Machine Learning for econometrics

Reminders of potential outcomes and Directed Acyclic Graphs

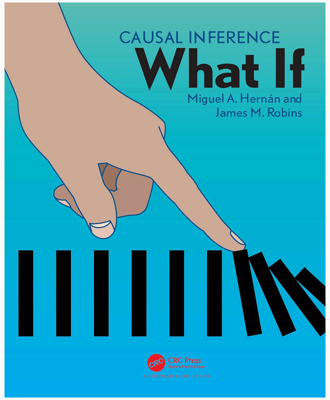
Matthieu Doutreligne

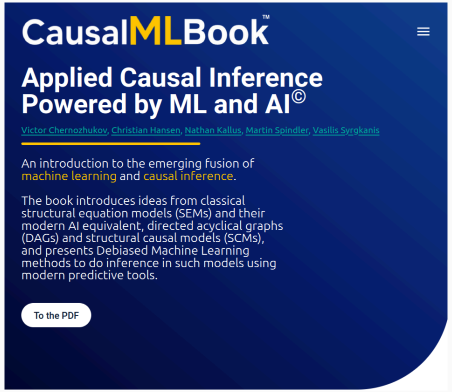
Thanks to Judith Abecassis for the slides on DAGs

Contents

Introduction

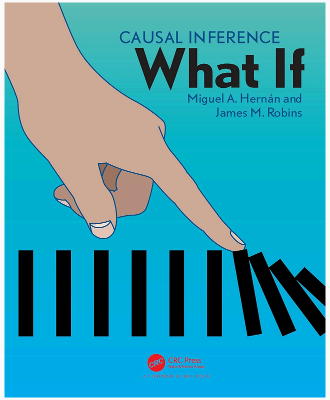
Causal inference: subfield of statistics dealing with "why questions"

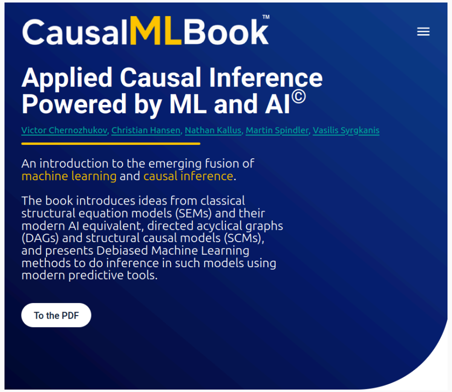




At the center of epidemiology (Hernan & Robins, 2020), econometrics (Chernozhukov et al., 2024), social sciences,

Causal inference: subfield of statistics dealing with "why questions"





At the center of epidemiology (Hernan & Robins, 2020), econometrics (Chernozhukov et al., 2024), social sciences, machine learning...

Now, bridging with machine learning (Kaddour et al., 2022) : Fairness, reinforcement learning, causal discovery, causal inference for LLM, causal representations...

What is a "why question"?

Economics: How does supply and demand (causally) depend on price?

What is a "why question"?

Economics: How does supply and demand (causally) depend on price?

Policy: Are job training programmes actually effective?

What is a "why question"?

Economics: How does supply and demand (causally) depend on price?

Policy: Are job training programmes actually effective?

Epidemiology: How does this threatment affect the patient's health?

What is a "why question"?

Economics: How does supply and demand (causally) depend on price?

Policy: Are job training programmes actually effective?

Epidemiology: How does this threatment affect the patient’s health?

Public health : Is this prevention campaign effective?

What is a "why question"?

Economics: How does supply and demand (causally) depend on price?

Policy: Are job training programmes actually effective?

Epidemiology: How does this threatment affect the patient’s health?

Public health : Is this prevention campaign effective?

Psychology: What is the effect of family structure on children’s outcome?

What is a "why question"?

Economics: How does supply and demand (causally) depend on price?

Policy: Are job training programmes actually effective?

Epidemiology: How does this threatment affect the patient's health?

Public health : Is this prevention campaign effective?

Psychology: What is the effect of family structure on children's outcome?

Sociology: What is the effect of social media on political opinions?

This is different from predictive questions

Prediction (ML): What usually happens in a given situation?

$$X \longrightarrow Y$$

Prediction models (X, Y)

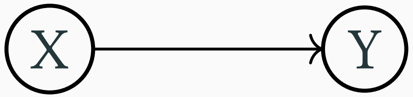
Prediction (ML): What usually happens in a given situation?

$$X \rightarrow Y$$

Prediction models (X, Y)

What will be the weather tomorrow?

Prediction (ML): What usually happens in a given situation?



Prediction models (X, Y)

What will be the weather tomorrow?

What will be the outcome of the next election?

This is different from predictive questions

Prediction (ML): What usually happens in a given situation?

$$X \longrightarrow Y$$

Prediction models (X, Y)

What will be the weather tomorrow?

What will be the outcome of the next election?

How many people will get infected by flue next season?

This is different from predictive questions

Prediction (ML): What usually happens in a given situation?

$$X \longrightarrow Y$$

Prediction models (X, Y)

What will be the weather tomorrow?

What will be the outcome of the next election?

How many people will get infected by flue next season?

What is the cardio-vacular risk of this patient?

Prediction (ML): What usually happens in a given situation?

$$X \longrightarrow Y$$

Prediction models (X, Y)

What will be the weather tomorrow?

What will be the outcome of the next election?

How many people will get infected by flue next season?

What is the cardio-vacular risk of this patient?

How much will the price of a stock be tomorrow?

This is different from predictive questions

Prediction (ML): What usually happens in a given situation?

$$X \longrightarrow Y$$

Prediction models (X, Y)

What will be the weather tomorrow?

What will be the outcome of the next election?

How many people will get infected by flue next season?

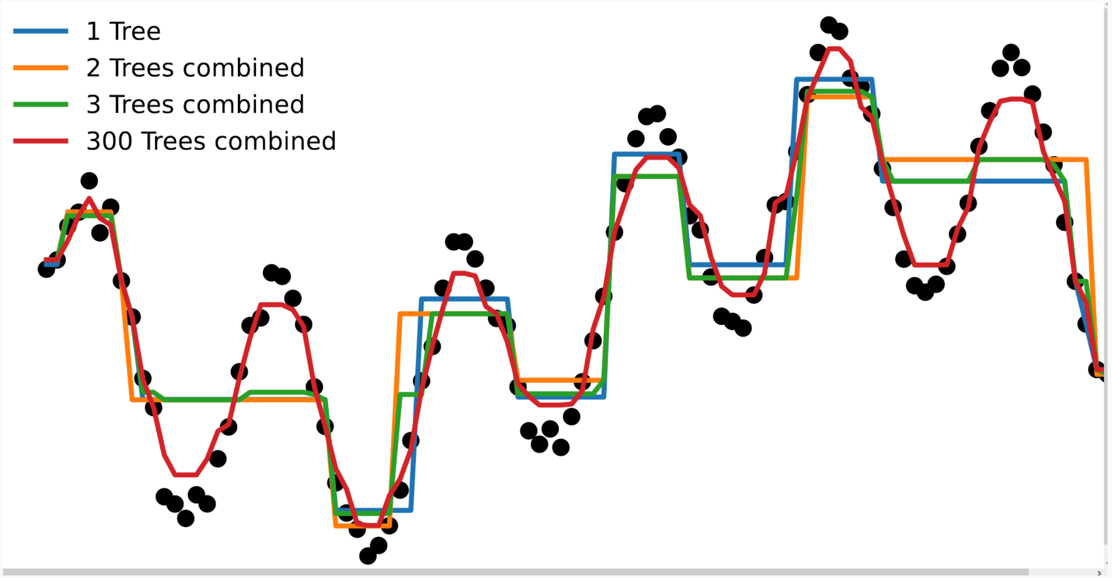
What is the cardio-vacular risk of this patient?

How much will the price of a stock be tomorrow?

Assumption Train and test data are drawn from the same distribution.

Machine learning is pattern matching

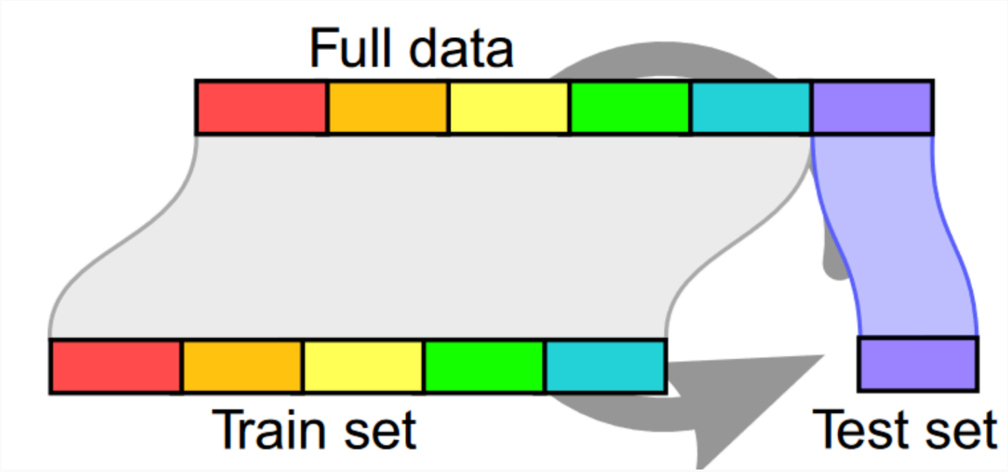
Find an estimator $f : x \to y$ that approximates the true value of y so that $f(x) \approx y$



Boosted trees : iterative ensemble of decision trees

Machine learning is pattern matching that generalizes to new data

Select models based on their ability to generalize to new data : (train, test) splits and cross validation (Stone, 1974).



“Cross validation” (Varoquaux et al., 2017)

Machine learning is great for prediction on complex data

Images: Image classification with deep convolutional neural networks (Krizhevsky et al., 2012)

ImageNet 1K: 1.5 million images, 1000 classes

Machine learning is great for prediction on complex data

Images: Image classification with deep convolutional neural networks (Krizhevsky et al., 2012)

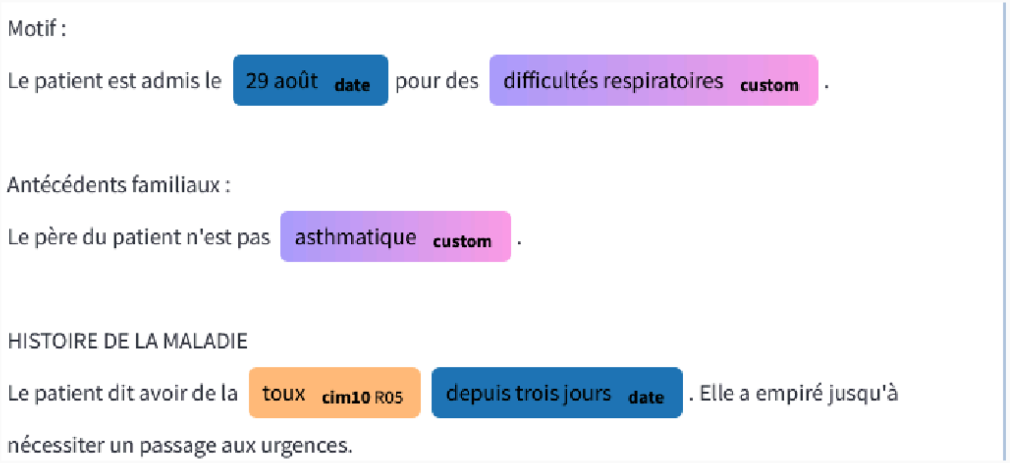
Speech-to-text: Towards end-to-end speech recognition with recurrent neural networks (Graves & Jaitly, 2014)

Machine learning is great for prediction on complex data

Images: Image classification with deep convolutional neural networks (Krizhevsky et al., 2012)

Speech-to-text: Towards end-to-end speech recognition with recurrent neural networks (Graves & Jaitly, 2014)

Text: Attention is all you need (Vaswani, 2017)



Named entity recognition

Machine learning is not driven by causal mechanisms

- For example people that go to the hospital die more than people who do not[1]:
- Naive data analysis might conclude that hospitals are bad for health.

[1] Example from https://inria.github.io/scikit-learn-mooc/concluding_remarks.html?highlight=causality

Machine learning is not driven by causal mechanisms

- For example people that go to the hospital die more than people who do not[1]:
- Naive data analysis might conclude that hospitals are bad for health.
- The fallacy is that we are comparing different populations: people who go to the hospital typically have a worse baseline health than people who do not.

Definition: Confounding factor

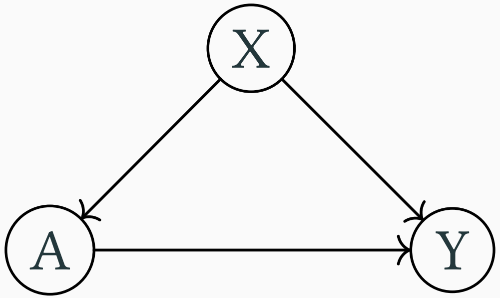
A variable that influences both the treatment and the outcome.

[1] Example from https://inria.github.io/scikit-learn-mooc/concluding_remarks.html?highlight=causality

Why is prediction different from causation? (2/2)

Causal inference (most part of economists) : What would happen if we changed the system ie. under an intervention?

Causal inference (most part of economists) : What would happen if we changed the system ie. under an intervention?



Causal inference models $(X, A, Y(A=1), Y(A=0))$ ie. the covariate shift between treated and control units.

Assumption: No counfounders

No unmeasured variables influencing both treatment and outcome.

