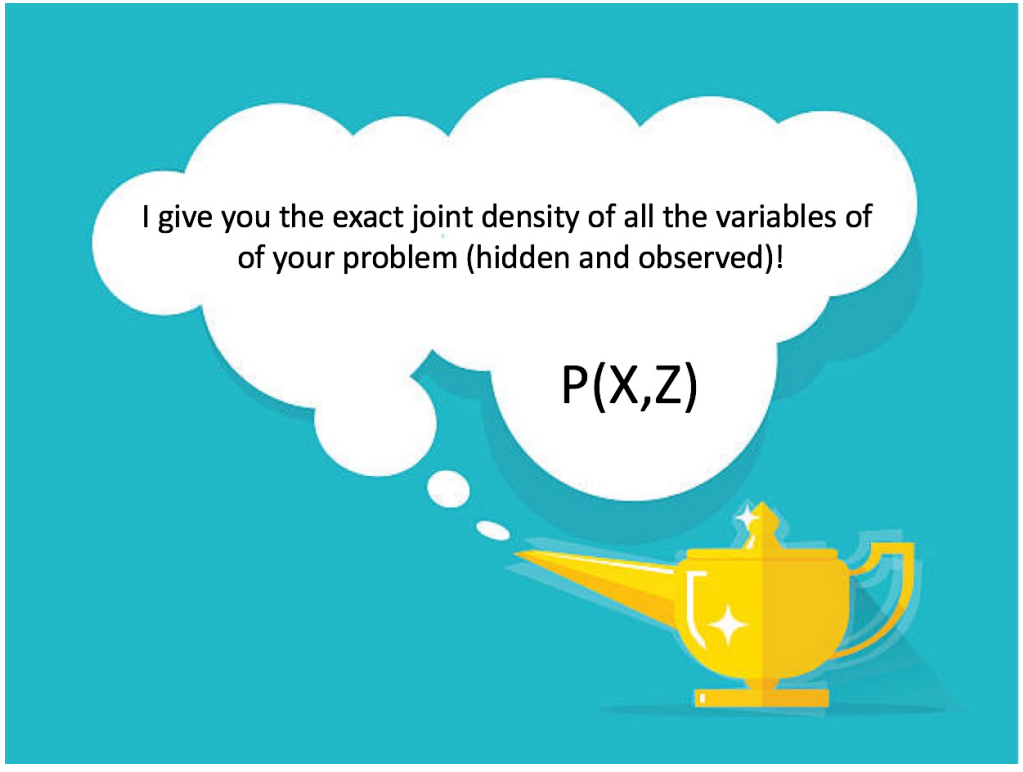
It is your lucky day!

You want to 'solve' statistics/machine learning (before taking a causality tutorial, of course!) and you can ask the Aladin lamp for one desire and one desire only. What would you ask?



It is your lucky day!

You want to 'solve' machine learning and you can ask the Aladin lamp for one desire (and one desire only), what would you ask?



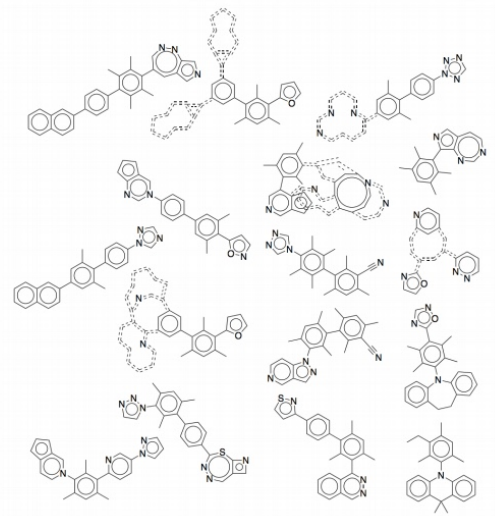
What can I do If I know $\mathbb{P}(X, Y)$?

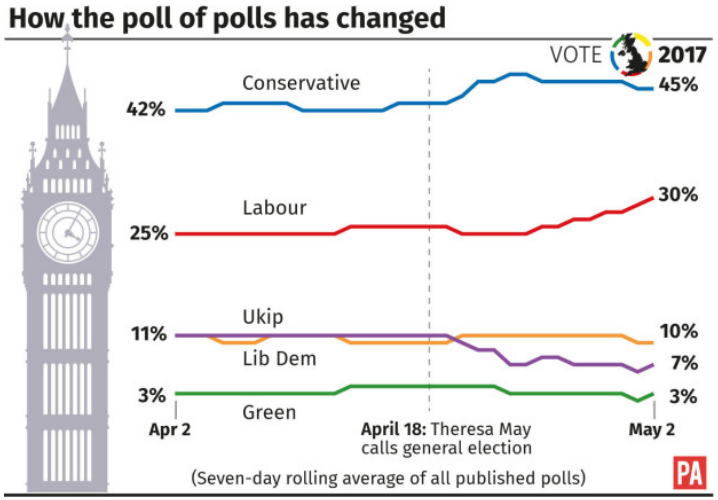
I can do a lot:

1. Compute marginals: $\mathbb{P}(X) = \int \mathbb{P}(X, Y) dY$.
2. Compute conditionals: $\mathbb{P}(Y|X) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(X)}$.
3. Make predictions: $\mathbb{P}(y^*|X, Y, x^*)$.

A person riding a motorcycle on a dirt road.

A group of young people playing a game of frisbee.



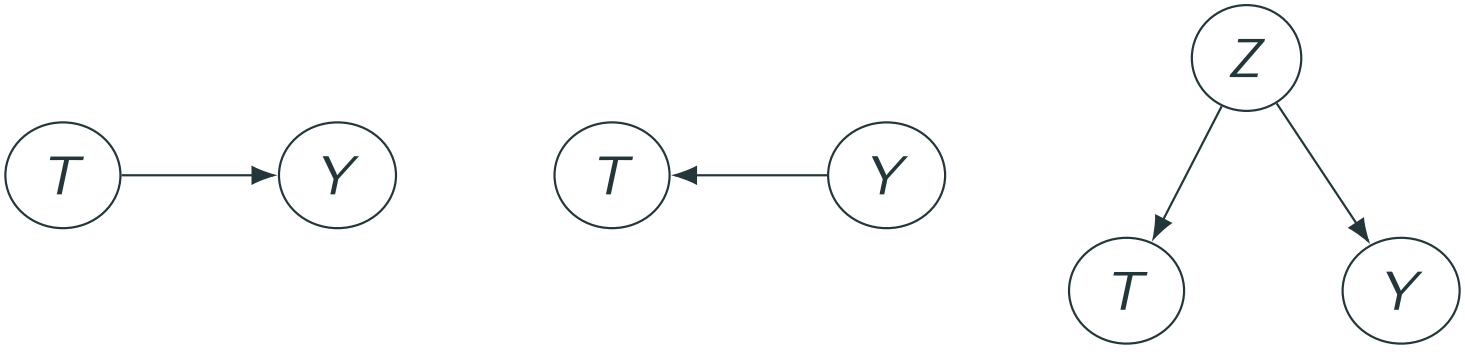


Link images and text, predict the effects of drugs, marginalize the vote intention, etc.

Correlation is not causation

What can't I do If I only know $\mathbb{P}(X, Y)$?

$\mathbb{P}(X, Y)$ helps to describe the world but not to understand how it works.



Reichenbach's Common Cause Principle

'If two events are correlated, then either there is a causal connection between them or there is a third event, a so called common cause, which brings about the correlation.'

Structural Causal models

Structural equation models

Question

How do we express mathematically that the dose of treatment (T) affects the recovery speed (Y)?

$$y = \beta t + \epsilon_Y$$

- t stands for the dose of the treatment.
- y stands for recovery speed.
- ϵ_Y stands for all factors that could affect Y when T is held constant.

Modelling the directionality of the causal effect

$y = \beta t + \epsilon_Y$ does not properly express the causal relationship.

- Algebraic equations are symmetric.
- $t = (y - \epsilon_y)/\beta$, the recovery speed causes the dose (?!).
- We need a diagram to disambiguate the situation.

Modelling the directionality of the causal effect

$y = \beta t + \epsilon_Y$ does not properly express the causal relationship.

- Algebraic equations are symmetric.
- $t = (y - \epsilon_y)/\beta$, the recovery speed causes the dose (?!).
- We need a diagram to disambiguate the situation.

Structural Causal model: $\mathbb{P}(T, Y)$ + causal graph G.

Full model for treatment and disease

$$t = \epsilon_t$$
$$y = \beta t + \epsilon_y$$

ϵ_t and ϵ_y are 'exogenous variables' (unobserved background factors).

Principle of independent mechanisms

[Haavelmo 1943, Frisch 1948]: *'A structural relation not only explains the observed data, it captures a structure connecting the variables; related to autonomy and invariance.'*

Elements of causal inference book: *'The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each others. In the probabilistic case this means that the conditional distribution of each variable given its causes (mechanism) does not inform or influence the other conditional distributions. In the case of two variables this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.'*

Intervening on the mechanism f_i does not affect any other f_j

From linear to non-parametric models

Let's have H, that represents the type of hospital.

Causal graph:

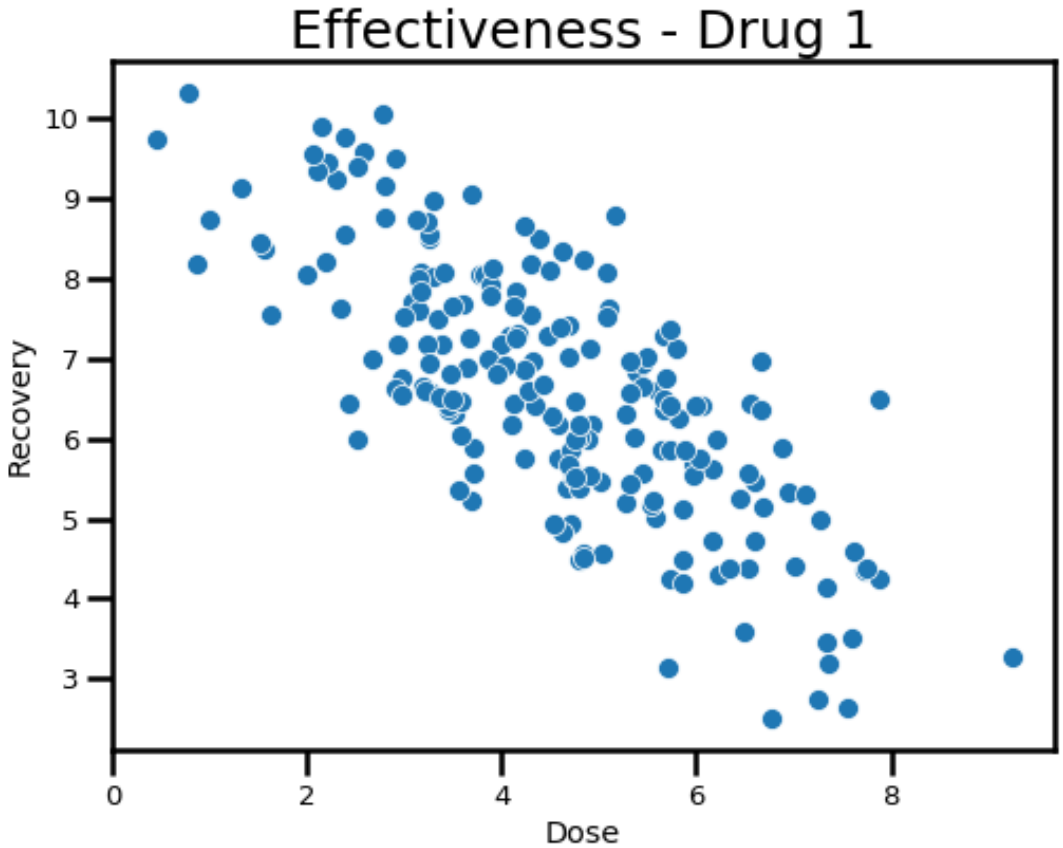
Structural equation mode that induces $\mathbb{P}(H, T, Y)$:

$$\begin{aligned} h &= \epsilon_h \\ t &= f_t(h, \epsilon_t) \\ y &= f_y(t, \epsilon_y) \end{aligned}$$

f_t and f_y represent independent mechanisms.
ϵ_h, ϵ_t and ϵ_y are jointly independent.