Illustration of the fundamental problem of causal inference (epidemiology)

Population: patients experiencing a stroke

Illustration of the fundamental problem of causal inference (epidemiology)

Population: patients experiencing a stroke

Intervention $A = 1$: Patients had access to a MRI scan in less than 3 hours after the first symptoms

Comparator $A = 0$: Patients had access to a MRI scan in more than 3 hours after the first symptoms

Illustration of the fundamental problem of causal inference (epidemiology)

Population: patients experiencing a stroke

Intervention $A = 1$: Patients had access to a MRI scan in less than 3 hours after the first symptoms

Comparator $A = 0$: Patients had access to a MRI scan in more than 3 hours after the first symptoms

$Y = \mathbb{P}[\text{Mortality}]$: **the mortality at 7 days**

Illustration of the fundamental problem of causal inference (epidemiology)

Population: patients experiencing a stroke

Intervention $A = 1$: Patients had access to a MRI scan in less than 3 hours after the first symptoms

Comparator $A = 0$: Patients had access to a MRI scan in more than 3 hours after the first symptoms

$Y = \mathbb{P}[\text{Mortality}]$: **the mortality at 7 days**

$X = \mathbb{P}[\text{Charlson score}]$: **a comorbidity index summarizing the overall health state of the patient. Higher is bad for the patient.**

Illustration of the fundamental problem of causal inference (epidemiology)

Population: patients experiencing a stroke

Intervention $A = 1$: Patients had access to a MRI scan in less than 3 hours after the first symptoms

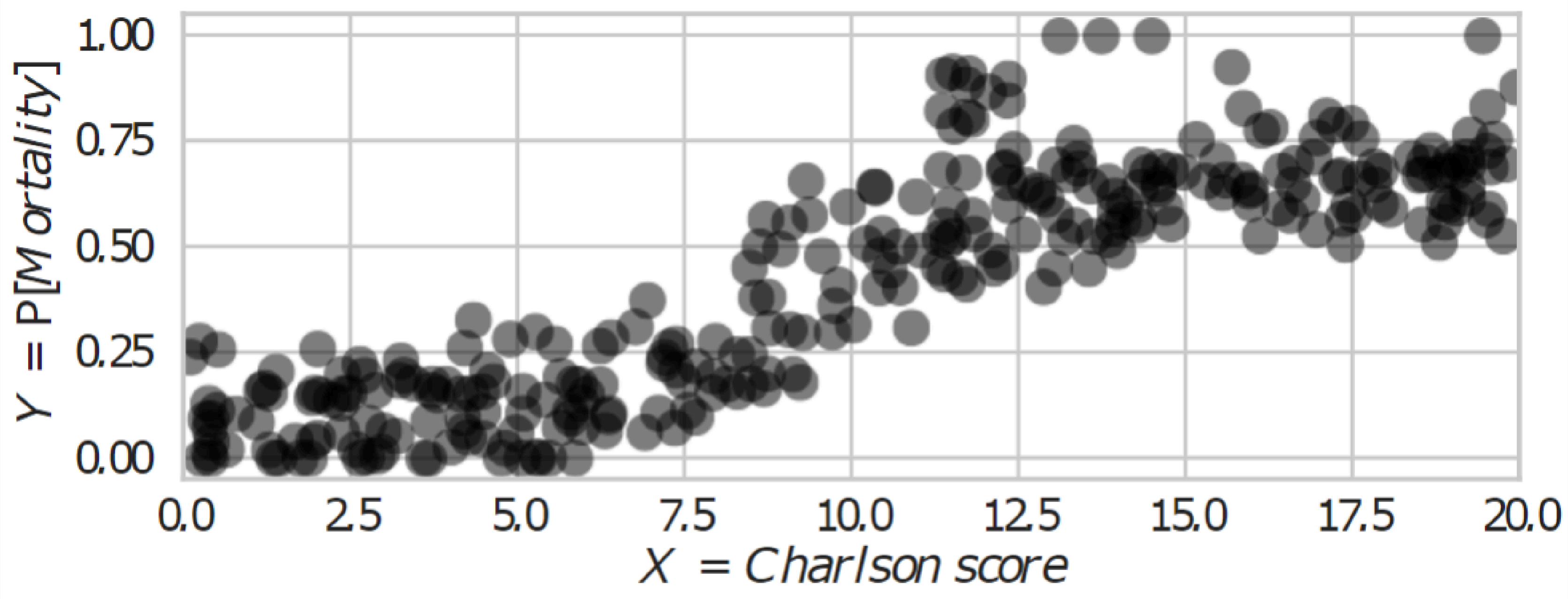
Comparator $A = 0$: Patients had access to a MRI scan in more than 3 hours after the first symptoms

$Y = \mathbb{P}[\text{Mortality}]$**: the mortality at 7 days**

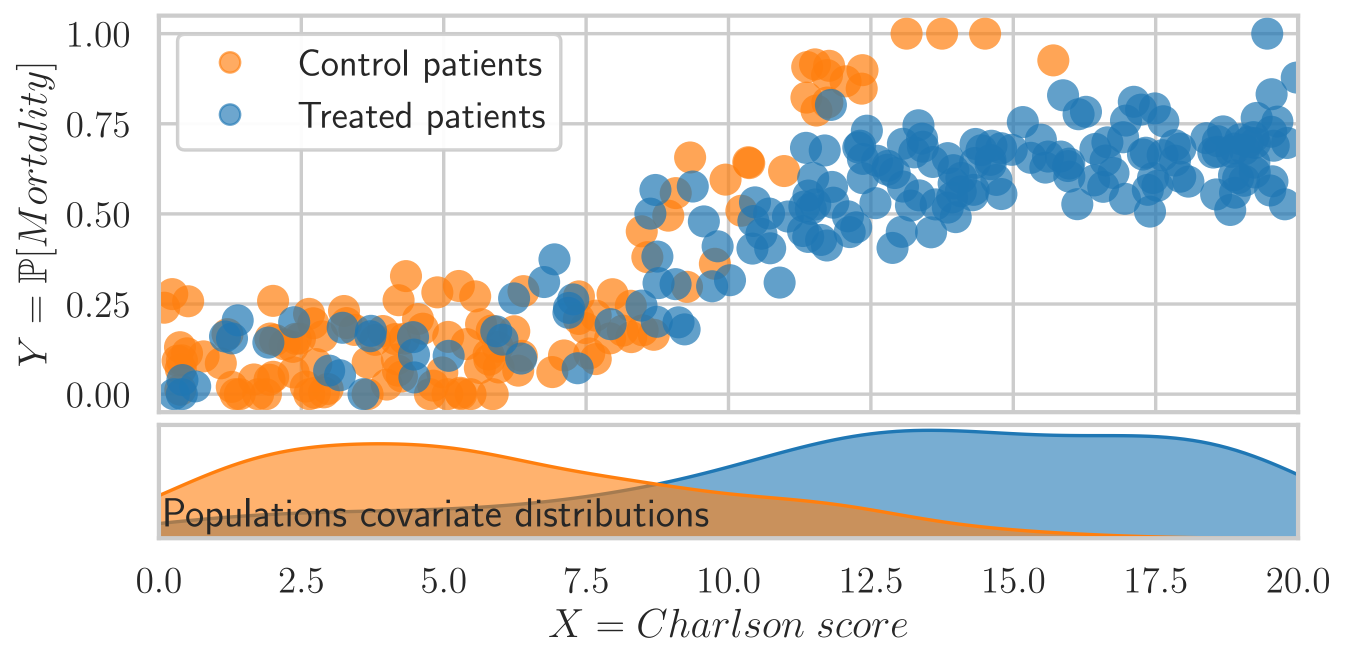
$X = \mathbb{P}[\text{Charlson score}]$**: a comorbidity index summarizing the overall health state of the patient. Higher is bad for the patient.**

❓ What is the effect of early access to MRI on the mortality at 7 days for stroke patients?

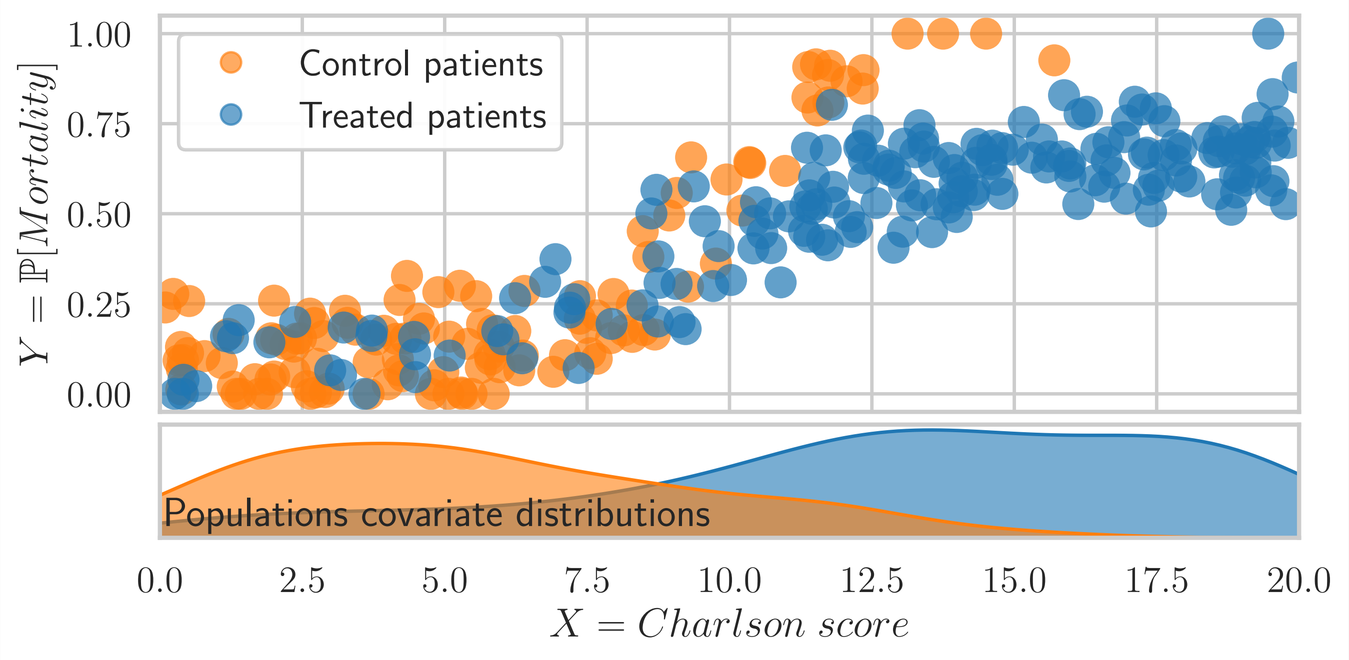
Draw a population sample without treatment status



Draw a population sample with treatment status



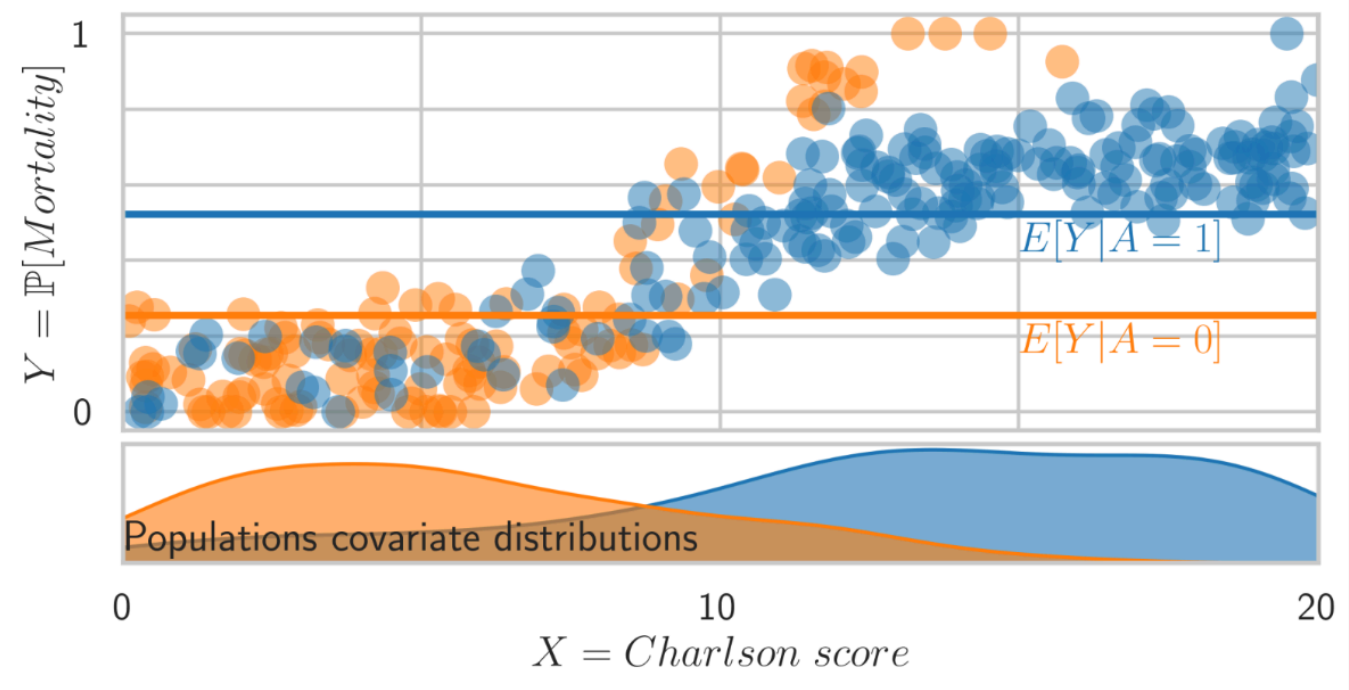
Draw a population sample with treatment status



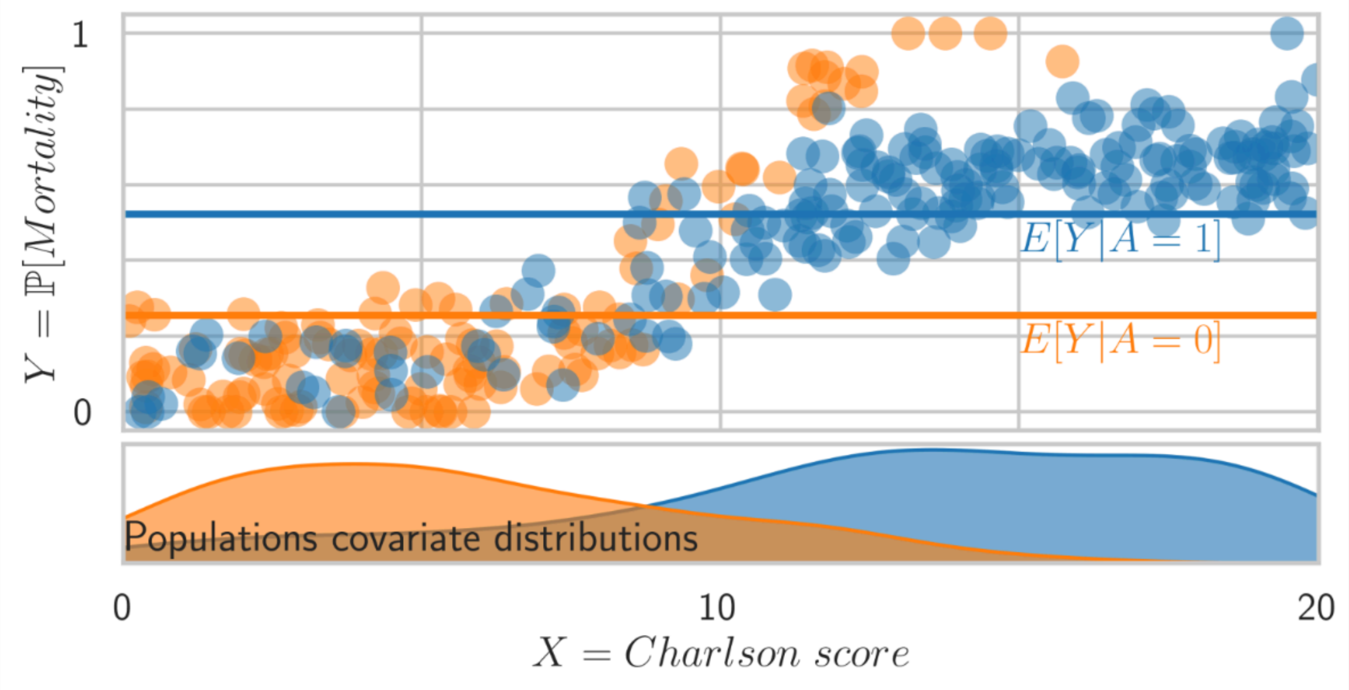
👀 Patient with higher risks have early access to MRI.

Compute the difference in mean (DM): $\tau_{\mathrm{DM}} = \mathbb{E}[Y(1)]$ - $\mathbb{E}[Y(0)]$

Compute the difference in mean (DM): $\tau_{\mathrm{DM}} = \mathbb{E}[Y(1)]$ - $\mathbb{E}[Y(0)]$



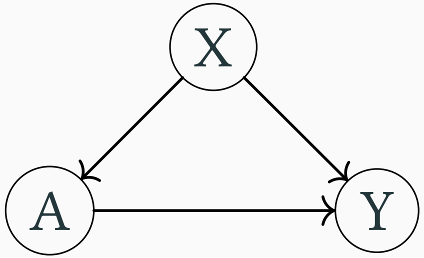
Compute the difference in mean (DM): $\tau_{\text{DM}} = \mathbb{E}[Y(1)]$ - $\mathbb{E}[Y(0)]$



(False) conclusion: early access to MRI is associated with a higher mortality at 7 days.

RCT case: No problem of confounding

Observational data

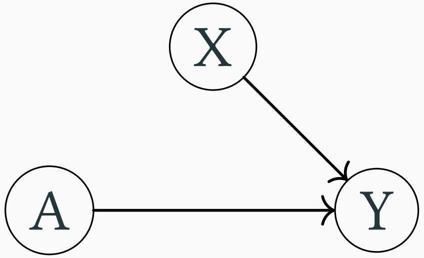


$$Y(1), Y(0) \not\perp\!\!\!\perp A$$

Intervention is not random

(with respect to the confounders)

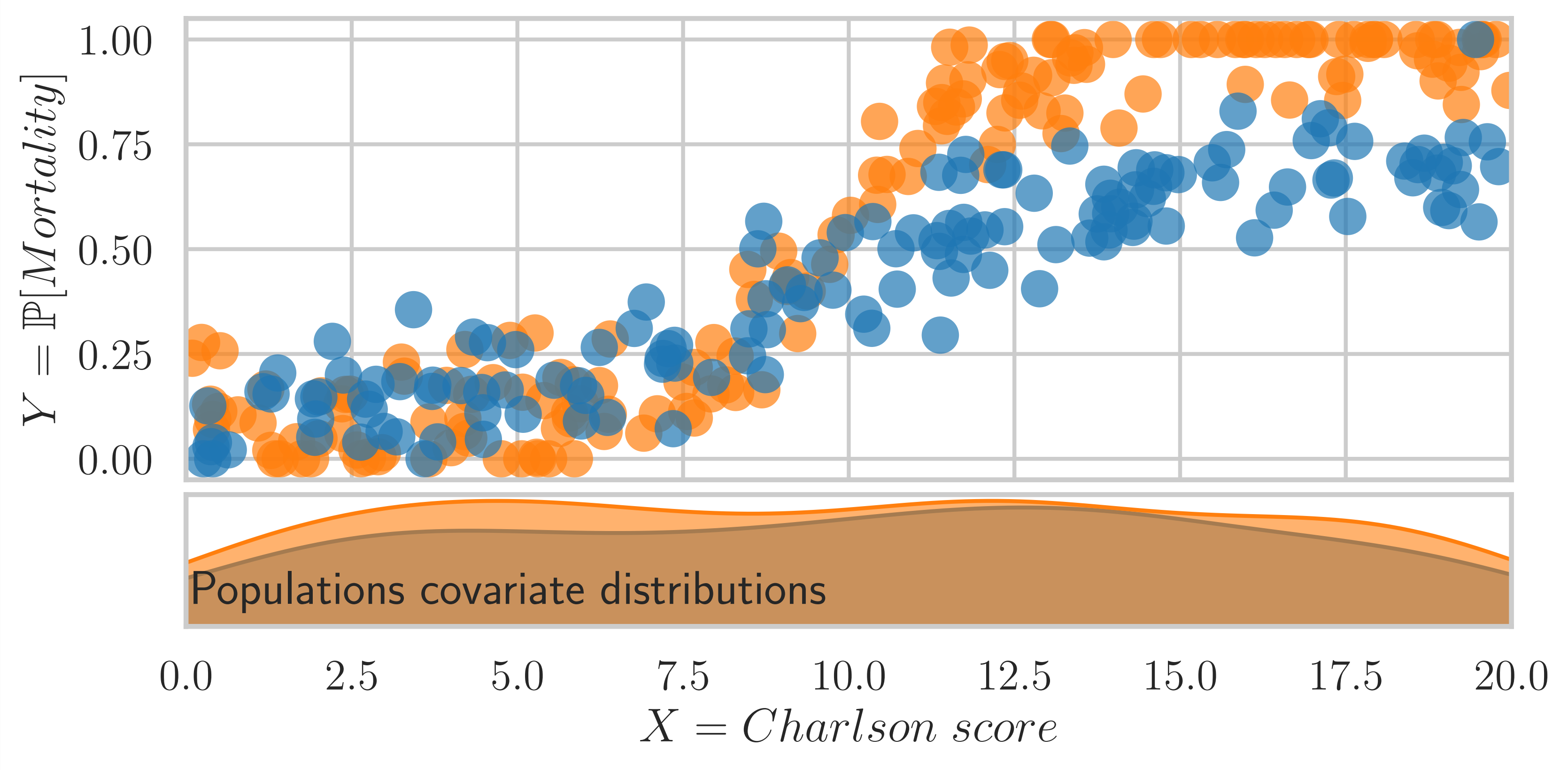
RCT data



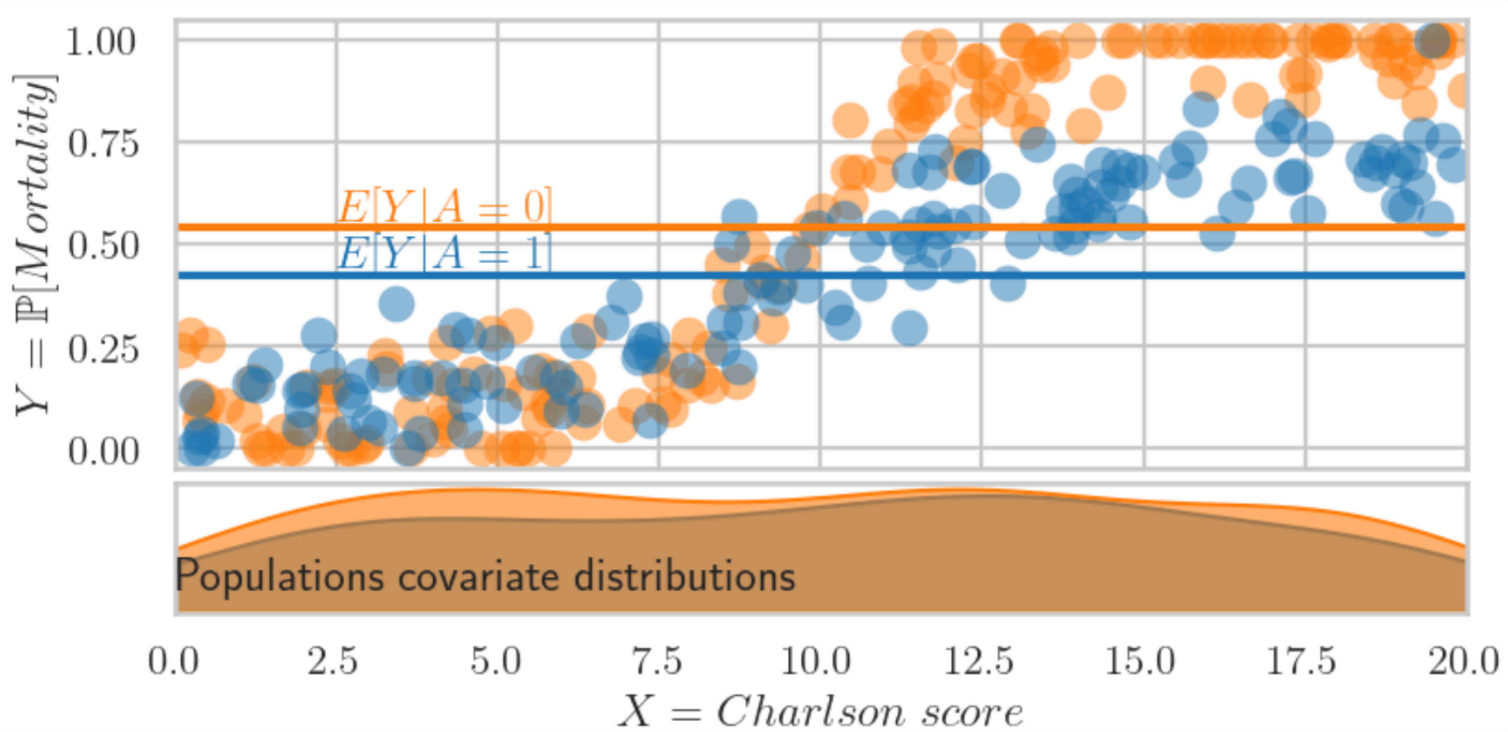
$$Y(1), Y(0) \perp\!\!\!\perp A$$

Force random assignment of the intervention

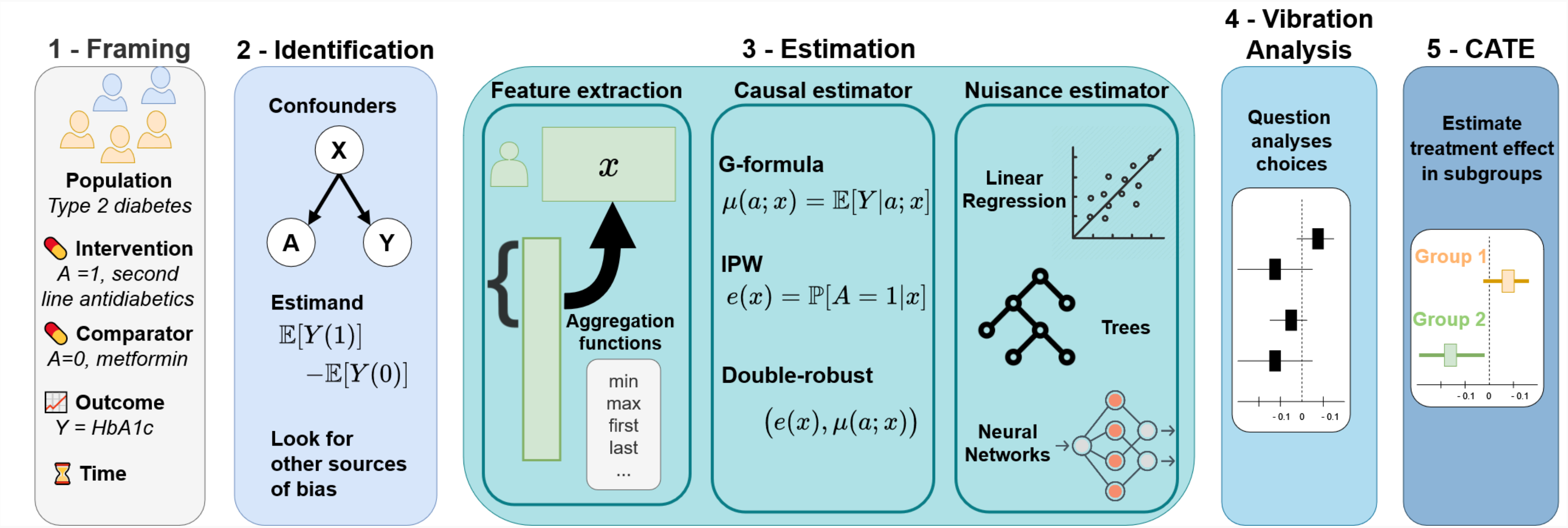
Illustration: RCT data



Compute the difference in mean (DM): $\tau_{\mathrm{DM}} = \mathbb{E}[Y(1)]$ - $\mathbb{E}[Y(0)]$

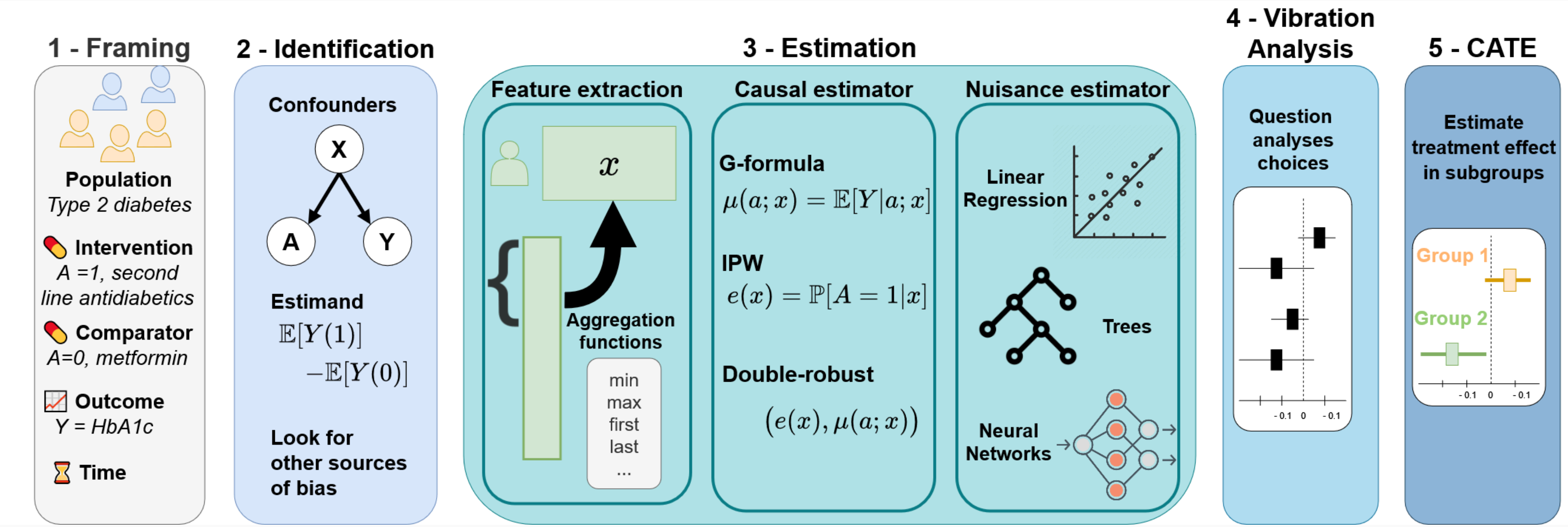


Causal inference: framing with PICO and identification with DAGs



Five steps for causal inference: an applied tutorial (Doutreligne et al., 2025, Figure 1)

Causal inference: framing with PICO and identification with DAGs



Five steps for causal inference: an applied tutorial (Doutreligne et al., 2025, Figure 1)

Further references:

- Gentle introduction from ML and epidemiologists (Abécassis et al., 2024)
- Formal statistical point of view (Wager, 2024)

Framing: How to ask a sound causal question

PICO framework (Richardson et al., 1995)

Originally designed for clinical research. It is a structured approach to formulate a research question. Critical for health technology assessment (eg. Haute Autorité de santé).