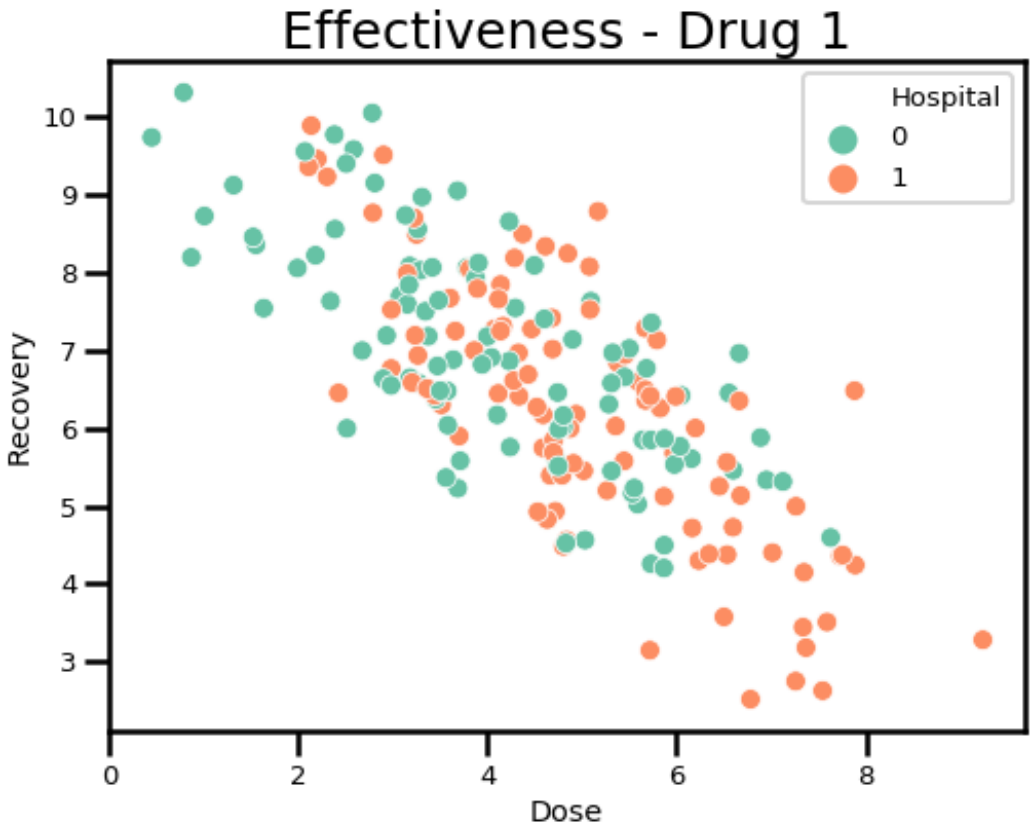
Example with data

Effectiveness of a drug in the recovery time of patients.



Example with data

Effectiveness of a drug in the recovery time of patients.



Observing vs. doing (intervening)

Patients sampled from hospitals, treatment dose and recovery time.

Observational distribution: $\mathbb{P}(\textit{recovery} < 8\,\textit{days} | \textit{dose} = 2\,\textit{grs.})$

Distribution of *recovery* when we **observe** that the *dose* takes value 1gr.

'Probability that someone recovers in less that 8 days given that **we know** *that he/she had a dose of 3 grs."*

Observing vs. doing (intervening)

Patients sampled from hospitals, treatment dose and recovery time.

Observational distribution: $\mathbb{P}(recovery < 8\,days | dose = 2\,grs.)$

Distribution of *recovery* when we **observe** that the *dose* takes value 1gr.

'Probability that someone recovers in less that 8 days given that **we know** *that he/she had a dose of 3 grs."*

Interventional distribution: $\mathbb{P}(recovery < 8 days | do(dose = 3gr))$

Distribution of *recovery* when we **force** everyone to take *doses* of 3gr.

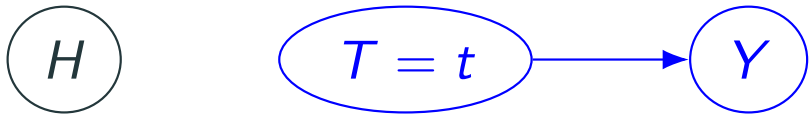
'Probability that someone recovers in less that 8 days given that **we force** *he/she to have a dose of 3 grs."*

Representing interventions

Fantasise an 'intervened universe' in which everyone gets a fixed treatment.

Observed universe

Intervened universe



$$
\begin{aligned}
H &= f_h(\epsilon_h) \\
T &= f_t(T, \epsilon_t) \\
Y &= f_y(X, \epsilon_y)
\end{aligned}
$$

$$
\begin{aligned}
H &= f_h(\epsilon_h) \\
T &= t \\
Y &= f_y(t, \epsilon_y)
\end{aligned}
$$

This is represented by the mathematical operator $do(T = t)$.

Representing interventions

Observed universe

$$\begin{aligned} H &= f_h(\epsilon_h) \\ T &= f_t(T, \epsilon_t) \\ Y &= f_y(X, \epsilon_y) \end{aligned}$$

Intervened universe

$$\begin{aligned} H &= f_h(\epsilon_h) \\ T &= t \\ Y &= f_y(t, \epsilon_y) \end{aligned}$$

Observational distribution

$$\mathbb{P}(H, T, Y)$$

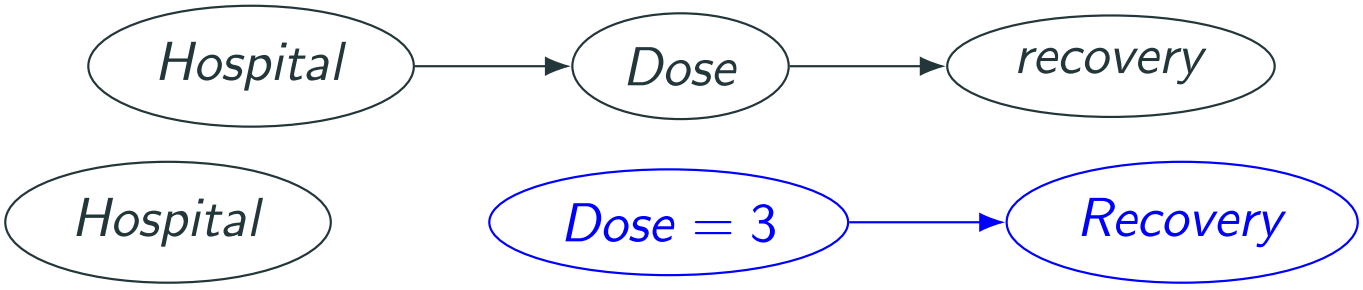
Post-intervention distribution

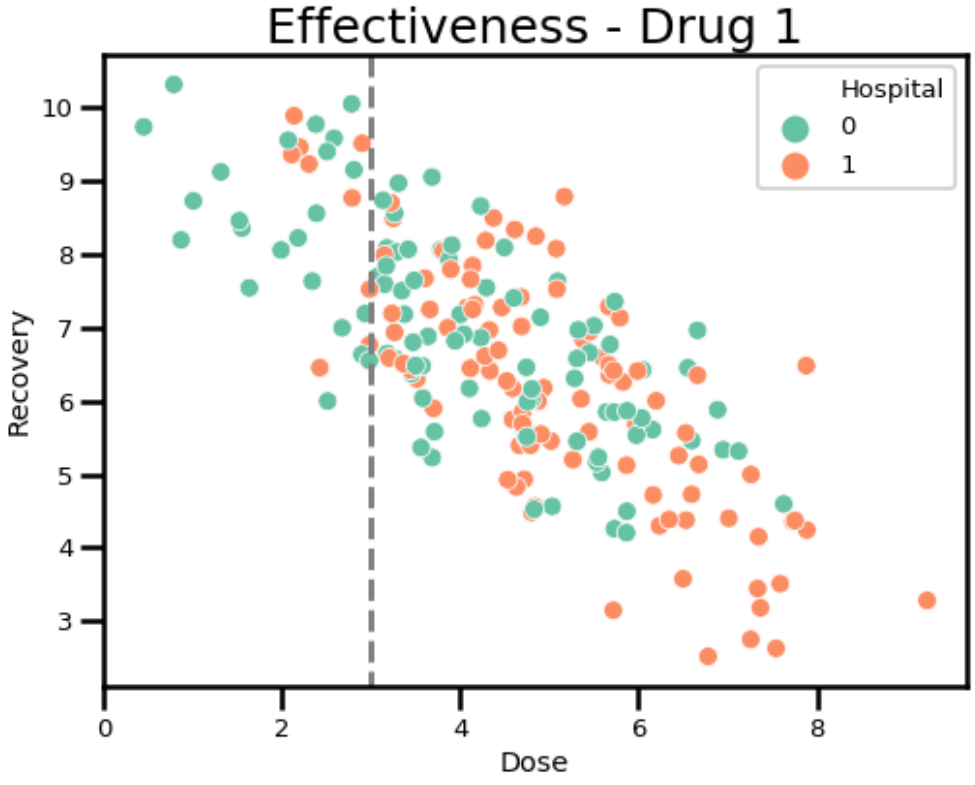
$$P^{do(T=t)}(H, Y)$$

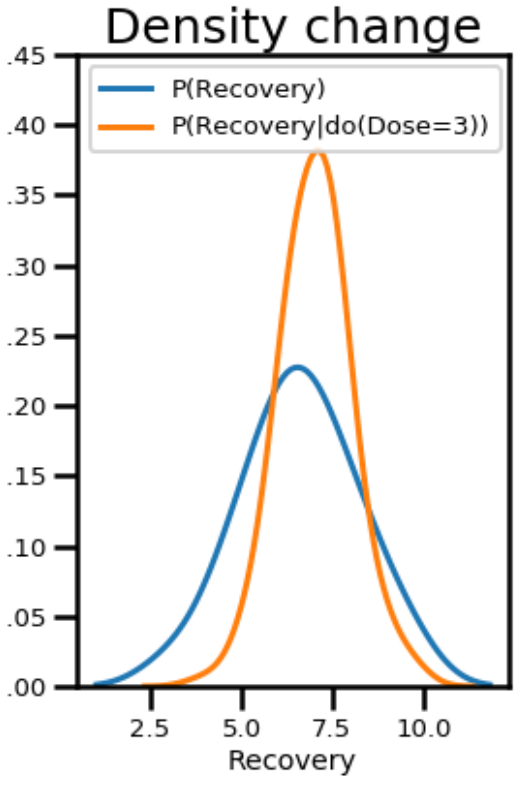
$$\boxed{\mathbb{P}(Y|do(T=t)) := P^{do(T=t)}(Y|T=t)}$$

The *do* operator is a conditional in a post-intervention 'parallel universe'.

Effect of an intervention







First formal definition of a causal effect

Given an SCM $\mathcal{S} := (G, \mathbb{P})$ the following statements are equivalent:

1. There is a causal effect from T to Y.
2. There is t' such that $\mathbb{P}(Y|do(T = t')) \neq \mathbb{P}(Y)$.

Confounding factors

Example: Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Example: Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Treatment B.

Example: Kidney stones

Success recovery rates of two treatments for kidney stones:

Treatment A	**Treatment B**
78% (273/350)	**83% (289/350)**

Which treatment is better?

Treatment B.

Ok, wait, are we sure? let's have a look to the data again....

Confounders

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**

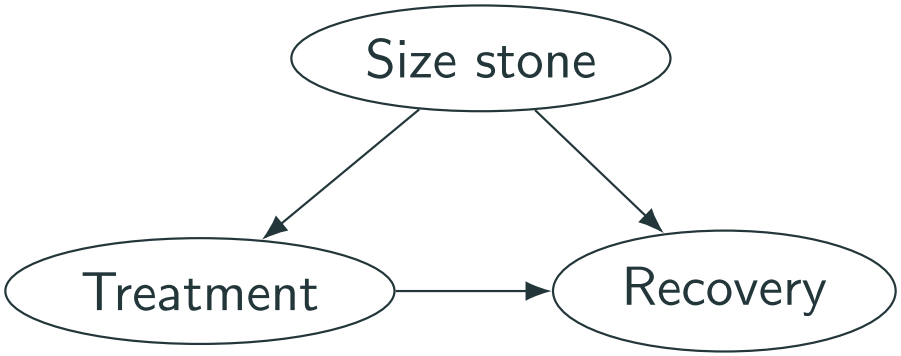
The size of the stone is what is called a confounding variable.

Confounders

When the less effective treatment (B) is applied more frequently to less severe cases, it can appear to be a more effective treatment.

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**

The size of the stone is what is called a confounding variable.



Solution

Compute the effect of each treatment by weighting the effect by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

Solution

Compute the effect of each treatment by weighting the effect by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

$$\begin{aligned}\mathbb{P}(Recover|do(T=A)) &= \mathbb{P}(small)\mathbb{P}(Recover|small,A)\\ &+ \mathbb{P}(big)\mathbb{P}(Recover|big,A)\\ &= \mathbf{0.8325}\end{aligned}$$

Solution

Compute the effect of each treatment by weighting the effect by the number of cases.