PICO stands for

- Population : Who are we interested in?
- Intervention : What treatment/intervention do we study?
- Comparison : What are we comparing it to?
- Outcome : What are we interested in?

PICO framework (Richardson et al., 1995)

PICO framework (Richardson et al., 1995)

Originally designed for clinical research. It is a structured approach to formulate a research question. Critical for health technology assessment (eg. Haute Autorité de santé).

PICO stands for

- Population : Who are we interested in?
- Intervention : What treatment/intervention do we study?
- Comparison : What are we comparing it to?
- Outcome : What are we interested in?

Example with the job dataset (LaLonde, 1986)

Built to evaluate the impact of the National Supported Work (NSW) program. The NSW is a transitional, subsidized work experience program targeted towards people with longstanding employment problems.

The PICO framework

Component	Description	Example
Population	What is the target population of interest?	People with longstanding employment problems
Intervention	What is the intervention?	On-the-job training lasting between nine months and a year
Control	What is the relevant comparator?	No training
Outcome	What are the outcomes?	Earnings in 1978
Time	Is the start of follow-up aligned with intervention assignment?	The period of follow-up for the earning is the year after the intervention.

PICO: other examples in econometrics

The Oregon Health Insurance Experiment (Finkelstein et al., 2012) : A randomized experiment by lottery assessing the impact of Medicaid on low-income adults in Oregon.

- P: Low-income adults in Oregon
- I: Medicaid
- C: No insurance
- O: Healthcare uses and expenditures, health outcomes

PICO: other examples in econometrics

The economic impact of climate change on US agricultural land. (Deschênes & Greenstone, 2007): difference-in-differences design assessing the impact of climate change on agricultural profits.

- P: US agricultural land
- I: Climate change
- C: No climate change
- O: Agricultural profits

PICO: other examples in econometrics

The impact of class size on test scores. (Angrist & Lavy, 1999): regression discontinuity design.

- P: Fourth and fifth grades school in Israel
- I: Class size increases by one unit
- C: No class size increase
- O: Test scores (math and reading)

Identification: List necessary information to answer the causal question

Identification: Build the causal model

> A causal effect is said to be identified if it is possible, with ideal data (infinite sample size and no measurement error), to purge an observed association of all noncausal components such that only the causal effect of interest remains.
>
> — (Elwert & Winship, 2014)

Steps

- Potential outcome framework : mathematical tool to reason about causality
- Directed acyclic graphs (DAG) : graphical tool to reason about causality
- Causal estimand : what is the targeted quantity?

Potential outcomes, (Neyman, 1923; Rubin, 1974)

The Neyman-Rubin model, let:

- Y be the outcome,
- A the (binary) treatment

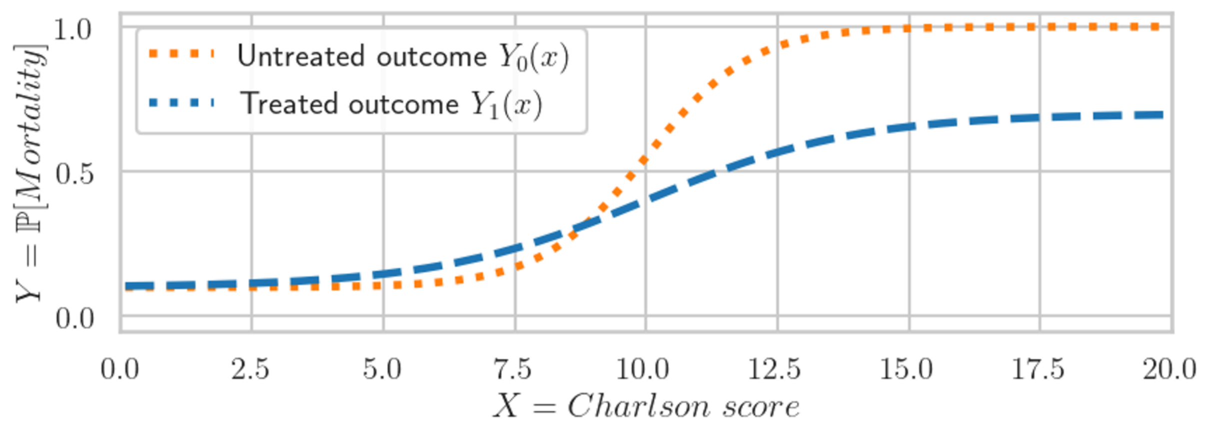
For each individual, we have two potential outcomes: $Y(1)$ and $Y(0)$. But only one is observed, depending on the treatment assignment: $Y(A)$.

Potential outcomes, (Neyman, 1923; Rubin, 1974)

The Neyman-Rubin model, let:

- Y be the outcome,
- A the (binary) treatment

For each individual, we have two potential outcomes: $Y(1)$ and $Y(0)$. But only one is observed, depending on the treatment assignment: $Y(A)$.

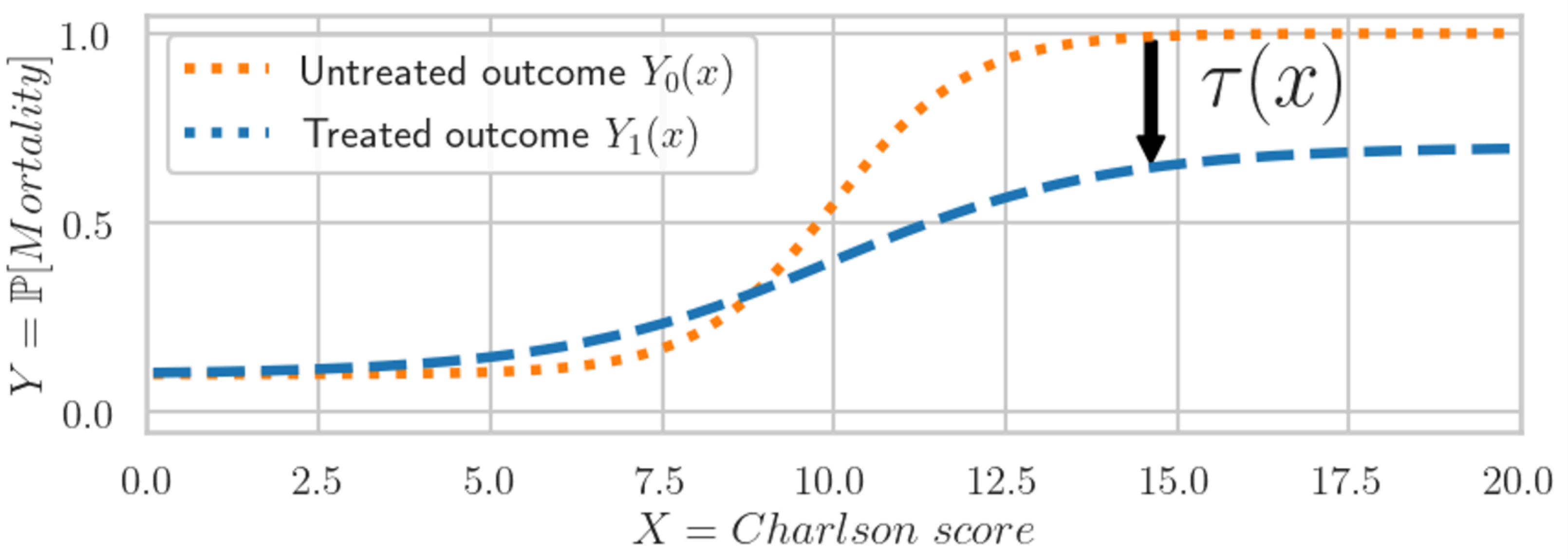


Causal estimand: What is the targeted quantity (with potential outcomes)?

- Average treatment effect (ATE): $\mathbb{E}[Y(1) - Y(0)]$
- Conditional average treatment effect (CATE): $\mathbb{E}[Y(1) - Y(0) \mid X]$

Causal estimand: What is the targeted quantity (with potential outcomes)?

- Average treatment effect (ATE): $\mathbb{E}[Y(1) - Y(0)]$
- Conditional average treatment effect (CATE): $\mathbb{E}[Y(1) - Y(0) \mid X]$



Other estimands

- Average treatment effect on the treated (ATT): $\mathbb{E}[Y(1) - Y(0) \mid A = 1]$
- Conditional average treatment effect on the treated (CATT):

$$\mathbb{E}[Y(1) - Y(0) \mid A = 1, X]$$

Other estimands

- Average treatment effect on the treated (ATT): $\mathbb{E}[Y(1) - Y(0) \mid A = 1]$
- Conditional average treatment effect on the treated (CATT):

$$\mathbb{E}[Y(1) - Y(0) \mid A = 1, X]$$

Other estimands more used in epidemiology

- Risk ratio (RR): $\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]}$
- Odd ratio (OR) for binary outcome: $\left(\frac{\mathbb{P}[Y(1)=1]}{\mathbb{P}[Y(1)=0]}\right) / \left(\frac{\mathbb{P}[Y(0)=1]}{\mathbb{P}[Y(0)=0]}\right)$

See (Colnet et al., 2023) for a review of the different estimands and the impact on generalization.

PICO framework, link to the potential outcomes

Component	Description	Notation	Example
Population	What is the target population of interest?	$X \sim P(X)$	People with longstanding employment problems
Intervention	What is the intervention?	$A \sim P(A=1) = p_A$	On-the-job training lasting between nine months and a year
Control	What is the relevant comparator?	$1 - A \sim 1 - p_A$	No training
Outcome	What are the outcomes?	$Y(1), Y(0) \sim P(Y(1), Y(0))$	Earnings in 1978
Time	Is the start of follow-up aligned with intervention assignment?	N/A	The period of follow-up for the earning is the year after the intervention

What can we learn from the data?

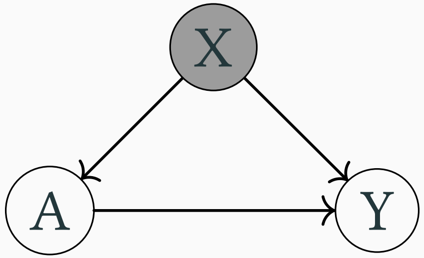
Four assumptions, referred as strong ignorability

Required for identifiability of the causal estimands with observational data (Rubin, 2005)

Assumption 1: Unconfoundedness, also called ignorability

Treatment assignment is as good as random given the covariates X

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$$



Assumption 1: Unconfoundedness, also called ignorability

Treatment assignment as good as random given the covariates X

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$$

- Equivalent to conditional independence on the propensity score: $e(X) \stackrel{\text{def}}{=} \mathbb{P}(A = 1|X)$ (Rosenbaum & Rubin, 1983):

$$\{Y(1), Y(0)\} A \mid e(X)$$

Assumption 1: Unconfoundedness, also called ignorability

Treatment assignment as good as random given the covariates X

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$$

- Equivalent to conditional independence on the propensity score: $e(X) \stackrel{\text{def}}{=} \mathbb{P}(A = 1|X)$ (Rosenbaum & Rubin, 1983):

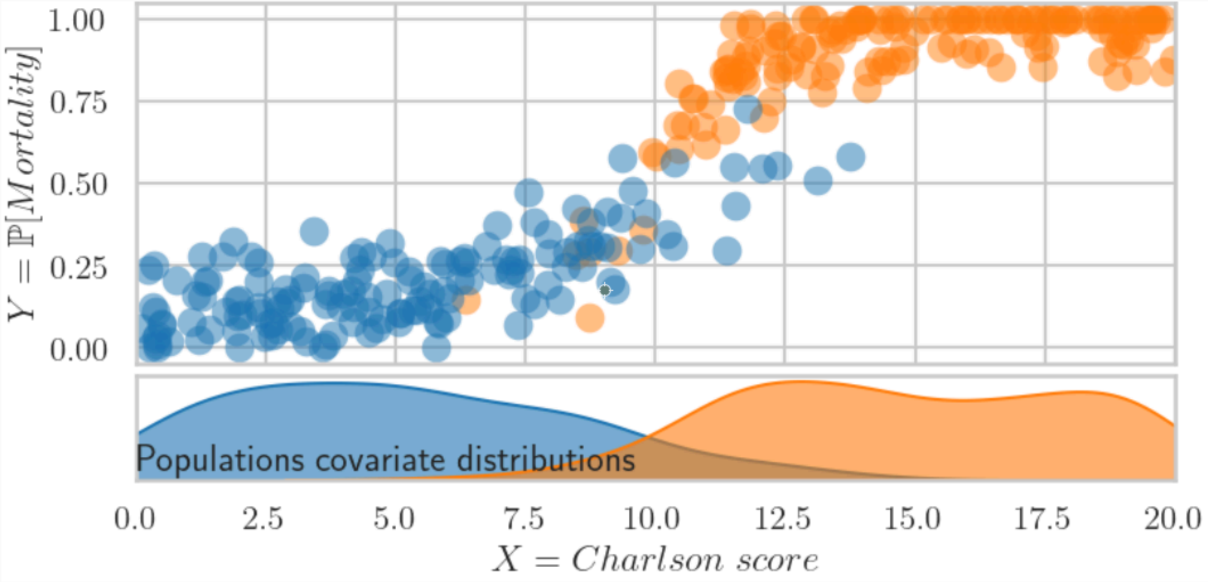
$$\{Y(1), Y(0)\} A \mid e(X)$$

- Knowledge based ie. cannot be validated with data
  - Because of possibly unmeasured confounders
  - In practice: ask yourself if you have measured all the relevant variables that could influence both the treatment and the outcome.