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

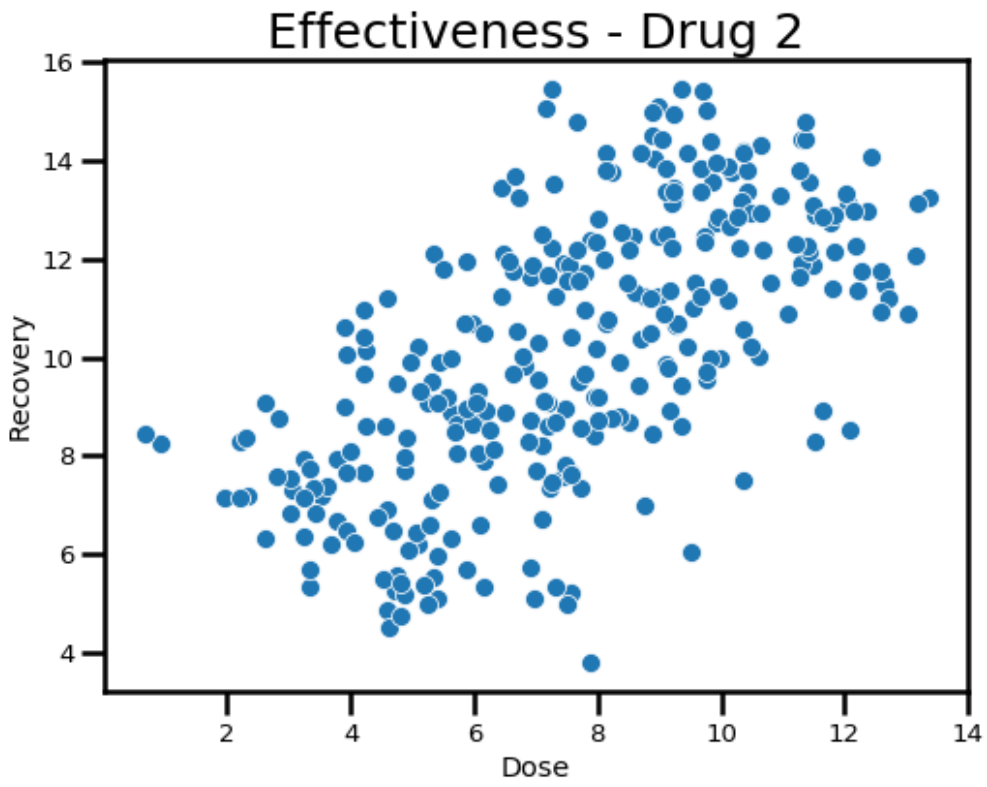
$$
\begin{aligned}
\mathbb{P}(Recover|do(T=A)) &= \mathbb{P}(small)\mathbb{P}(Recover|small,A) \\
&+ \mathbb{P}(big)\mathbb{P}(Recover|big,A) \\
&= \mathbf{0.8325} \\
\mathbb{P}(Recover|do(T=B)) &= \mathbb{P}(small)\mathbb{P}(Recover|small,B) \\
&+ \mathbb{P}(big)\mathbb{P}(Recover|big,B) \\
&= \mathbf{0.7788}
\end{aligned}
$$

Treatment A is indeed better.

Simpson's paradox

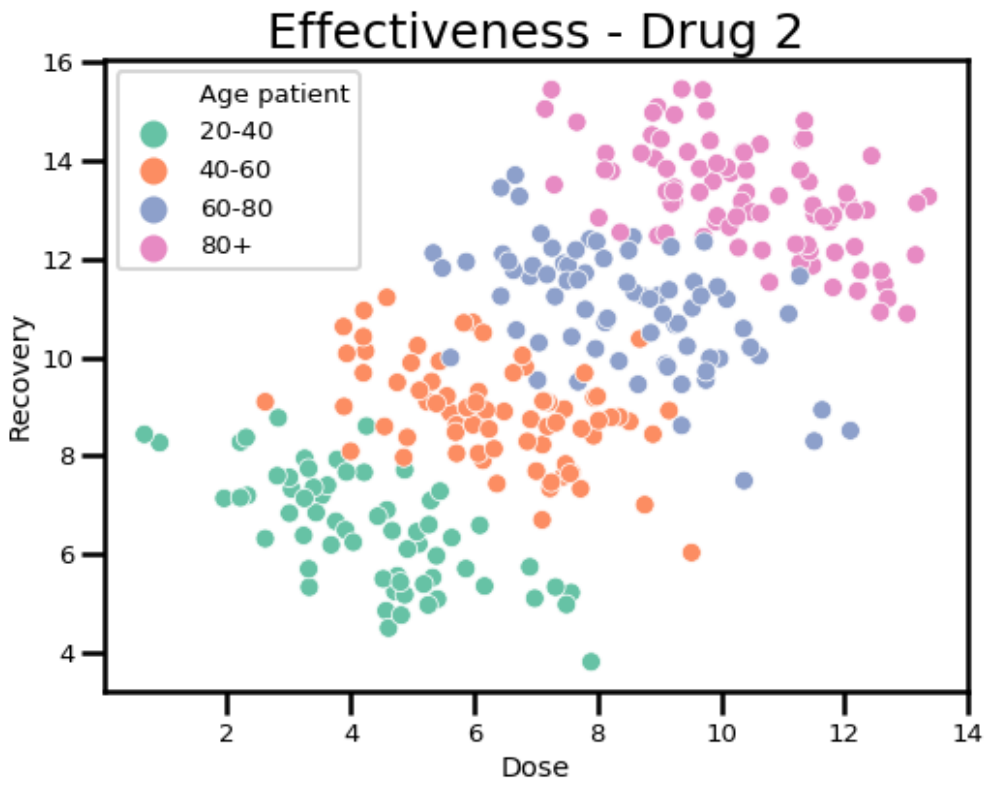
'A trend that appears in several different groups of data may disappear or reverse when these groups are combined.'

Days of recovery revisited - drug 2



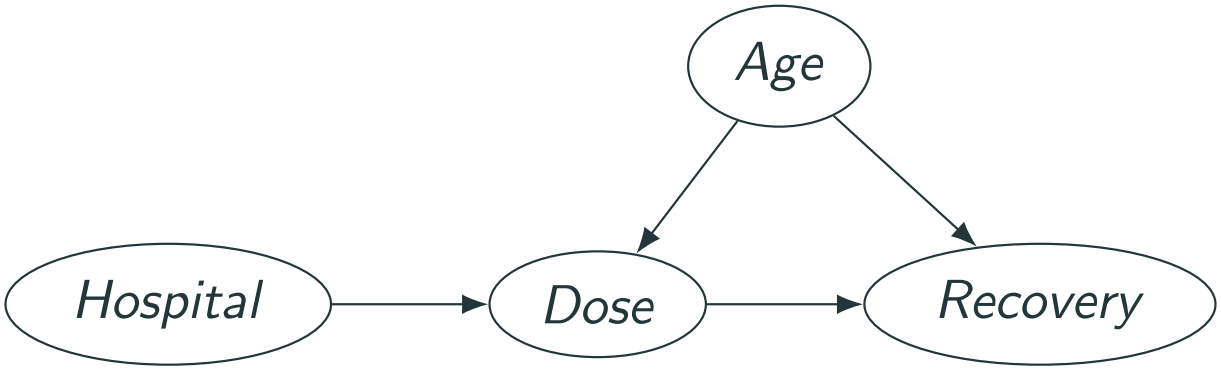
What is happening? Increasing the dose in drug 2 seem to make patient to spend more time at the hospital.

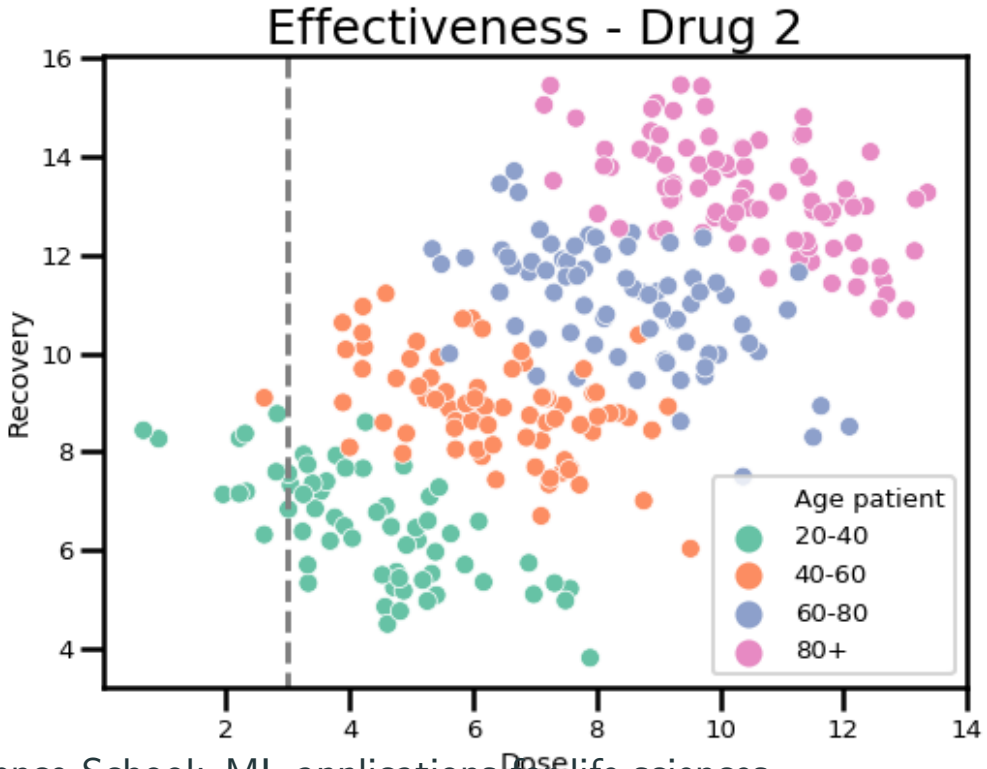
Days of recovery vs Dose - drug 2

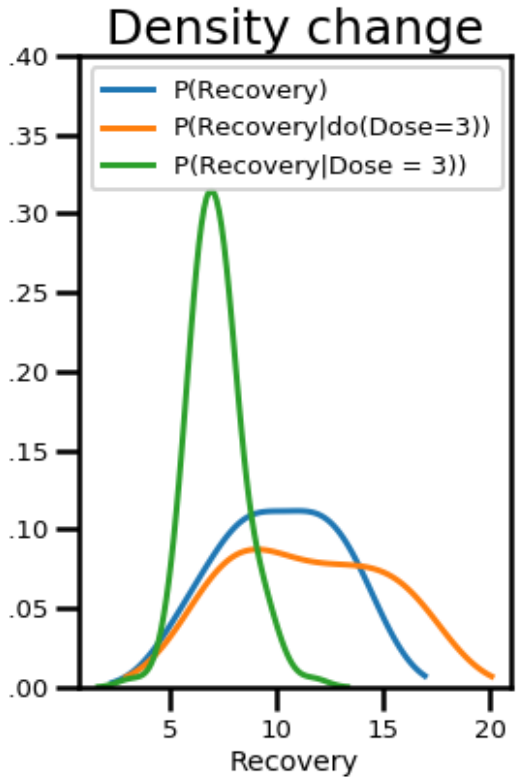


Age is acting as a confounder. The drug indeed is effective but older people suffer the disease more severely and requires a larger dose

Days of recovery vs. Dose - drug 2







Remarks

- The age is a confounder: affects both Dose and Recovery.
- With confounders: $\mathbb{P}(Y) \neq \mathbb{P}(Y|T = t) \neq \mathbb{P}(Y|do(T = t))$.
- Selection bias due to Age: doctors select small doses for young patients and large doses for old patients.
- We will make this more explicit later when we talk about propensity scores.
- We can compute the causal effect by averaging across age groups.

Second definition of a causal effect

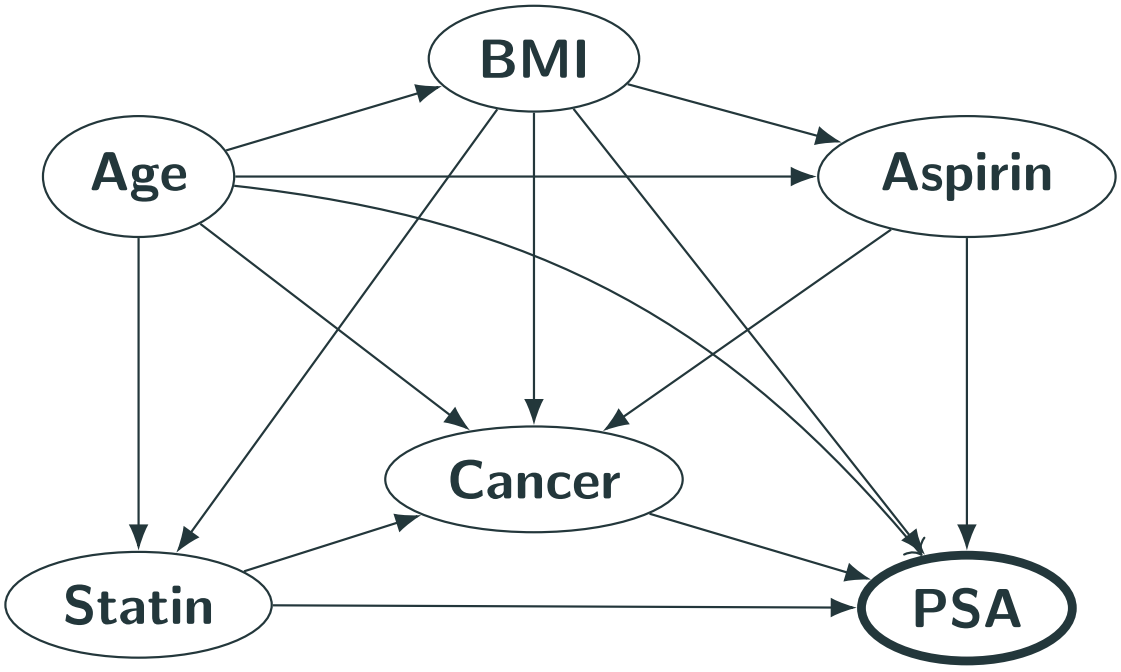
Given an SEM $\mathcal{S} := (G, P_\epsilon)$ the following statements are equivalent:

1. There is a causal effect from T to Y.
2. There is t' such that $\mathbb{P}(Y|do(T = t')) \neq \mathbb{P}(Y)$.
3. There is $t^\star$ and t' such that $\mathbb{P}(Y|do(T = t^\star) \neq \mathbb{P}(Y|do(T = t'))$.

do-calculus

What to do in general?

Which is the effect of aspirin in the Prostate-specific Antigen (PSA)?



We need to compute:

$$\mathbb{P}(PSA|do(Aspirin = x)$$

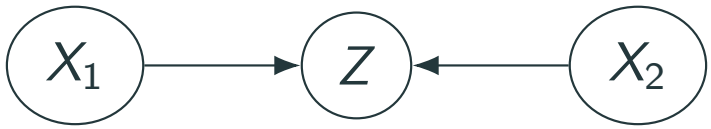
Do-calculus

- Proposed by Judea Pearl.
- Systematic way of knowing how to deconfound and effect given a causal graph (identification).
- Clearly separates identification from estimation (identification strategies are applicable with multiple models).

Confounders, colliders and mediators

Confounder:

Collider:



Descendants and mediators:

Back-door adjustment

Back-door adjustment

If Z is a **admissible adjustment set** then:

$$\mathbb{P}(Y|do(X=x)) = \sum_z \mathbb{P}(Y|X=x,Z=z)\mathbb{P}(Z=z)$$

$$\mathbb{P}(Y|do(X=x)) = \int \mathbb{P}(Y|X=x,Z=z)\mathbb{P}(Z=z)dz$$

- We can identify causal effects with observational distributions.
- We only need to control by Z, nothing else.
- Indeed, controlling by variables not in Z can be a terrible idea.

What is Z?

Admissible adjustment sets

A set Z is admissible ('sufficient') for adjustment if:

1. No element of Z is a descendant of X
2. The elements of Z 'block' all 'back-door' paths from X to Y (all paths that end with an arrow pointing to X).

D-separation

A set Z of nodes is said to block a path between X and Y if either:

- The path contains at least one arrow-emitting node that is in Z or
- The path contains at least one collider that is outside Z and has no descendant in Z.

D-separation

If Z blocks all paths from X to Y , it is said to "d-separate X and Y," and then, X and Y are independent given Z, written $X \perp\!\!\!\perp Y|Z$

Examples

Example 1:

Admissible sets: $\{H_1\}$ and $\{H_2\}$.

$$X \perp\!\!\!\perp Y | H_1$$

$$X \perp\!\!\!\perp Y | H_2$$

Example 2:

Admissible sets: $\{\emptyset\}$

$$X \perp\!\!\!\perp Y$$

(conditioning on colliders is a bad idea...)

Berkson's paradox

We know that the is no causal effect between the two diseases:

$$\mathbb{P}(Bone|do(Respiratory = Yes)) = \mathbb{P}(Bone)$$

	General population		
	Bone disease		
Respiratory disease	Yes	No	%Yes
Yes	17	207	**8.4%**
No	184	2376	**7.7%**

Berkson's paradox. On conditioning on colliders.

	General population			Hospitalizations last 6 months		
	Bone disease			Bone disease		
Respiratory disease	Yes	No	%Yes	Yes	No	%Yes
Yes	17	207	**7.6%**	5	15	**25%**
No	184	2376	**7.2%**	18	219	**7.6%**

Berkson's paradox

- The respiratory and bone diseases are independent.
- But they are conditionally dependent given hospitalization.

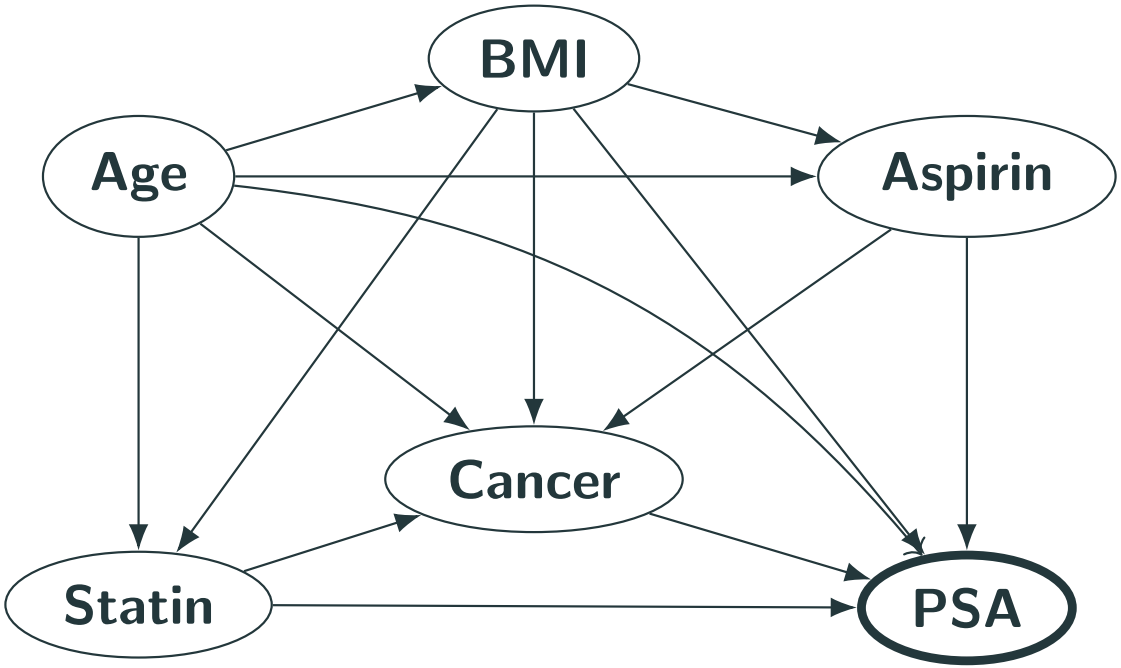
Controlling by hospitalization we would compute

$$\mathbb{P}(Bone|do(Resp. = Yes)) = \int \mathbb{P}(Bone|Resp. = Yes, Hosp.)\mathbb{P}(Hosp.) \neq 0$$

which is **wrong!** because we know that are causally independent.

Example in the Prostate-specific Antigen revisited

$$\mathbb{P}(PSA|do(Statin = 1))?$$



Cancer, Age and BMI d-separate PSA and Statin:

$$\mathbb{P}(PSA|do(St = 1)) = \int \mathbb{P}(PSA|S = 1, Can., Age, BMI)\mathbb{P}(Can., Age, BMI)$$

Steps when computing the back-door adjustment

$$\mathbb{P}(Y|do(X=x)) = \int \mathbb{P}(Y|X=x, Z=z)\mathbb{P}(Z=z)dz$$

1. Obtain the set Z via do-calculus (identification).
2. Fit $\mathbb{P}(Y|X=x, Z=z)$ (estimation).
3. Integrate with respect to $\mathbb{P}(Z=z)$ (estimation).
4. Validate the causal estimator (refutation)

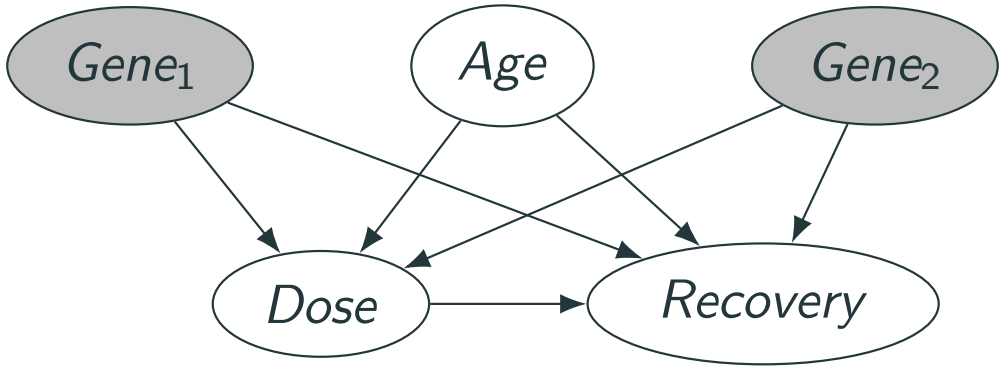
Open issues and questions

$$\mathbb{P}(Y|do(X=x)) = \int \mathbb{P}(Y|X=x, Z=z)\mathbb{P}(Z=z)dz$$

- The back-door provides a way to identify the effect but not to estimate it.
- Wrong causal graph $\rightarrow$ wrong Z (sol: RCT, instrumental variables).
- Z is high dimensional (sol: propensity scores).
- Some unobserved confounders (sol: front-door adj., LVMs).

Randomized control trials (RCT)

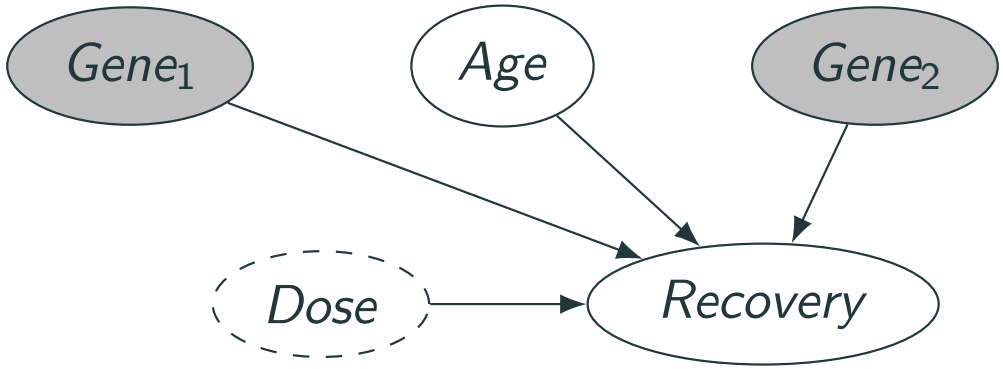
Randomized control trials (RCT)



- What to do if we don't know the causal graph?
- Or we have unobserved confounders ($Gene_1$ and $Gene_2$)?

In this example $Gene_1$ and $Gene_2$ are unobserved. We cannot control the confounders when computing the effect of *Dose* on *Recovery*.

Randomized control trials (RCT)



- Randomize dose: assign random levels independent of characteristics.
- This 'kills' all the incoming arrows from confounders.
- On the new data the adjustment set is the empty set!

In the RCT, no matter how complex the world is:

$$\mathbb{P}(Recovery|do(Dose = x)) = \mathbb{P}(Recovery|Dose = x)$$

Third definition of causal effect

Given an SEM $\mathcal{S}$ the following statements are equivalent:

1. There is a causal effect from X to Y.
2. There is x' such that $\mathbb{P}(Y|do(X = x')) \neq \mathbb{P}(Y)$.
3. Exist $x^\star$, x' such that $\mathbb{P}(Y|do(X = x^\star) \neq \mathbb{P}(Y|do(X = x'))$.
4. X and Y are not independent in $\mathbb{P}(X, Y|do(X = N_x))$ for any N_x whose distribution has full support.

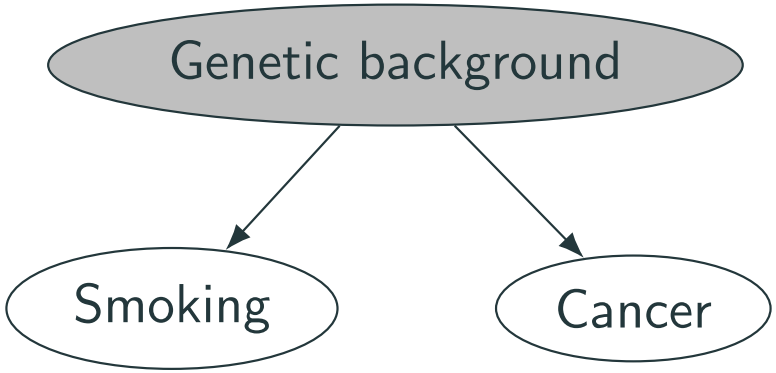
Instrumental variables

Does smoking cause cancer?