Assumption 2: Overlap, also known as positivity

The treatment is not deterministic given X

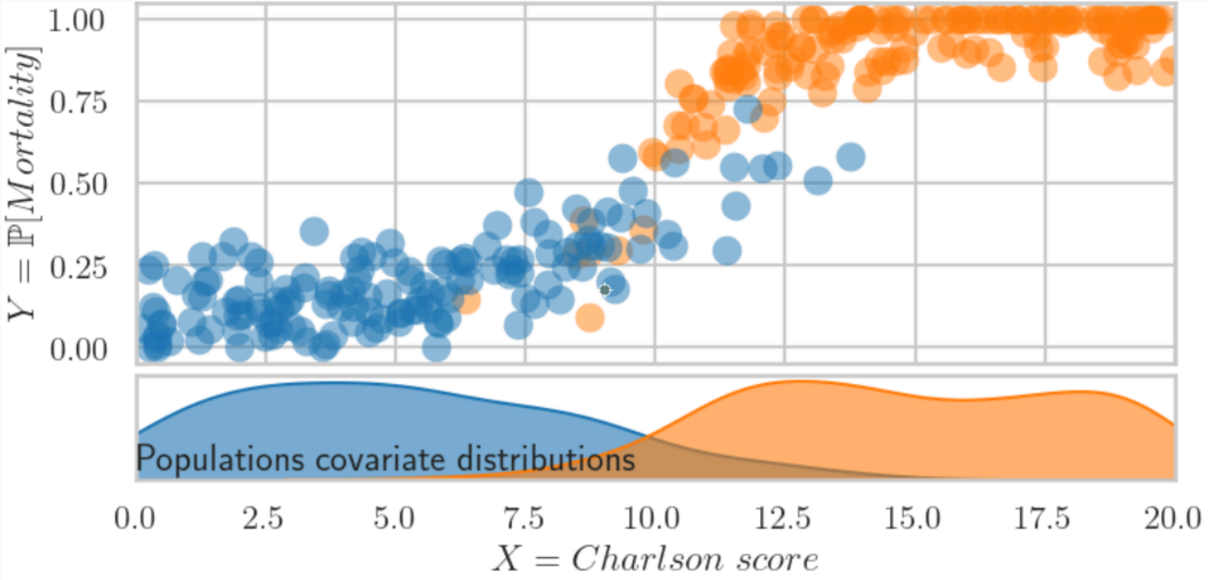
$$\eta < e(x) < 1 - \eta \text{ with } e(X) \stackrel{\text{def}}{=} \mathbb{P}(A = 1|X)$$



Assumption 2: Overlap, also known as positivity

The treatment is not deterministic given X

$$\eta < e(x) < 1 - \eta \text{ with } e(X) \overset{\text{def}}{=} \mathbb{P}(A = 1|X)$$



Choice of covariates X: trade-off between ignorability and overlap (D'Amour et al., 2021)

Consistency, also called Stable Unit Treatment Values (SUTVA)

The observed outcome is the potential outcome of the assigned treatment for each unit i.

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

Assumption 3 and 4: Consistency and generalization

Consistency, also called Stable Unit Treatment Values (SUTVA)

The observed outcome is the potential outcome of the assigned treatment for each unit i.

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

- The intervention A is well defined (Hernan & Robins, 2020)
- There is no interfrence ie. network effect

Consistency, also called Stable Unit Treatment Values (SUTVA)

The observed outcome is the potential outcome of the assigned treatment for each unit i.

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

- The intervention A is well defined (Hernan & Robins, 2020)
- There is no interfrence ie. network effect

Generalization, also called no-covariate shift

Training and test data are are drawn from the same distribution

Directed acyclic graphs (DAGs)

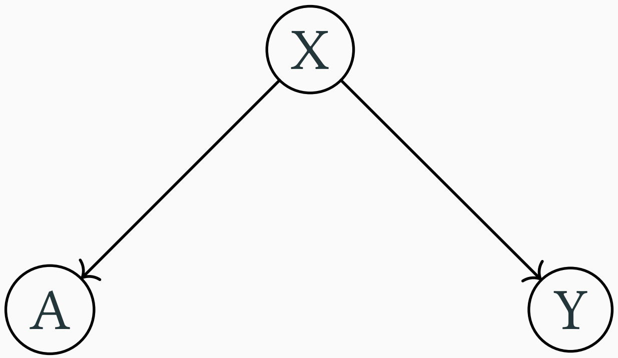
DAGs encode the causal structure of the data generating process

Introduced by (Pearl, 1995), (Pearl & others, 2000). Good practical overview in (VanderWeele, 2019).

Motivation

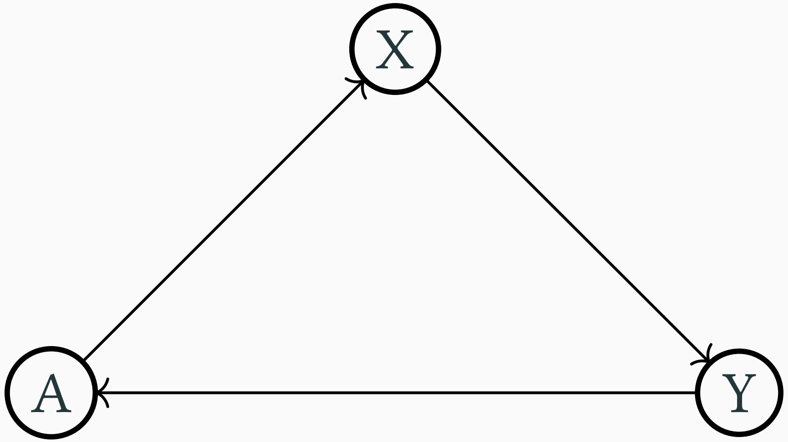
- Reason about the relation between variables.
- Help identify for which (minimal) set of variables, the ATE is identifiable.

Directed acyclic graphs (DAG), definitions



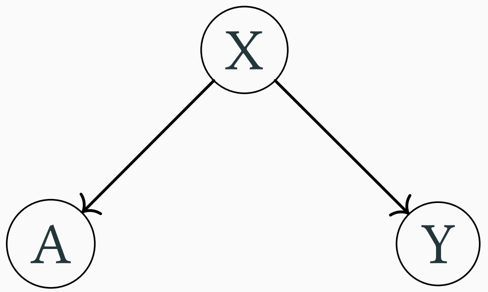
- Graph: A set of relations between nodes described by edges between those nodes.
- Directed: Edges between nodes have direction. The direction of the arrow represents a cause-effect relationship.
- Acyclic: : There are no cycles or loops in the causal structure. A variable can't be a cause of itself.

⚠️ This is not a DAG



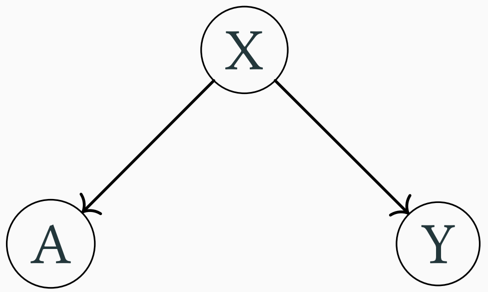
DAGs: nodes

- Nodes represent random variables.
- Edges between nodes symbolize causal effects (i.e. difference in potential outcomes). Here, $Y_i(a) \neq Y_i(a')$ for two different levels of A_i because of the arrow from A to Y.
- Lack of edges between nodes denotes the absence of a causal relationships.



DAGs: nodes

- Nodes represent random variables.
- Edges between nodes symbolize causal effects (i.e. difference in potential outcomes). Here, $Y_i(a) \neq Y_i(a')$ for two different levels of A_i because of the arrow from A to Y.
- Lack of edges between nodes denotes the absence of a causal relationships.



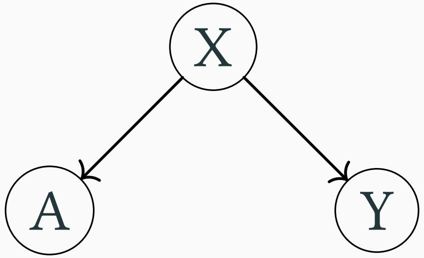
Important

Not drawing an arrow makes a stronger assumption about the relationship between those two variables than drawing an arrow.

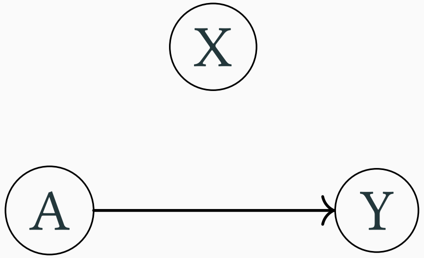
DAGs: paths

- A path between two nodes is a route that connects the two nodes following non-intersecting edges.
- A path exists even if the arrows are not pointing in the good direction.
- Two examples of paths between A and Y:

$$A \leftarrow X \rightarrow Y$$



$$A \rightarrow Y$$



DAGs: causal paths

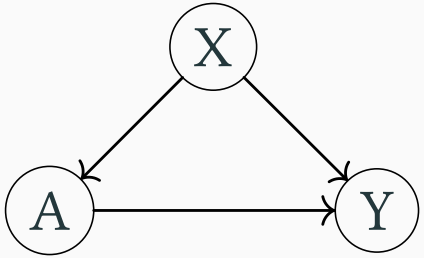
- Paths encode dependencies between random variables: not necessarily causal dependecies, it can be mere associations.

DAGs: causal paths

- Paths encode dependencies between random variables: not necessarily causal dependecies, it can be mere associations.
- We distinguish:
  - Causal paths: arrows are all in the same direction.
  - From Non-causal paths: arrows pointing in different directions
- When there is a causal path between two variables A and B, we say that B is a descendant of A (it is causally impacted by A)

DAGs: causal paths

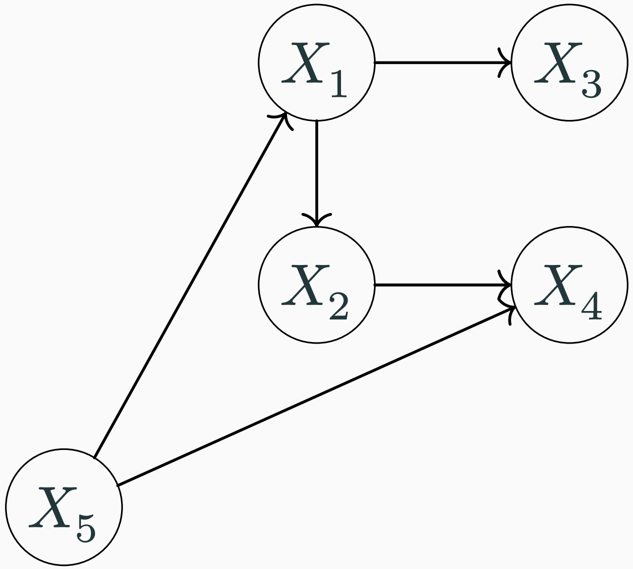
- Paths encode dependencies between random variables: not necessarily causal dependecies, it can be mere associations.
- We distinguish:
  - Causal paths: arrows are all in the same direction.
  - From Non-causal paths: arrows pointing in different directions
- When there is a causal path between two variables A and B, we say that B is a descendant of A (it is causally impacted by A)



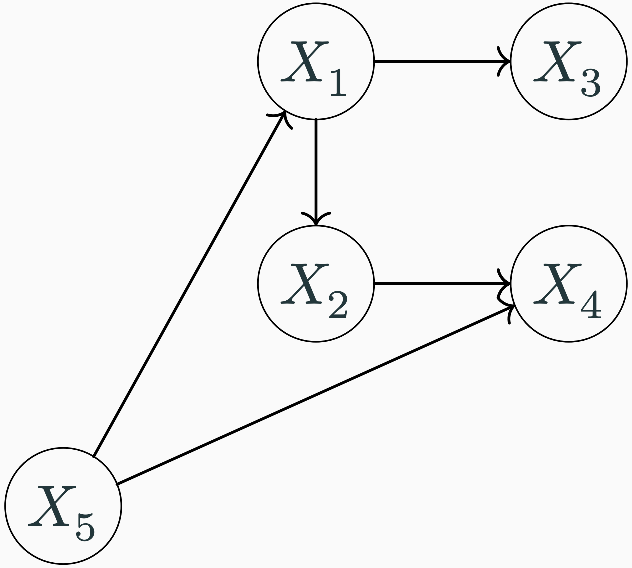
$A \to Y$ is causal

$Y \to A$ is non-causal

Your turn: Paths between X_1 and X_4? Which of them are causal?



Your turn: Paths between X_1 and X_4? Which of them are causal?



- $X_5 \to X_1 \to X_2 \to X_4$
- $X_5 \to X_4$
- $X_1 \leftarrow X_5 \to X_4$: non causal
- $X_3 \leftarrow X_1 \to X_2 \to X_4$: non causal

Three types of directed edges: path

There are three kinds of “triples“ or paths with three nodes: These constitute the most basic building blocks for causal DAGs.

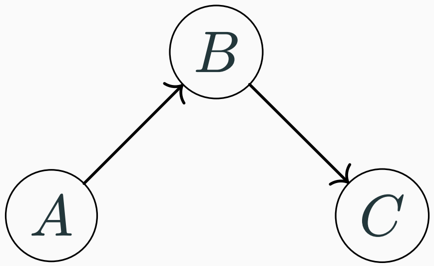
First, a causal path (or chain): $A \to B \to C$

Three types of directed edges: path

There are three kinds of "triples" or paths with three nodes: These constitute the most basic building blocks for causal DAGs.

First, a causal path (or chain): $A \rightarrow B \rightarrow C$

- The effect of A "flows" through B. A and C are not independent and the relationship is causal.

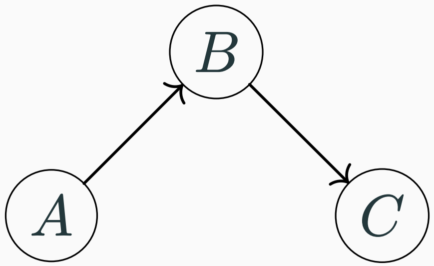


Three types of directed edges: path

There are three kinds of “triples“ or paths with three nodes: These constitute the most basic building blocks for causal DAGs.

First, a causal path (or chain): $A \rightarrow B \rightarrow C$

- The effect of A “flows“ through B. A and C are not independent and the relationship is causal.



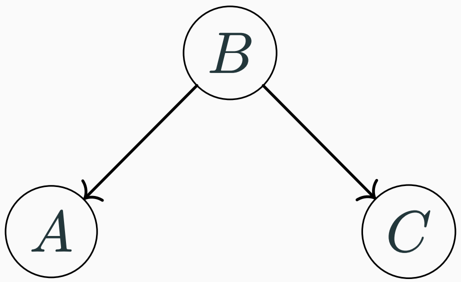
Example

An individual receiving a message (A) encouraging them to vote causes that individual to actually vote (C) only if the individual actually reads (B) the message.