Seems obvious, but how can we be so sure?

Does smoking cause cancer?

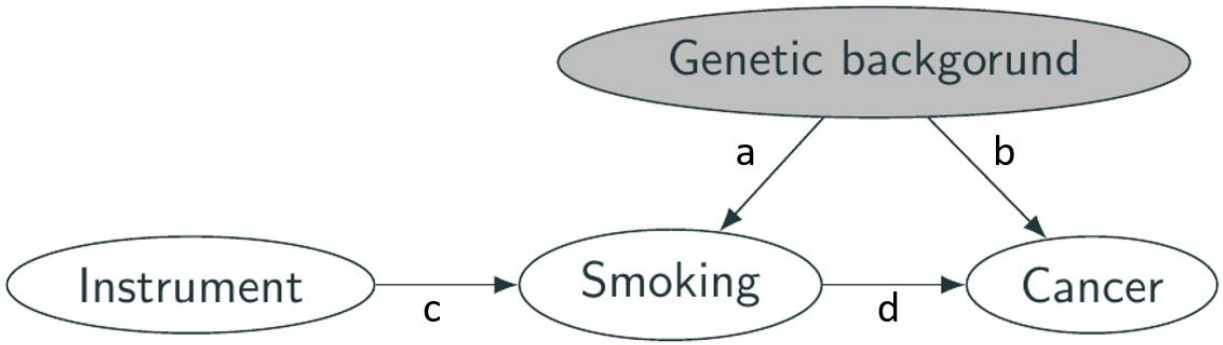
Controversy: What if there exist a latent confounder that affects both smoking and cancer?



Hard to prove:

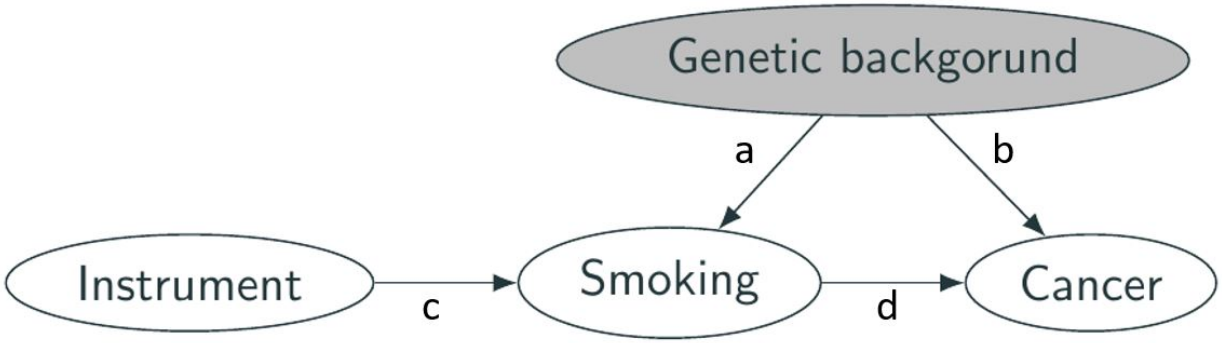
- We cannot control by the unobserved potential confounders.
- We cannot do a RCT because it would be unethical.

Instrumental variable



- **Instrument:** Observable variable that causes the treatment (smoking) but is independent of the effect and the confounders.
- Justified based on background knowledge.
- Causal effects need to be linear.

Instrumental variable



How does it works? two-stage regression:

1. Run linear regression $S = cI + e_s$ and obtain $\hat{c}$.
2. Compute predictions $\hat{S} = \hat{c}I$.
3. Run linear regression $C = d\hat{S} + e_c$ and obtain $\hat{d}$.
4. $\hat{d}$ is the estimated causal effect.

Instrumental variable: text on the box

Why does it work?

- The test on the box affect the smoking habit but nothing else.
- The predictions on smoking are a pseudo-random experiment.
- They contain information from the instrument but not from the genetic background.

Instrumental variable: text on the box

Why does it work?

- The test on the box affect the smoking habit but nothing else.
- The predictions on smoking are a pseudo-random experiment.
- They contain information from the instrument but not from the genetic background.

Counterfactuals

Counterfactuals

Counterfactuals.

Counterfactual statements

- Had Hilary Clinton been a man, she would have won the general election to Donald Trump
- Had the doctor offered drug A, the patient would be now healthy.
- If my friend Pablo hadn't watered my plants, today they would be all dead.

Common aspects of counterfactuals

- They are personalized: unit-level counterfactuals.
- Probability of a unit-level event in a parallel universe.

Joint model of both universes

Observed universe

$$\begin{aligned} H &= f_h(\epsilon_h) \\ T &= f_t(T, \epsilon_t) \\ Y &= f_y(X, \epsilon_y) \end{aligned}$$

Intervened universe

$$\begin{aligned} H &= f_h(\epsilon_h) \\ T &= t \\ Y &= f_y(t, \epsilon_y) \end{aligned}$$

Observational distribution

$$\mathbb{P}(H, T, Y)$$

Post-intervention distribution

$$P^{do(T=t)}(H, Y) = \mathbb{P}(H^*, Y^*)$$

Joint distribution over variables in both universes

$$\boxed{\mathbb{P}(H, T, Y, H^*, Y^*)}$$

Counterfactual: formal definition

1. Having observed the patient k (realization of ϵ_h, ϵ_t and ϵ_y):

   For patient k we know that $H = h_k$, $T = t_k$ and $Y = y_k$.

2. And had we set the dose level T to t:

   We created a parallel universe where $do(T = t)$.

3. Would the outcome (Y_k^*) had been different?

   Which is the probability of Y_k^*?

Unit level counterfactual, $Y_t(k)$

$$\mathbb{P}(Y^*|H = h_k, T = t_k, Y = y_k, T^* = t)$$

Fun fact

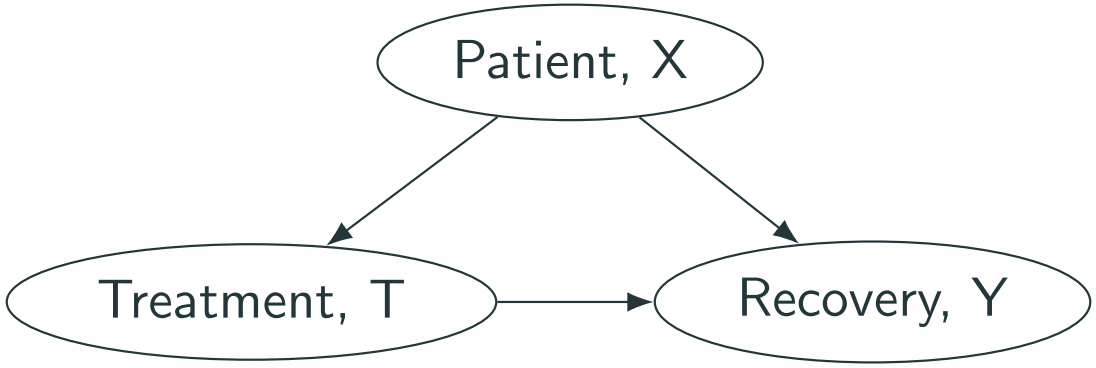
Causal effects are the averaged unit-level counterfactuals for all the individuals of the population:

$$\begin{aligned} \int \mathbb{P}(Y^*|H,T,Y,T^*=t)\mathbb{P}(H,T,Y) &= \mathbb{P}(Y^*|T^*=t) \\ &= \mathbb{P}(Y|do(T=t)) \end{aligned}$$

Potential outcomes framework

Common setup in all these examples

Common causal graph in all these examples:



- In Pearl's framework T, X, Y are the units of analysis of the theory.
- In Rubin's framework the potential outcomes $Y_0(x)$ and $Y_1(x)$ are.

Potential outcomes framework (Rubin-Neyman)

Each unit x_k has two potential outcomes of a treatment $t_k = \{0, 1\}$:

- $Y_0(x_k)$ if the unit is in the control group.
- $Y_1(x_k)$ if the unit is in the treatment group.

Observed outcome (given t_k):

$$y_k = t_k Y_1(x_k) + (1 - t_k) Y_0(x_k)$$

Counterfactual outcome:

$$y_k^* = (1 - t_k) Y_1(x_k) + t_k Y_0(x_k)$$

Quantities of interest

Individual treatment effect:

$$ITE(x_k) := \mathbb{E}_{\mathbb{P}(Y_1|x_k)}[Y_1(x_k)] - \mathbb{E}_{\mathbb{P}(Y_0|x_k)}[Y_0(x_k)]$$

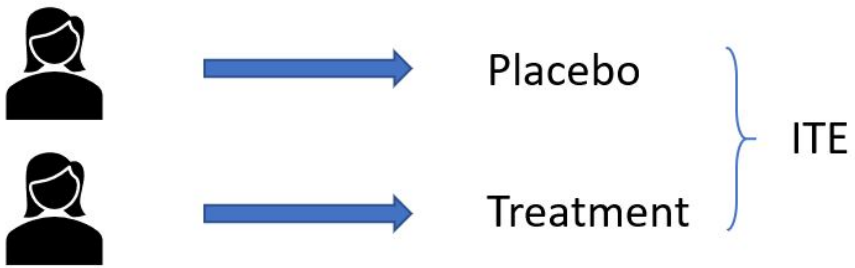
(Note that $ITE(x_k) \neq \mathbb{E}_{\mathbb{P}(Y|x_k)}[Y_1(x_k) - Y_0(x_k)]$)

Averaged treatment effect:

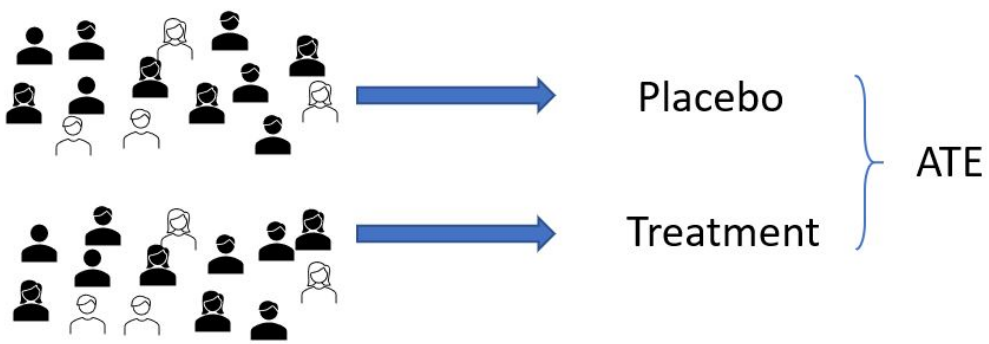
$$ATE := \mathbb{E}_{\mathbb{P}(X)}[Y_1(X) - Y_0(X)] = \mathbb{E}_{\mathbb{P}(X)}[ITE(X)]$$

How to compute these quantities?

Individual treatment effect: easy if x_k has a twin and that has been exposed to a different condition.



Averaged treatment effect: easy if I can run a RCT.



How to compute these quantities?

Individual treatment effect: easy if x_k has a twin and that has been exposed to a different condition.

$$ITE(x_k) = Y_1(x_k) - Y_0(x_k)$$

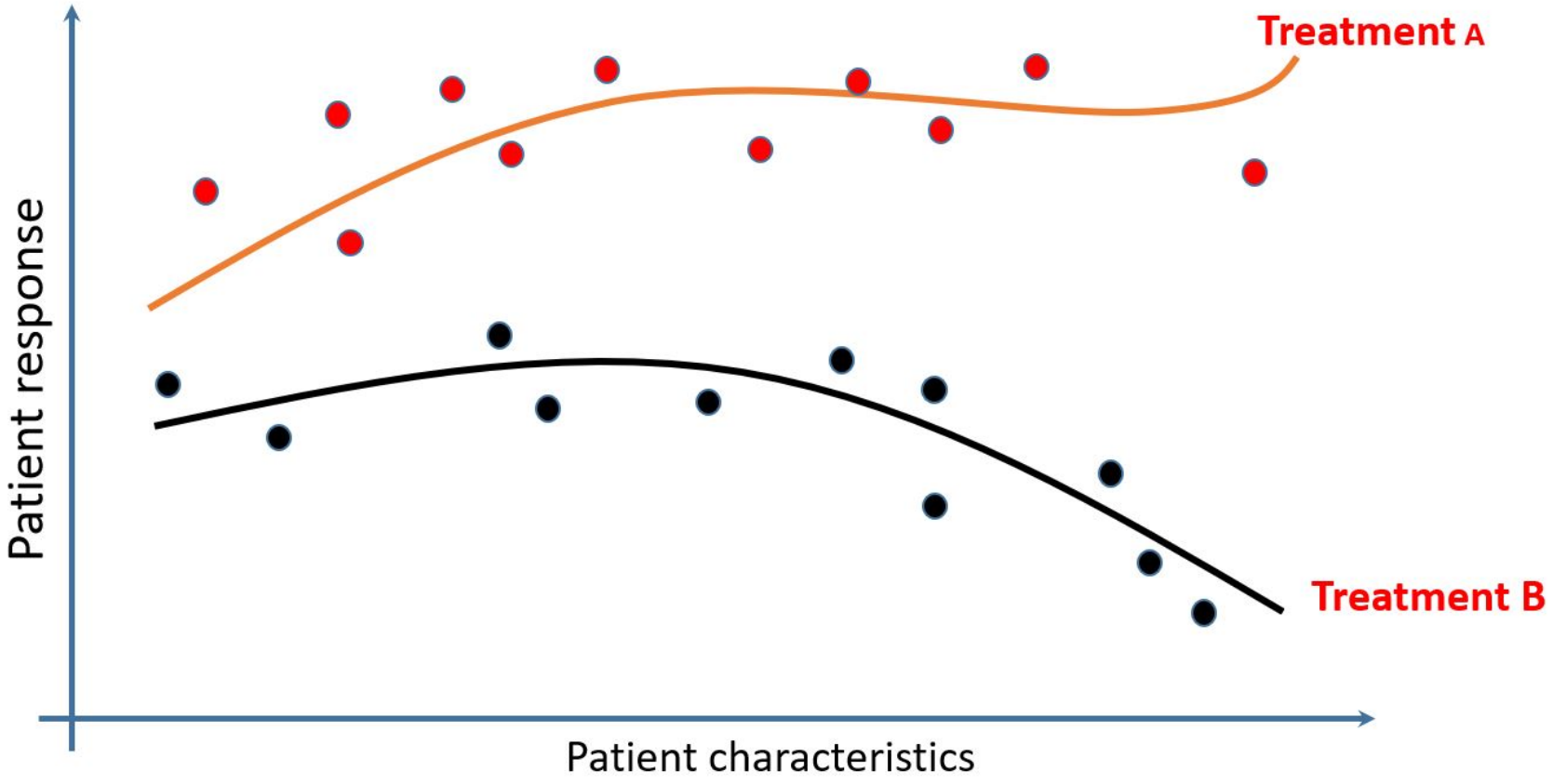
Averaged treatment effect: easy if I can run a RCT ($n/2$ treated)

$$ATE = \frac{2}{n} \sum_{x_k \in G_1} Y_1(x_k) - \frac{2}{n} \sum_{x_k \in G_0} Y_0(x_k)$$

Fundamental problem of causal inference

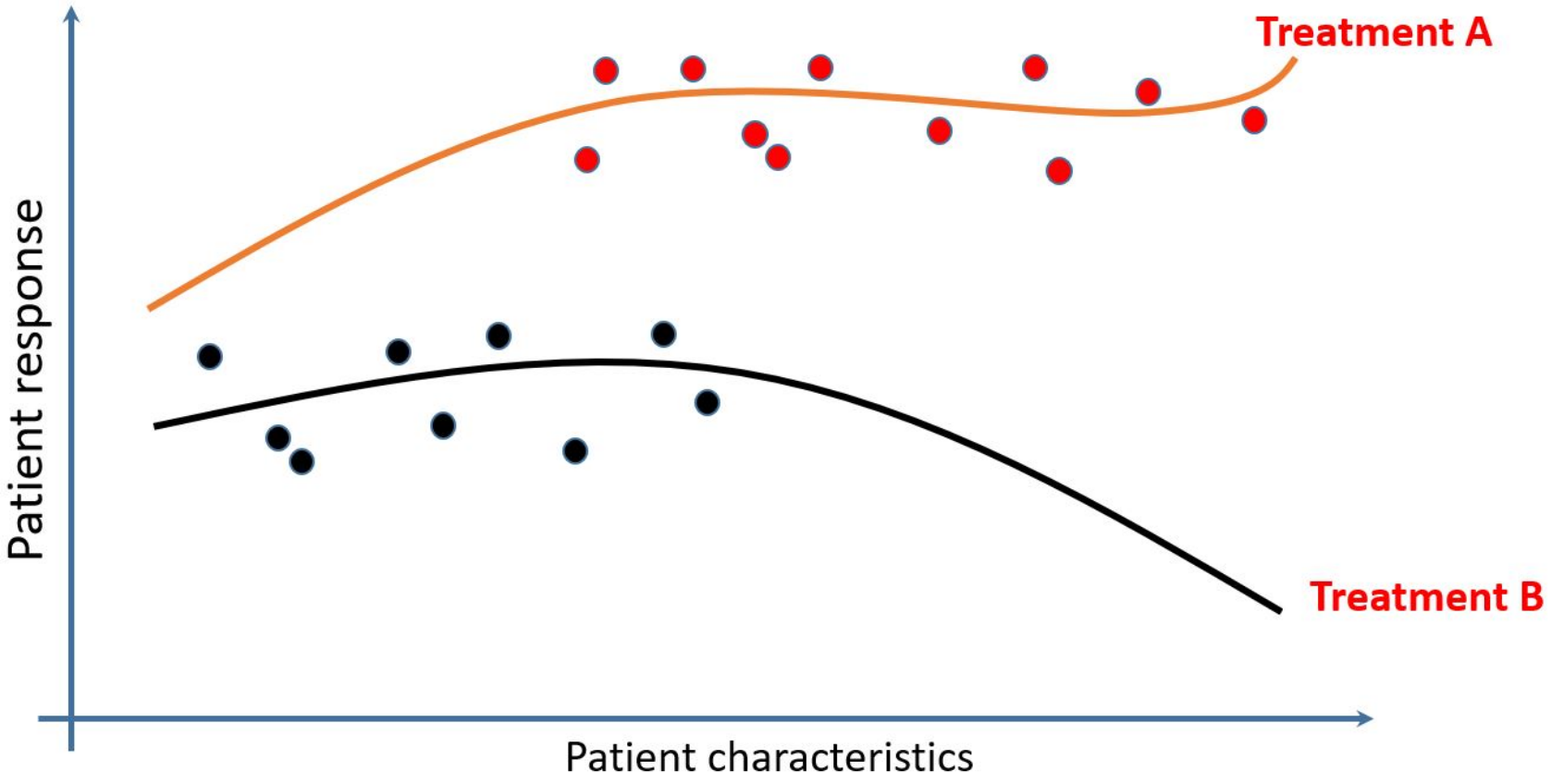
- In most cases we don't have twins and we cannot run RCT.
- We have observational data and we only observe one of the two outcomes.
- We have confounders: the value of X affects the probability of assignment.

Example



The patient characteristics don't affect the probability of assignment (like in RCT).

Example



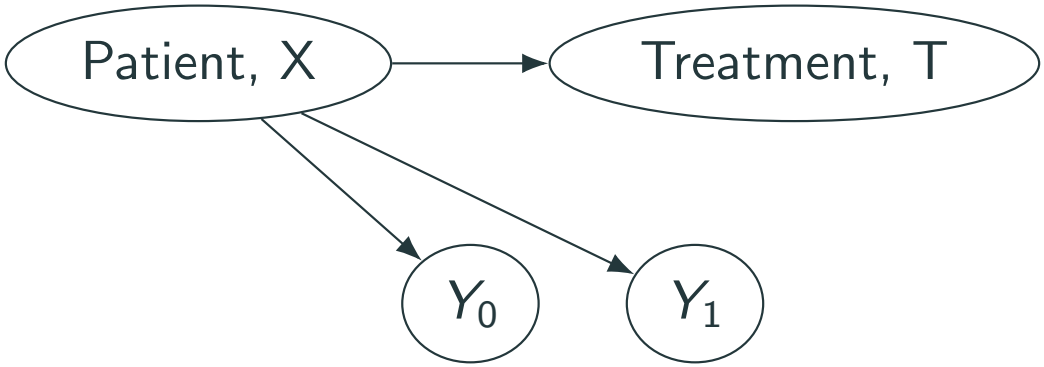
The patient characteristics don't affect the probability of assignment (confounding).

Assumption 1: ignorability

X: *age, BMI, clinical condition.*

T: *Cancer treatment or placebo.*

Y_0, Y_1: *response under placebo and treatment.*



$$(Y_0, Y_1) \perp\!\!\!\perp T|x$$

- X contains all the information about the assignment mechanism.
- Not unmeasured confounders.

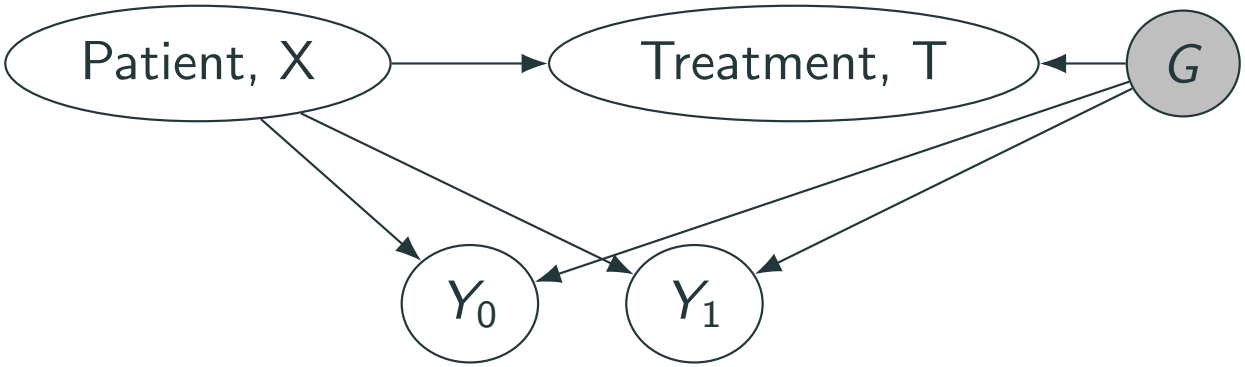
Assumption 1: ignorability

X: *age, BMI, clinical condition.*

T: *Cancer treatment or placebo.*

Y_0, Y_1: *response under placebo and treatment.*

G: *unobserved genetic background.*



$$(Y_0, Y_1) \not\perp\!\!\!\perp T|x$$

- Unmeasured confounders.

Assumption 2: common support

X: *age, BMI, clinical condition.*

T: *Cancer treatment or placebo.*

Y_0, Y_1: *response under placebo and treatment.*

$$\mathbb{P}(T = t|X = x) > 0, \forall t, x$$

Non-zero probability of assignment for all individuals in the population.

The adjustment formula

$$ATE := \mathbb{E}_{\mathbb{P}(x)}[Y_1(x) - Y_0(x)] = \mathbb{E}_{\mathbb{P}(x)}[Y_1(x)] - \mathbb{E}_{\mathbb{P}(x)}[Y_0(x)]$$

Under the assumption of ignorability:

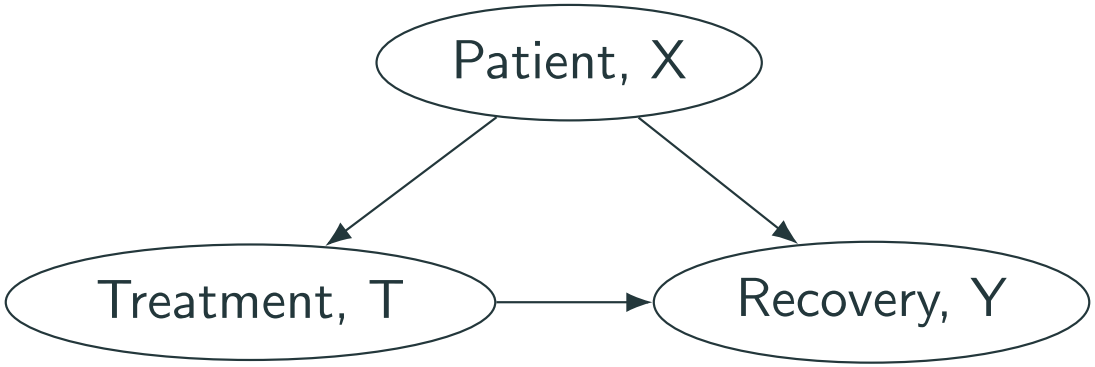
$$\begin{aligned}
\mathbb{E}_{\mathbb{P}(X)}[Y_0(x)] &= \mathbb{E}_{\mathbb{P}(X)}[Y_0(X)] \\
&= \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_0|x)}[Y_0|x]] \\
&= \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_0|x)}[Y_0|x, T=0]]
\end{aligned}$$

And therefore:

$$ATE = \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_1|x)}[Y_1|x, T=1]] - \mathbb{E}_{\mathbb{P}(X)}[\mathbb{E}_{\mathbb{P}(Y_0|x)}[Y_0|x, T=0]]$$