Three types of directed edges: mutual dependence

Second, mutual dependence or fork or confounder: $A \leftarrow B \rightarrow C$

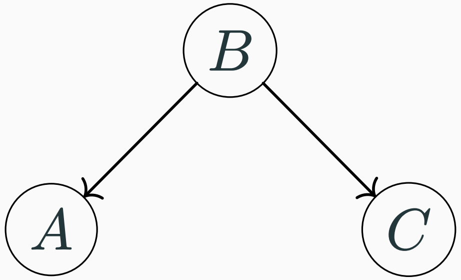
- A and C are not causally related but B is a common cause of both.
- A and C are not independent, but are independent conditional on B.
- Not conditioning on X introduces bias.



Three types of directed edges: mutual dependence

Second, mutual dependence or fork or confounder: $A \leftarrow B \rightarrow C$

- A and C are not causally related but B is a common cause of both.
- A and C are not independent, but are independent conditional on B.
- Not conditioning on X introduces bias.



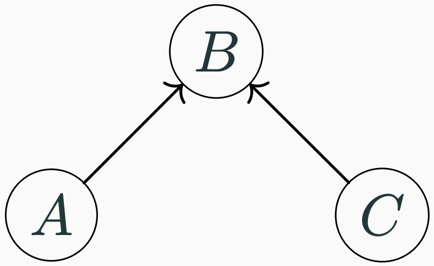
Examples

- Rise in temperature (B) causes both the thermometer (A) to change, and ice to melt (C), but the thermometer changing does not cause ice to melt.
- A is prostate cancer; B is age; and C is Alzheimer's disease.

Three types of directed edges: collider

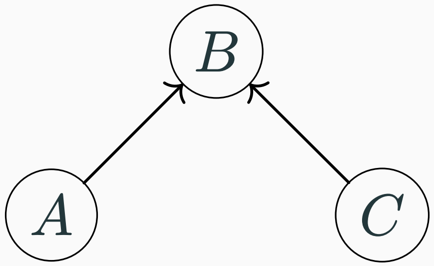
Third, collider: $A \rightarrow B \leftarrow C$

- A and C are both common causes of B : they collide into B.
- A and C are independent, but conditionnaly dependent given B.
- Conditionning on B introduces a spurious correlation between A and C.



Third, collider: $A \rightarrow B \leftarrow C$

- A and C are both common causes of B : they collide into B.
- A and C are independent, but conditionnaly dependent given B.
- Conditionning on B introduces a spurious correlation between A and C.



Examples

- A is result from dice 1, C is results from dice 2, B is sum of dice 1 and dice 2.
- A is height, C is speed, B is whether an athlete plays in the NBA.

Why do we need DAG?

Two data generating processes (DGP) yielding identical data

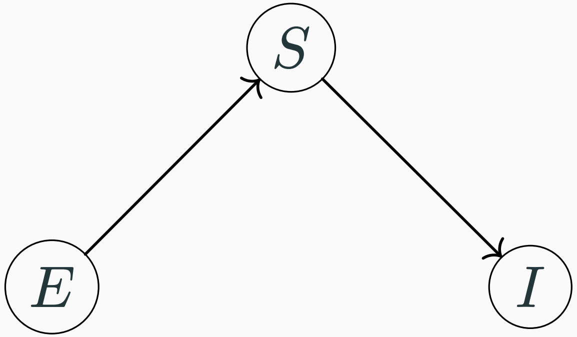
Setup

You are an HR analyst for a large tech company. You have data on three variables for thousands of employees:

- 📚 Education (E): eg. university years.
- 🧠 Skills (S): eg. technical tests.
- 💲 Income (I): current annual income.

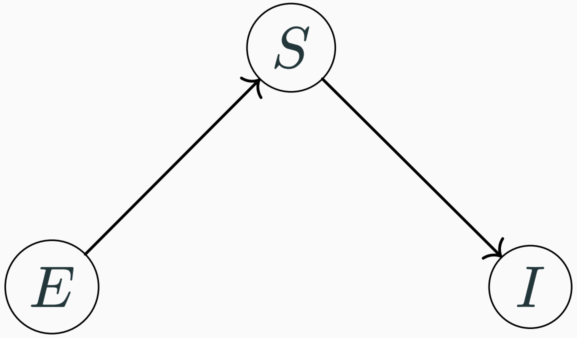
First DGP: education is useful

Is 📚 education improving 🧠 skills thus increasing 💲 income ?



First DGP: education is useful

Is 📚 education improving 🧠 skills thus increasing 💲 income ?



This is a chain

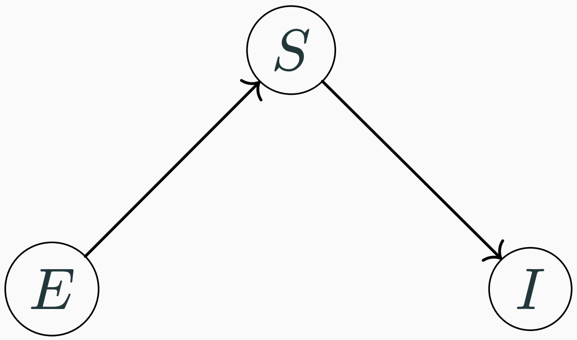
First DGP: education is useful

Is 📚 education improving 🧠 skills thus increasing 💲 income ?

$$E_A = \mathcal{N}(0,1)$$

$$S_A = 0.8E_A + \mathcal{N}(0,0.6^2)$$

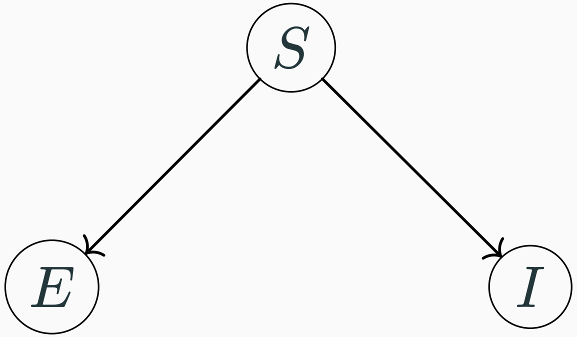
$$I_A = 0.8S_A + \mathcal{N}(0,0.6^2)$$



This is a chain

Second DGP: education is a signal for high skilled workers

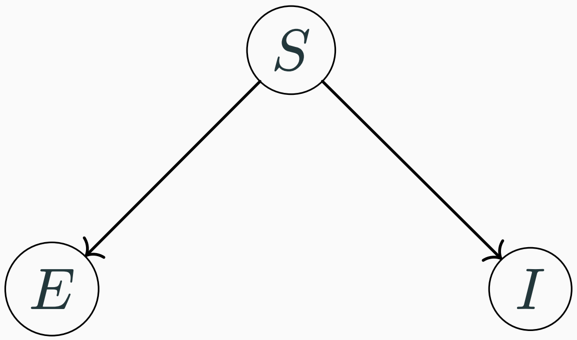
Is 📚 education a signal from people with high 🧠 skills and 💲 high income?



Equations

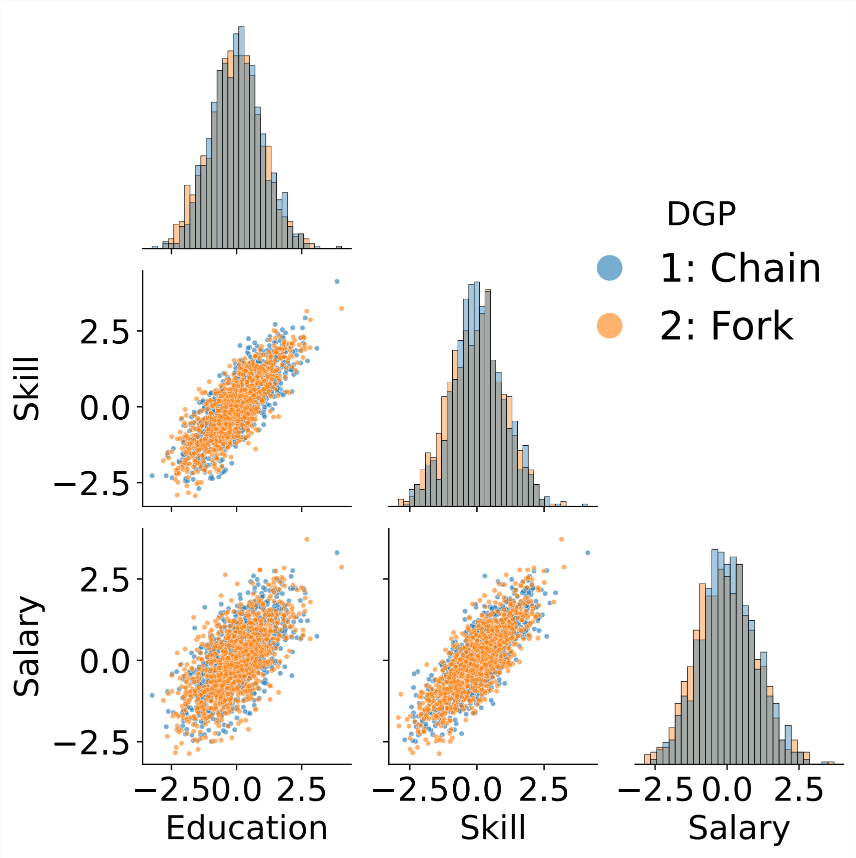
Is 📚 education a signal from people with high 🧠 skills and 💲 high income?

Equations



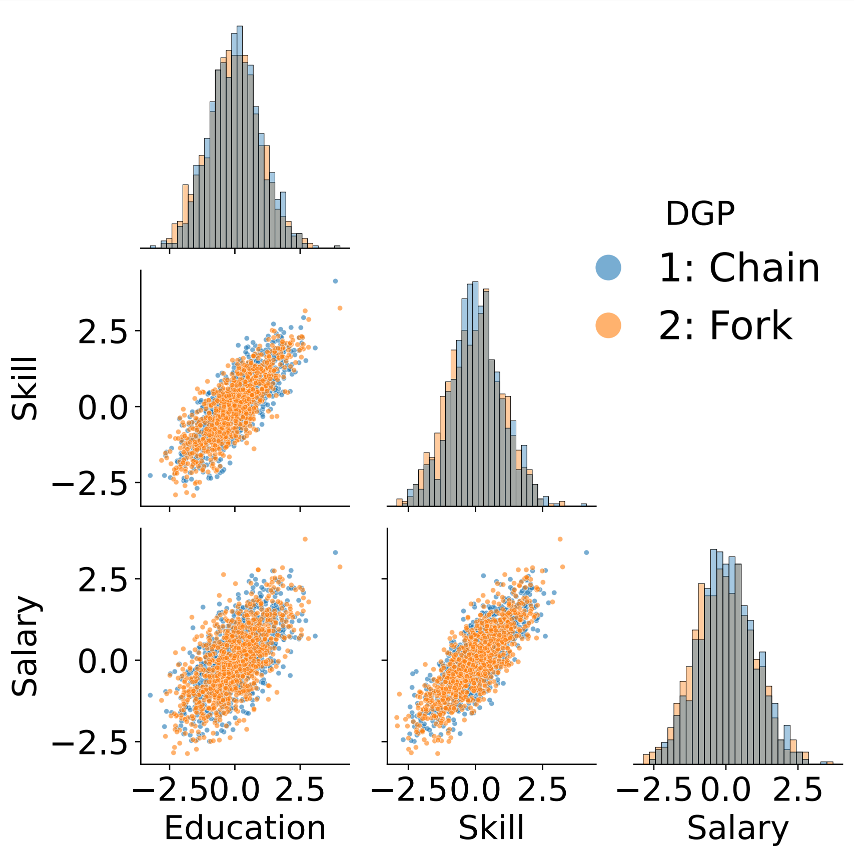
This is a fork

Same observed data, different causal effects



Simulated data from the two DGPs.

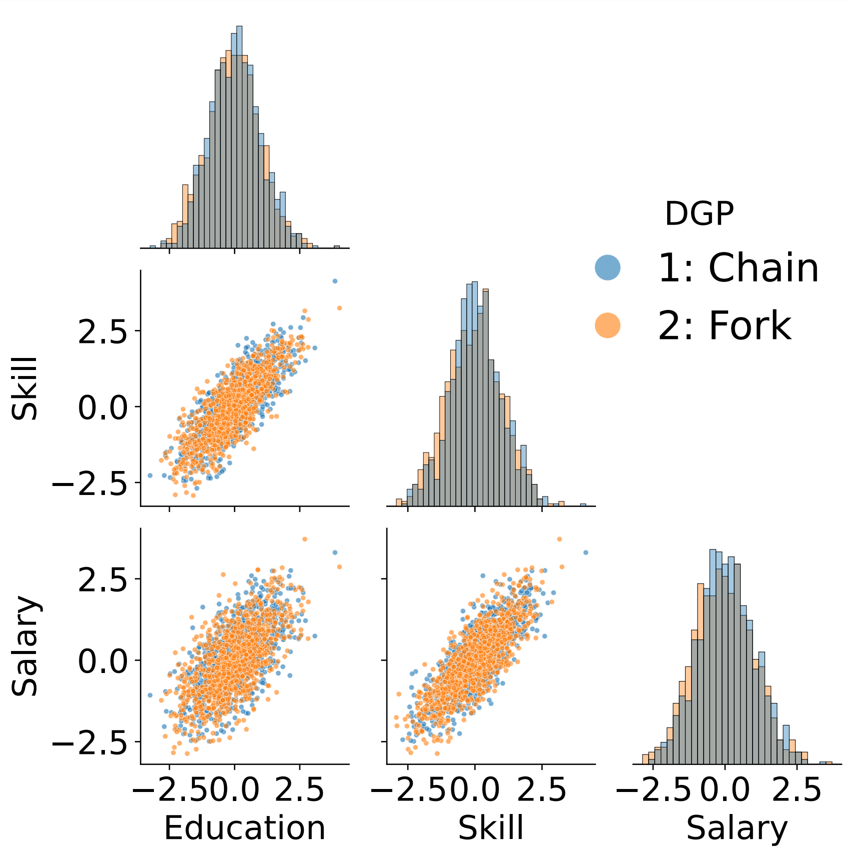
Same observed data, different causal effects



Simulated data from the two DGPs.

Both DGP yield the same data distribution but different causal effects.

Same observed data, different causal effects



Simulated data from the two DGPs.

Both DGP yield the same data distribution but different causal effects.

In the first DGP, increasing education increases skills and income

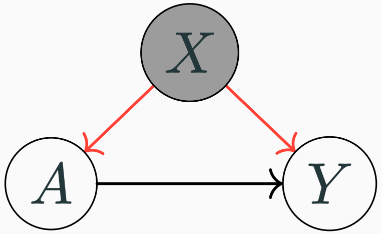
But not in the second !

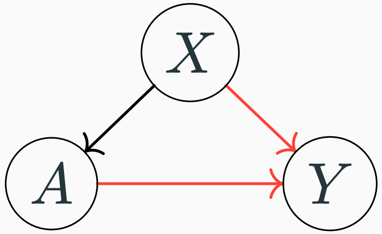
Using DAGs to identify causal effects

Open and blocked paths by conditioning

A path is blocked (or d-separated) if:

- the path contains a non-collider that has been conditioned on.
- or the path contains a collider that has not been conditioned on (and has no descendants that have been conditioned on).

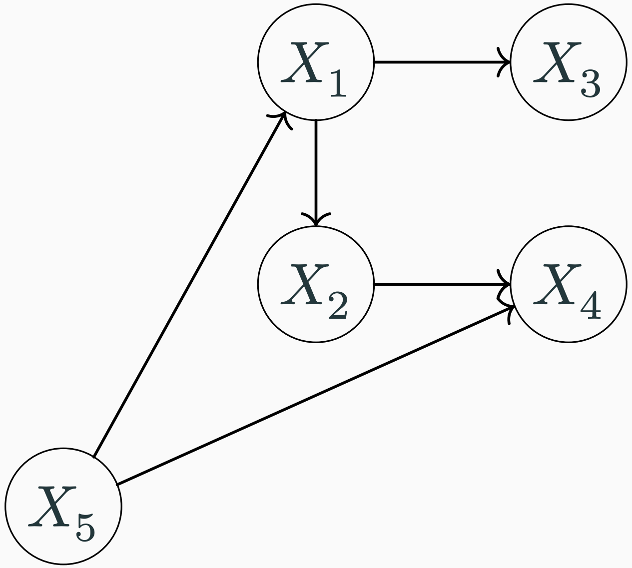




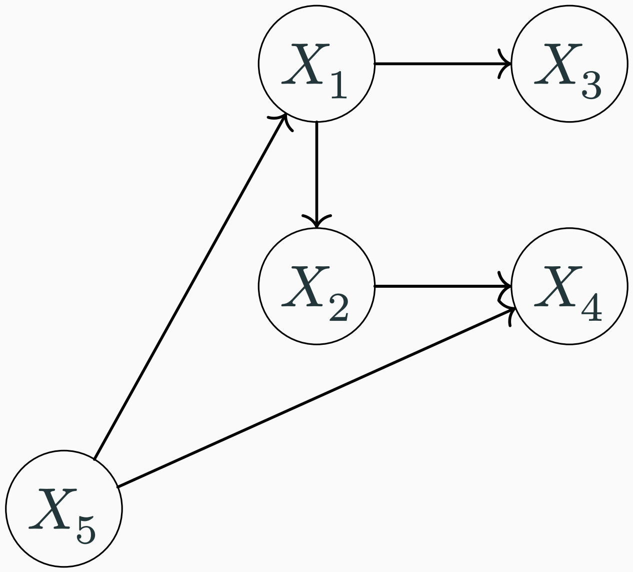
Conditioning on a variable:

- Blocks a path if that variable is not a collider on that path.
- Opens a path if that variable is a collider on that path.

Your turn: Paths from X_5 to X_2? Which of them are opened/blocked?



Your turn: Paths from X_5 to X_2? Which of them are opened/blocked?



$X_5 \rightarrow X_1 \rightarrow X_2$ (blocked by conditionning on X_1)

$X_5 \rightarrow X_1 \rightarrow X_2 \rightarrow X_4$ (opened by conditionning X_4)

Backdoor paths: a special type of paths

Def. Backdoor path from A to Y

Any non-causal path between A and Y that does not include descendants of A.

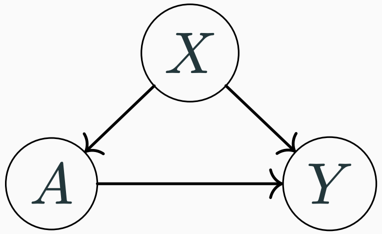
Backdoor paths: a special type of paths

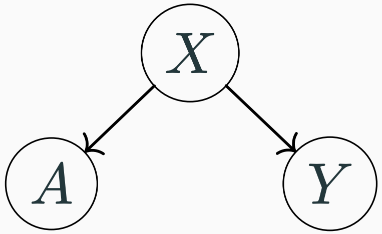
Def. Backdoor path from A to Y

Any non-causal path between A and Y that does not include descendants of A.

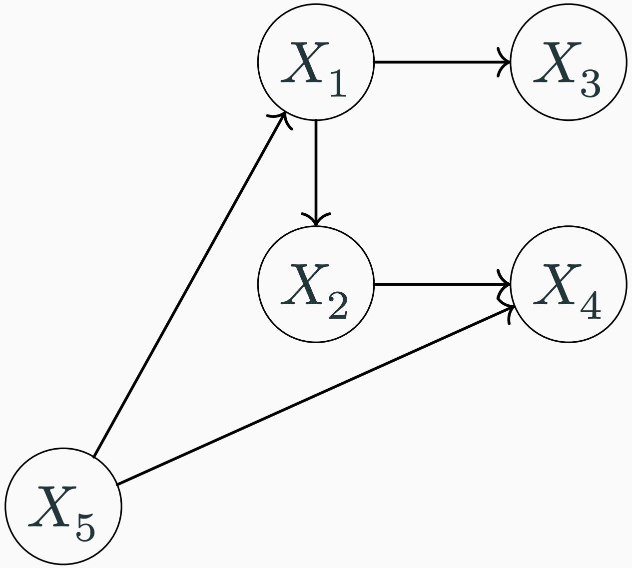
Identifying backdoor paths: Backdoor paths from A to Y are all those paths that remain between A and Y after removing all arrows coming out of A.

These paths are responsible for confounding bias: they imply association not causation.

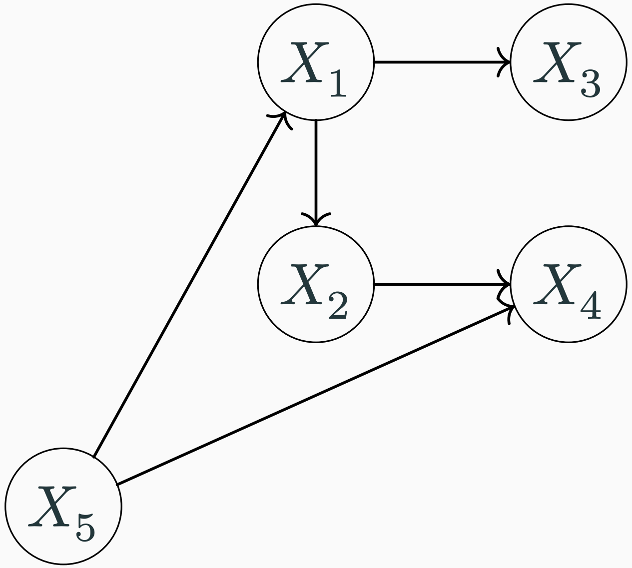




Your turn: What are the backdoor paths from X_1 to X_4? from X_2 tot X_4?



Your turn: What are the backdoor paths from X_1 to X_4? from X_2 tot X_4?

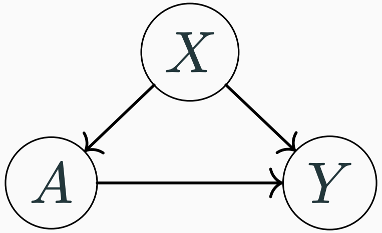


$$X_1 \leftarrow X_5 \rightarrow X_4$$

$$X_2 \leftarrow\leftarrow X_5 \rightarrow X_4$$

Graphical identification

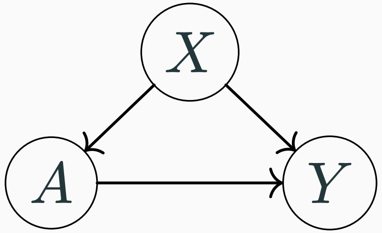
DAGs help us know whether observed covariates are enough to identify a treatment effect.



In other words, how can we make it so that there are no non-causal dependencies between treatment and outcome?

Graphical identification

DAGs help us know whether observed covariates are enough to identify a treatment effect.



In other words, how can we make it so that there are no non-causal dependencies between treatment and outcome?

Graphical identification (Pearl & others, 2000)

The effect of A on Y is identified if all backdoor paths from A to B are blocked, and no descendant of A is conditioned on.

On which variables should we condition? General rules

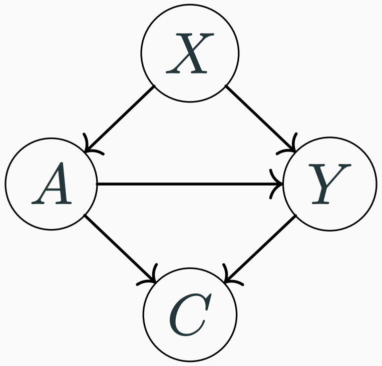
- Do not condition for variables on causal paths from treatment to outcome
- Condition on variables that block non-causal backdoor paths
- Don't condition on colliders! Eg. don't condition on post-treatment variables.

On which variables should we condition? General rules

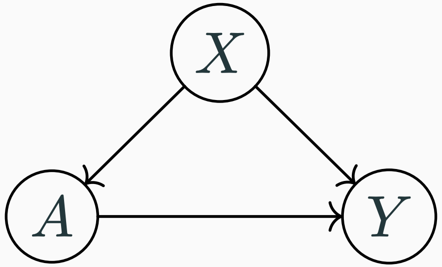
- Do not condition for variables on causal paths from treatment to outcome
- Condition on variables that block non-causal backdoor paths
- Don't condition on colliders! Eg. don't condition on post-treatment variables.

In the following example, to estimate the effect of T on Y, we should:

- Condition on X
- NOT condition on M because it is a descendant of T



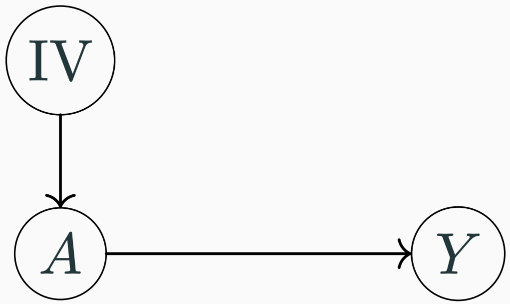
Famous examples of confounders



Effect of education on earnings

The family background can act as a confounder: Wealthier families may provide better education opportunities AND influence earnings independently of the education itself.

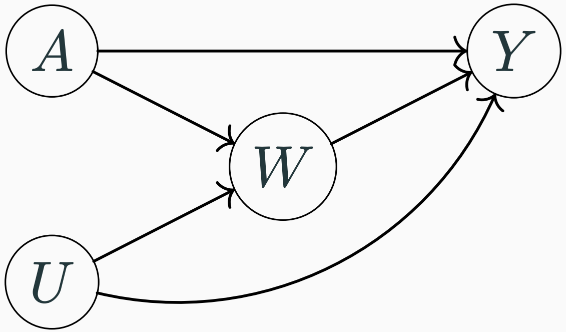
Instrumental variable (IV): influences only the treatment.



Effect of education on earnings (Angrist & Krueger, 1991)

Quarter of birth are randomly assigned but influence the lengths of schooling due to school entry laws.

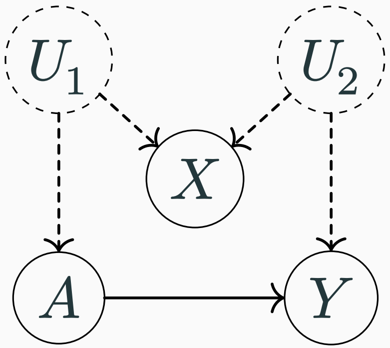
Famous examples of colliders: consequence of both the treatment and the outcome



Effect of smoking on mortality (Hernández-Díaz et al., 2006)

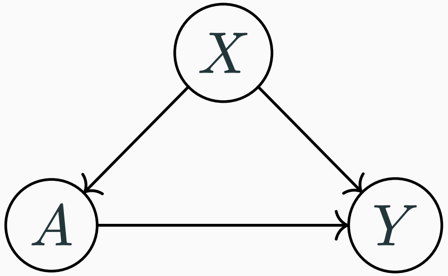
Birth weight is influenced by smoking and other factors. Conditioning on birth weight (a collider) creates a spurious negative correlation between smoking and other risk factors, leading to the paradoxical conclusion that smoking reduces infant mortality, even though it harms overall health.

More colliders: M-bias

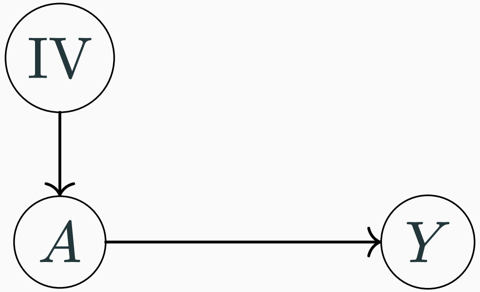


- Do not condition on any pre-exposure variable that you have at disposal!
- Should we condition on X in trying to estimate the effect of A on Y ?
- There is a backdoor path through two unobserved variables (U_1, U_2). But it is blocked because X is a collider along that path.
- Conditioning on X opens up that path, inducing a non-causal association between T and Y.

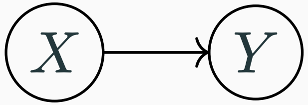
Which variable to include into your analysis?



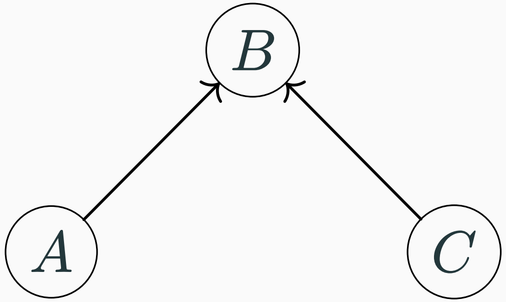
Confounder ✅



Instrumental variable ❌
(generally)

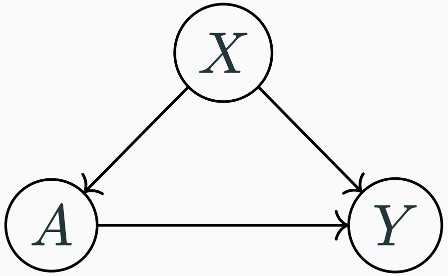


Outcome parent ✅
(generally)

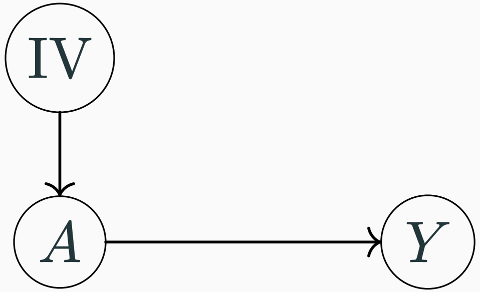


Collider ❌

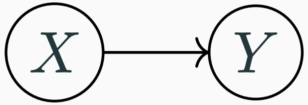
Which variable to include into your analysis?