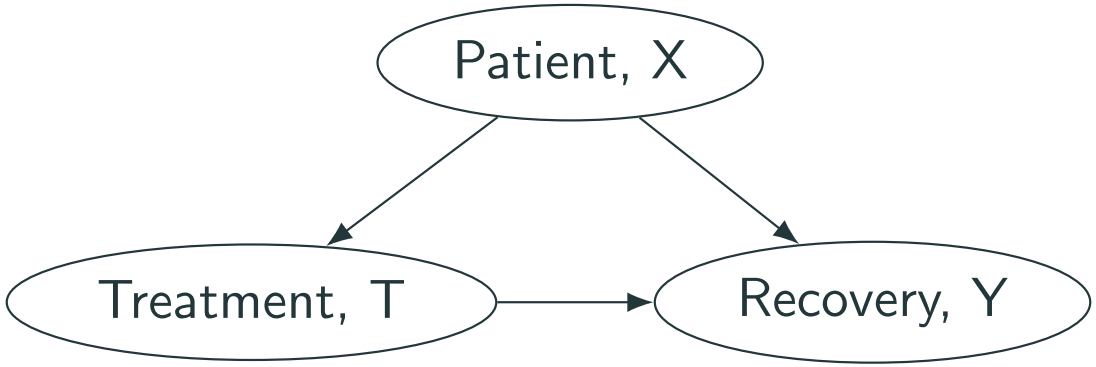
which can be estimated from data

The adjustment formula and the back door criterion



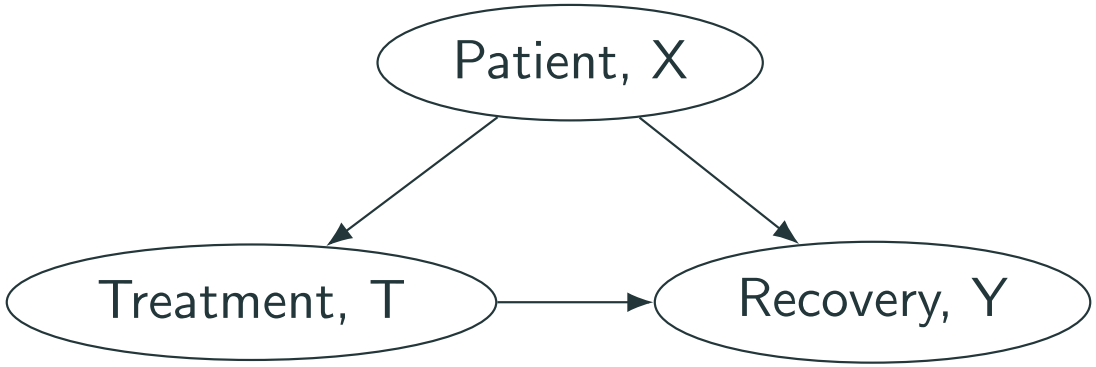
$$ATE \quad = \quad \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y]$$

The adjustment formula and the back door criterion



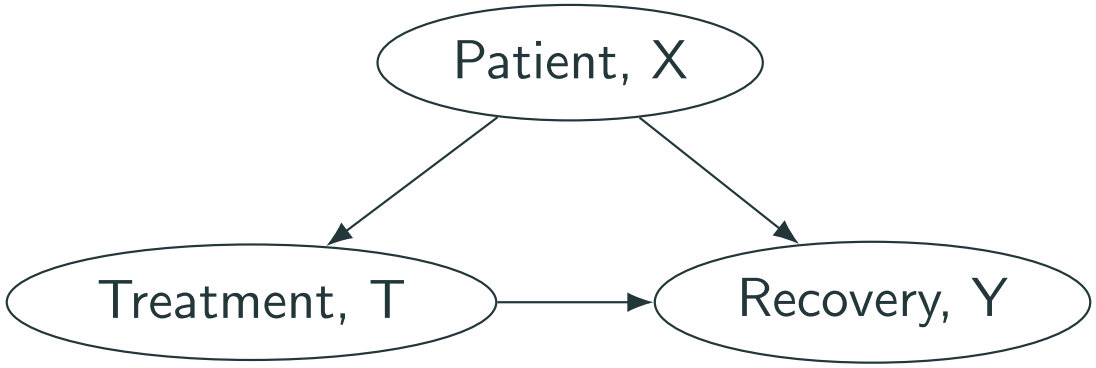
$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX
\end{aligned}
$$

The adjustment formula and the back door criterion



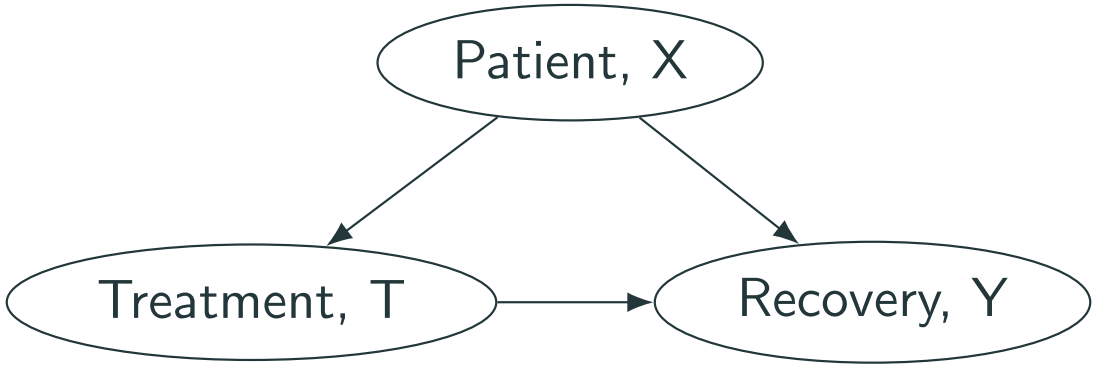
$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX \\
&= \int Y_1\mathbb{P}(Y_1|X, T=1)\mathbb{P}(X)dX - \int Y_0\mathbb{P}(Y_0|X, T=0)\mathbb{P}(X)dX
\end{aligned}
$$

The adjustment formula and the back door criterion



$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX \\
&= \int Y_1\mathbb{P}(Y_1|X, T=1)\mathbb{P}(X)dX - \int Y_0\mathbb{P}(Y_0|X, T=0)\mathbb{P}(X)dX \\
&= \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_1|x)}[Y_1|x, T=1]] - \mathbb{E}_{\mathbb{P}(X)}[\mathbb{E}_{\mathbb{P}(Y_0|x)}[Y_0|x, T=0]]
\end{aligned}
$$

The adjustment formula and the back door criterion



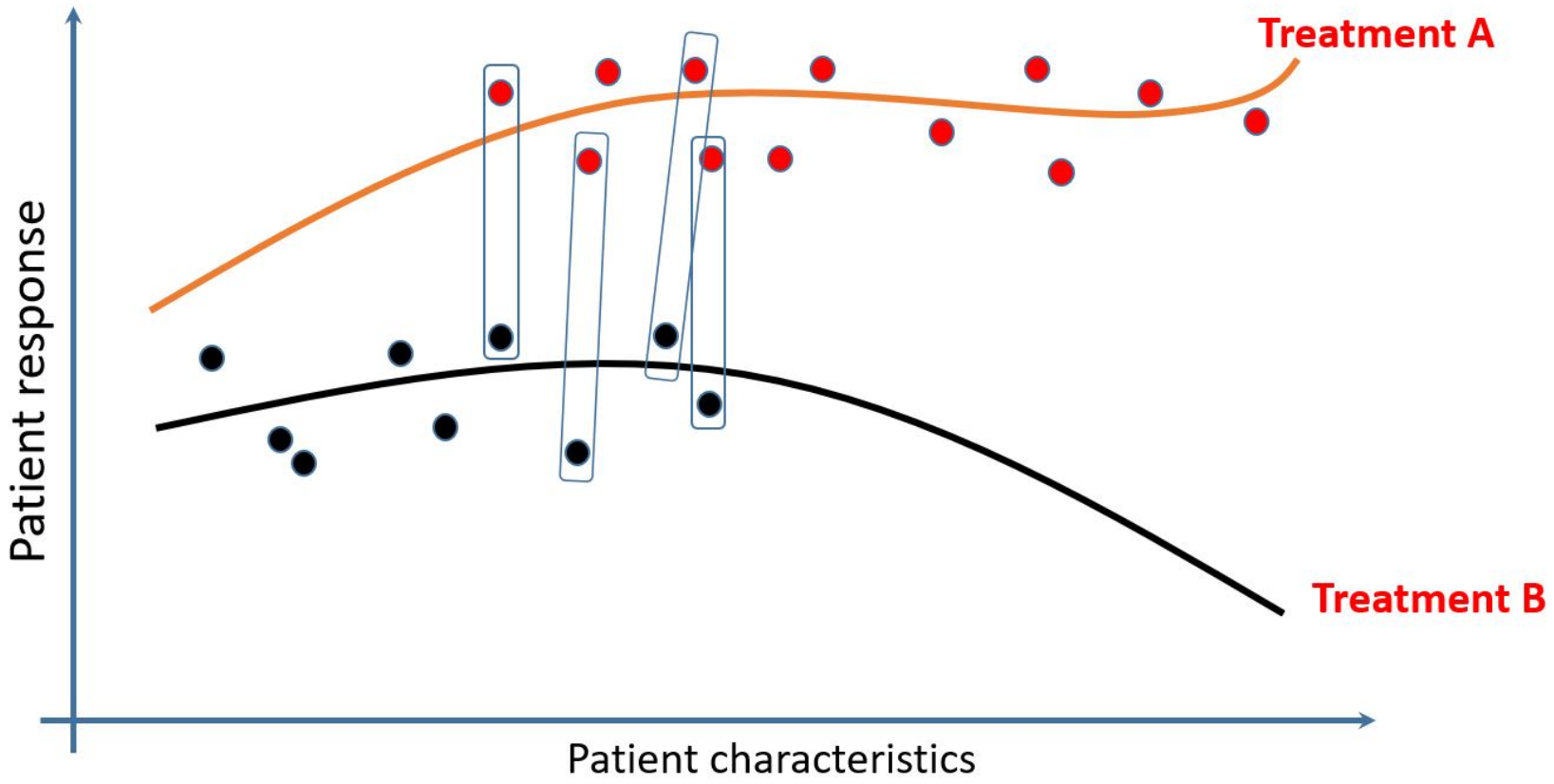
$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX \\
&= \int Y_1\mathbb{P}(Y_1|X, T=1)\mathbb{P}(X)dX - \int Y_0\mathbb{P}(Y_0|X, T=0)\mathbb{P}(X)dX \\
&= \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_1|x)}[Y_1|x, T=1]] - \mathbb{E}_{\mathbb{P}(X)}[\mathbb{E}_{\mathbb{P}(Y_0|x)}[Y_0|x, T=0]]
\end{aligned}
$$

The backdoor criterion and the adjustment formula and indeed equivalent in the binary treatment case.

Matching

Matching

Find 'twins' and average the difference between the twins.



Matching

1. Define a metric between the patients (1-NN for instance).
2. For each x_k find the closest twin $m = \arg\min d(x_m, x_k)$.

If unit k is treated:

$$I\hat{T}E(x_k) = y_k - y_m(k)$$

If unit k is control:

$$I\hat{T}E(x_k) = y_m(k) - y_k$$

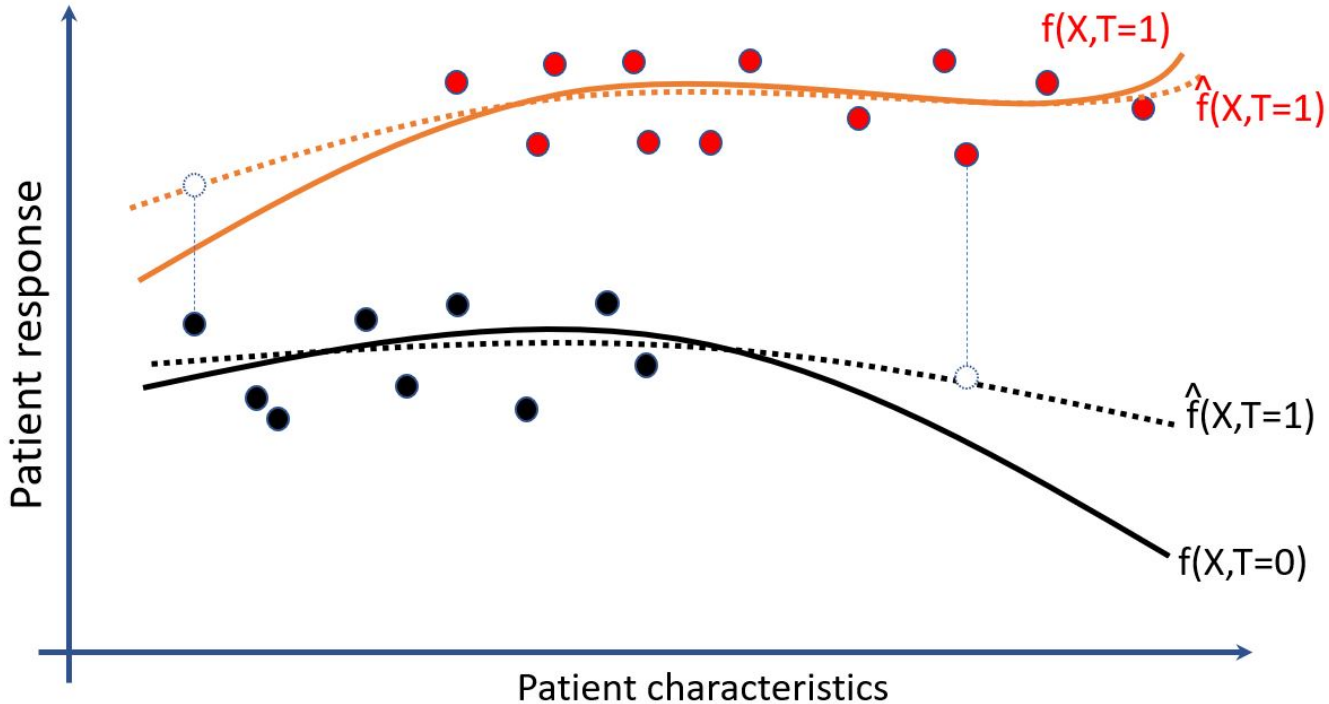
Averaged treatment effect:

$$A\hat{T}E = \frac{1}{n}\sum_{k=1}^{n} I\hat{T}E(x_k)$$

Covariate Adjustment

Covariate Adjustment

- Model the relationship between treatments, response and patients
- Treats the causal problem as supervised learning problem.
- Can be used to estimate both ITE and ATE.



$$Y = f(X, T)$$

Computing effects under covariate adjustment

Under ignorability we have that

$$ATE := \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_1|x)}[Y_1|x, T=1]] - \mathbb{E}_{\mathbb{P}(x)}[\mathbb{E}_{\mathbb{P}(Y_0|x)}[Y_0|x, T=0]]$$

Therefore, after fitting a model $f(x,t) = \mathbb{E}[Y_t|x, T=t]$

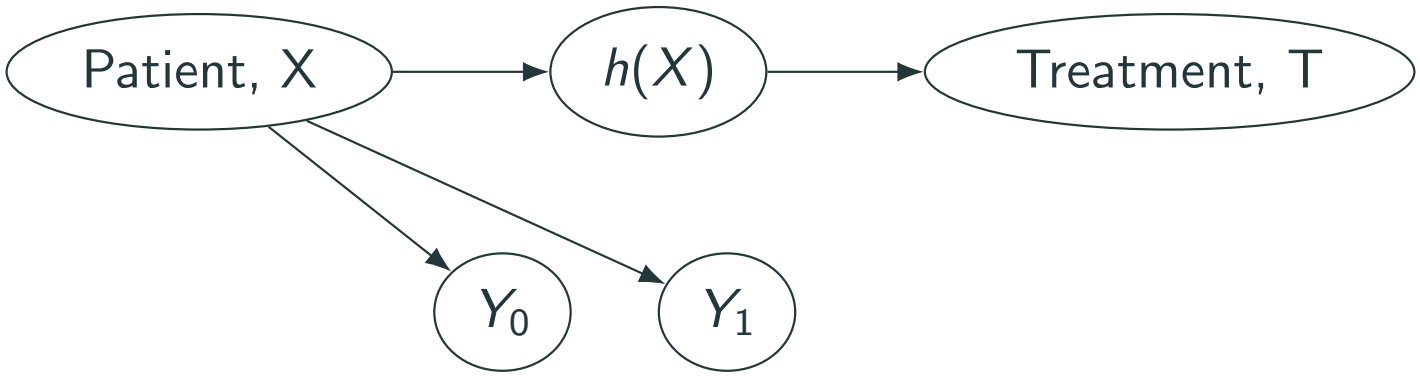
$$\hat{ITE}(x_k) = \hat{f}(x_k, T=1) - \hat{f}(x_k, T=0)$$

$$\hat{ATE} = \frac{1}{n}\sum_{k=1}^{n}[\hat{f}(x_k, T=1) - \hat{f}(x_k, T=0)]$$

Propensity scores

Propensity score

Balancing score: function $h(X)$ such that $X \perp\!\!\!\perp T | h(x)$.



Propensity score: probability of assignment to treatment.

$$e(x) := \mathbb{P}(T = 1 | X = x)$$

Propensity score

$$e(x) := \mathbb{P}(T = 1|X = x)$$

- The propensity $e(x)$ score is a balancing score.
- Coarsest balancing score function (takes a multidimensional object, x_k and transforms it into one dimension).
- Can be estimated with any supervised learning method.
- It is strong ignorable! $(Y_0, Y_1) \not\perp\!\!\!\perp T|h(x)$.
- Controlling by X and $e(X)$ is equivalent!

Propensity score re-weighting

In RCTs we now that:

$$\mathbb{P}(T = 1|X = x) = \mathbb{P}(T = 0|X = x) = 0.5$$

Under confounding variables, however:

$$\mathbb{P}(T = 1|X = x) \neq \mathbb{P}(T = 0|X = x) \neq 0.5$$

Idea of reweighing: find $w_0(x)$ and $w_1(x)$ such that:

$$\mathbb{P}(T = 1|X = x) \cdot w_1(x) \approx \mathbb{P}(T = 0|X = x) \cdot w_0(x) \approx 0.5$$

Propensity score re-weighting

Sample with n elements (x_k, t_k, y_k) ($n/2$ treated).

Step 1: Use ML to estimate $\mathbb{P}(T = 1|X = x)$

Step 2:

$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{\hat{\mathbb{P}}(T = 1|X = x_k)} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{\hat{\mathbb{P}}(T = 0|X = x_k)}$$

Propensity score re-weighting

Sample with n elements (x_k, t_k, y_k) ($n/2$ treated).

Step 1: Use ML to estimate $\mathbb{P}(T = 1|X = x)$

Step 2:

$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{\hat{\mathbb{P}}(T = 1|X = x_k)} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{\hat{\mathbb{P}}(T = 0|X = x_k)}$$

Reminder: In RCT we have $\mathbb{P}(T = 1|X = x) = 0.5$ and therefore:

$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{0.5} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{0.5}$$

Propensity score re-weighting

Sample with n elements (x_k, t_k, y_k) ($n/2$ treated).

Step 1: Use ML to estimate $\mathbb{P}(T = 1|X = x)$

Step 2:

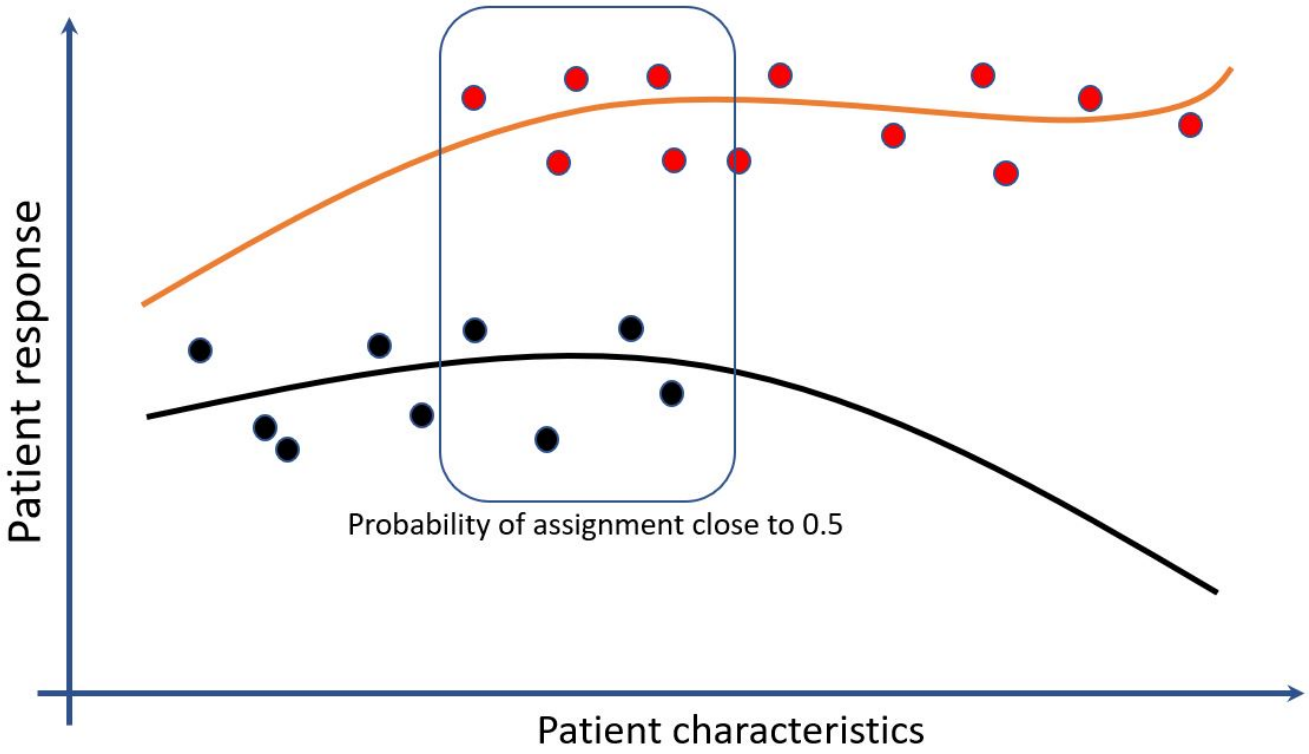
$$\hat{ATE} = \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{\hat{\mathbb{P}}(T = 1|X = x_k)} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{\hat{\mathbb{P}}(T = 0|X = x_k)}$$

Reminder: In RCT we have $\mathbb{P}(T = 1|X = x) = 0.5$ and therefore:

$$\begin{aligned} \hat{ATE} &= \frac{1}{n} \sum_{k \in G_1} \frac{y_k}{0.5} - \frac{1}{n} \sum_{k \in G_0} \frac{y_k}{0.5} \\ &= \frac{2}{n} \sum_{k \in G_1} y_k - \frac{2}{n} \sum_{k \in G_0} y_k \end{aligned}$$

Further intuition of the propensity score

- The harder the assignment $\mathbb{P}(T|X) \approx 0.5$ the higher the weight.
- Leverage those points that look more like a random assignment.



Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$ATE = \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y]$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX
\end{aligned}
$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX \\
&= \int Y_1\mathbb{P}(Y_1|X, T=1)\mathbb{P}(X)dX - \int Y_0\mathbb{P}(Y_0|X, T=0)\mathbb{P}(X)dX
\end{aligned}
$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X, T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X, T=0)\mathbb{P}(X)dX \\
&= \int Y_1\mathbb{P}(Y_1|X, T=1)\mathbb{P}(X)dX - \int Y_0\mathbb{P}(Y_0|X, T=0)\mathbb{P}(X)dX \\
&= \int Y\frac{\mathbb{P}(Y, X|T=1)}{\mathbb{P}(T=1|X)} - \int Y\frac{\mathbb{P}(Y, X|T=0)}{\mathbb{P}(T=0|X)}
\end{aligned}
$$

Propensity score re-weighting and the backdoor adjustment

Again, all can be derived from the back-door adjustment:

$$
\begin{aligned}
ATE &= \mathbb{E}_{\mathbb{P}(Y|do(T=1))}[Y] - \mathbb{E}_{\mathbb{P}(Y|do(T=0))}[Y] \\
&= \int Y\mathbb{P}(Y|X,T=1)\mathbb{P}(X)dX - \int Y\mathbb{P}(Y|X,T=0)\mathbb{P}(X)dX \\
&= \int Y_1\mathbb{P}(Y_1|X,T=1)\mathbb{P}(X)dX - \int Y_0\mathbb{P}(Y_0|X,T=0)\mathbb{P}(X)dX \\
&= \int Y\frac{\mathbb{P}(Y,X|T=1)}{\mathbb{P}(T=1|X)} - \int Y\frac{\mathbb{P}(Y,X|T=0)}{\mathbb{P}(T=0|X)} \\
&\approx \frac{1}{n}\sum_{k\in G_1}\frac{y_k}{\mathbb{P}(T=1|X=x_k)} - \frac{1}{n}\sum_{k\in G_0}\frac{y_k}{\mathbb{P}(T=0|X=x_k)}
\end{aligned}
$$

We use that

$$\mathbb{P}(Y|X,T) = \frac{\mathbb{P}(Y,X,T)}{\mathbb{P}(T|X)\mathbb{P}(X)}$$

Further comments about the propensity score

- Can be used as a metric for matching (propensity score matching).
- Can be used to directly control the causal effect in the back-door.
- Transforms the problem of computing a high dimensional integral, into learning a high dimensional mapping.
- Same idea as importance sampling in machine learning.

Questions?