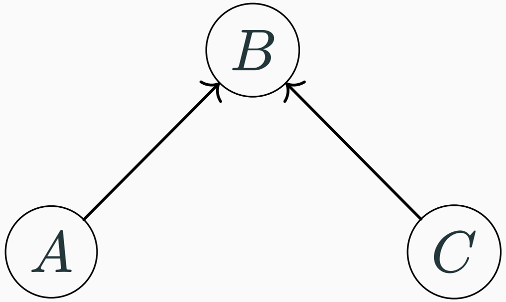
Confounder ✅



Instrumental variable ❌
(generally)



Outcome parent ✅
(generally)



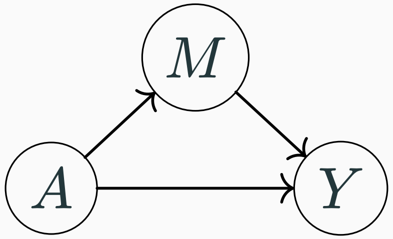
Collider ❌

High-level strategy

Control solely for pre-treatment variables that influences both the outcomes, the treatment or both.

Special types of variables: mediators

A mediator block the path from the treatment to the outcome.



Here, two causal paths from A:

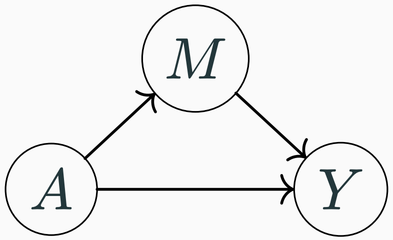
- $A \to Y$ - a "direct effect"
- $A \to M \to Y$ - an "indirect effect" through M

Effect of children poverty on economic outcomes (Bellani & Bia, 2019)

- Y is economic outcomes in adulthood, A is child poverty, M is education.
- What part of the effect of poverty on outcome is mediated by education?

Special types of variables: mediators

A mediator block the path from the treatment to the outcome.



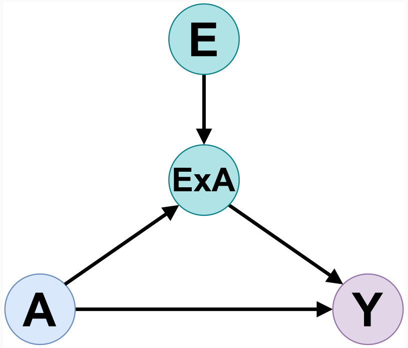
Here, two causal paths from A:

- $A \to Y$ - a “direct effect”
- $A \to M \to Y$ - an “indirect effect” through M

- All causal paths from a treatment capture its overall treatment effect.
- The average treatment effect of T combines both the “direct effect“ and the “indirect effect”.

Special types of variables: Effect modifier

An Effect modifier influences the treatment effect on the outcome.



DAGs

- DAGs are a powerful tool to reason about causality
- Useful to identify the variables to condition on / to include into the analysis
- Drawing the true DAG is often hard / not feasible

Take aways

DAGs

- DAGs are a powerful tool to reason about causality
- Useful to identify the variables to condition on / to include into the analysis
- Drawing the true DAG is often hard / not feasible

Covariate selection: What is important to ensure validity?

- The covariate included (an appropriate DAG)
- The design of the study
- The causal estimator (IPW, G-formula, AIPW...)
- The statistical estimator (Linear regression, Logistic regression...)

Illustration: Causal analysis of EHRs (Doutreligne et al., 2025)

Data: electronic health records from an hospital in Boston

Population: patients with sepsis in the intense care unit

Intervention: combination of crystalloids and albumin for fluid resuscitation

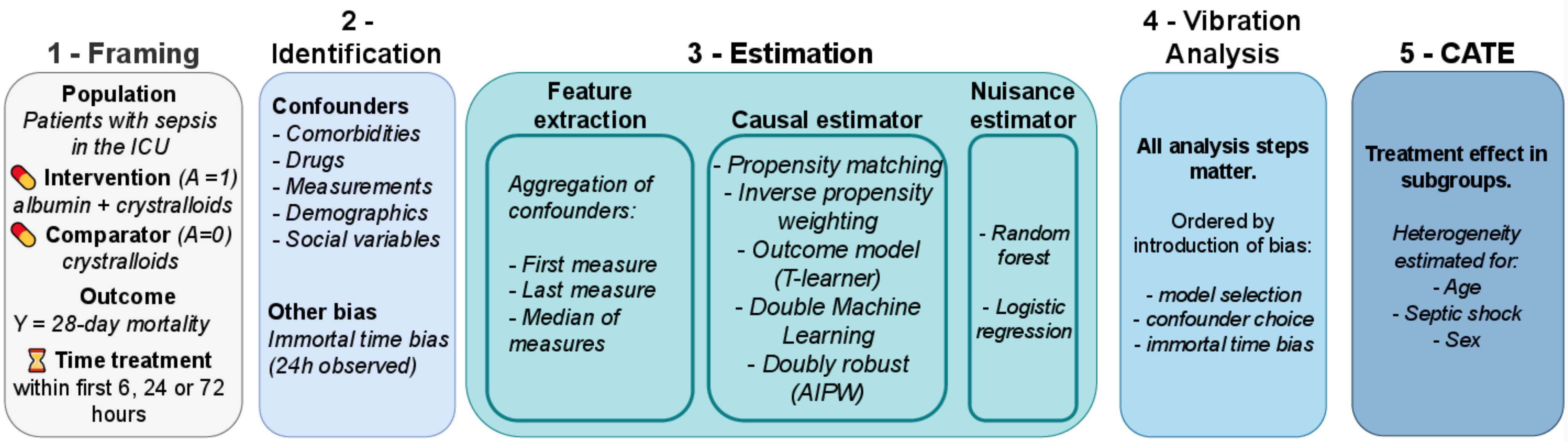
Control: crystalloids only

Outcome: 28-day mortality

Time: Intervention within the first day

💡 Question already answer with RCTs: we have a gold standard for the treatment effect ($\tau = 0$)

Full pipeline



Focus: studying the effect of incomplete DAGs

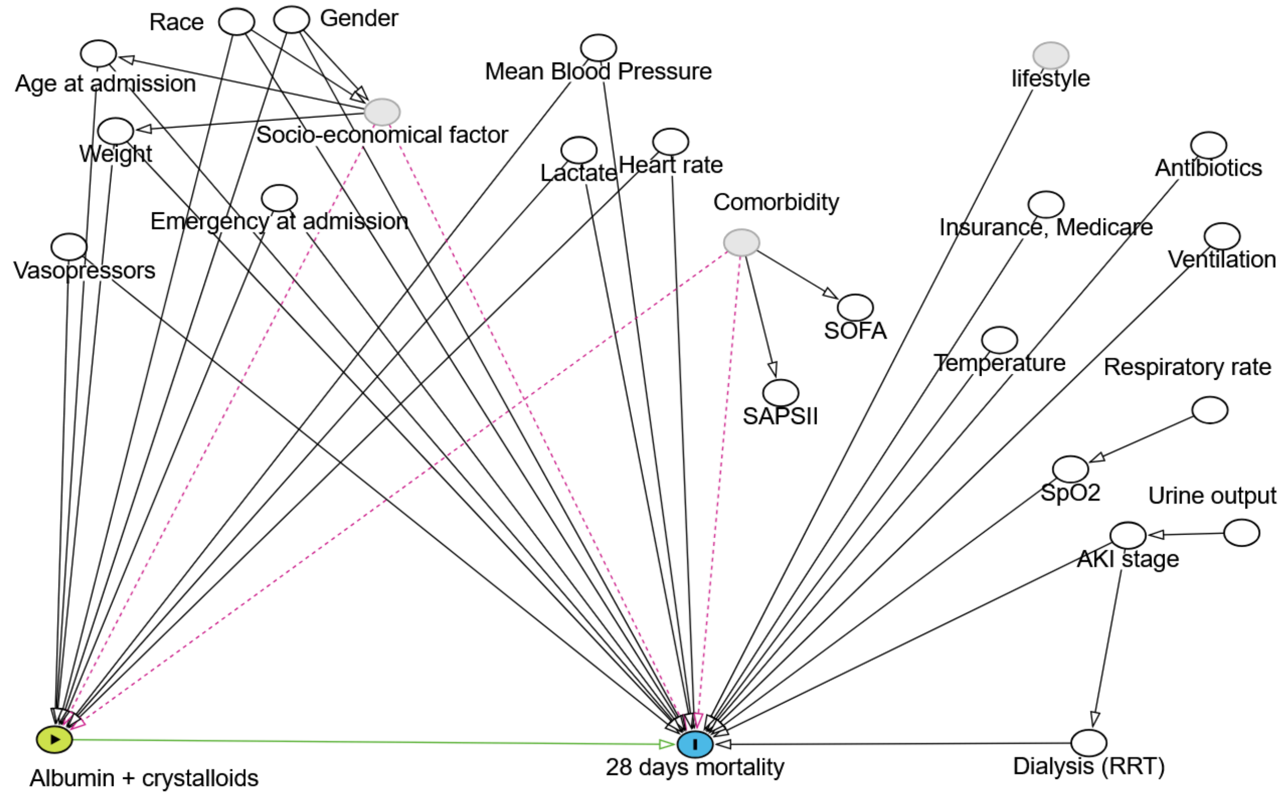
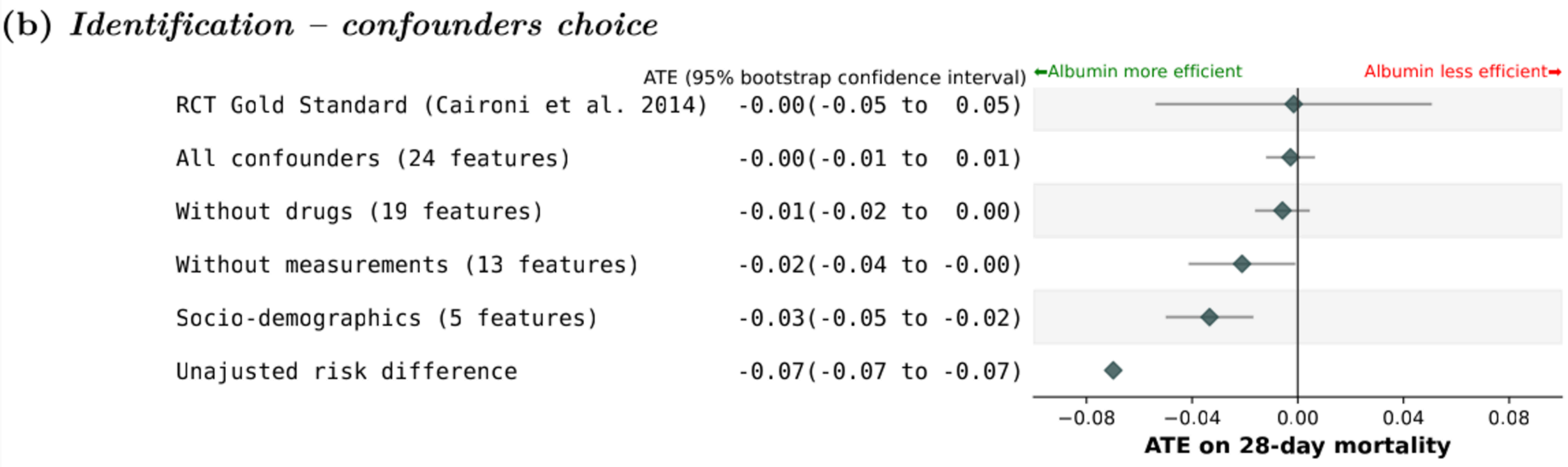
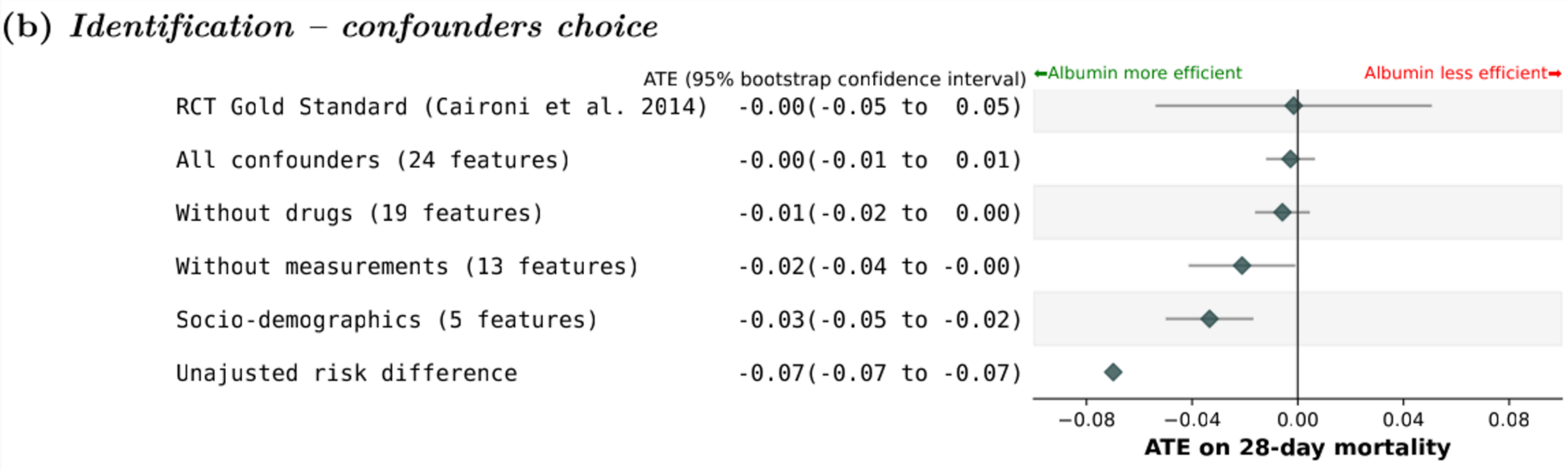


Fig 1. *Causal graph for the Albumin vs crystalloids emulated trial*

Results of an increasingly complete confounder set

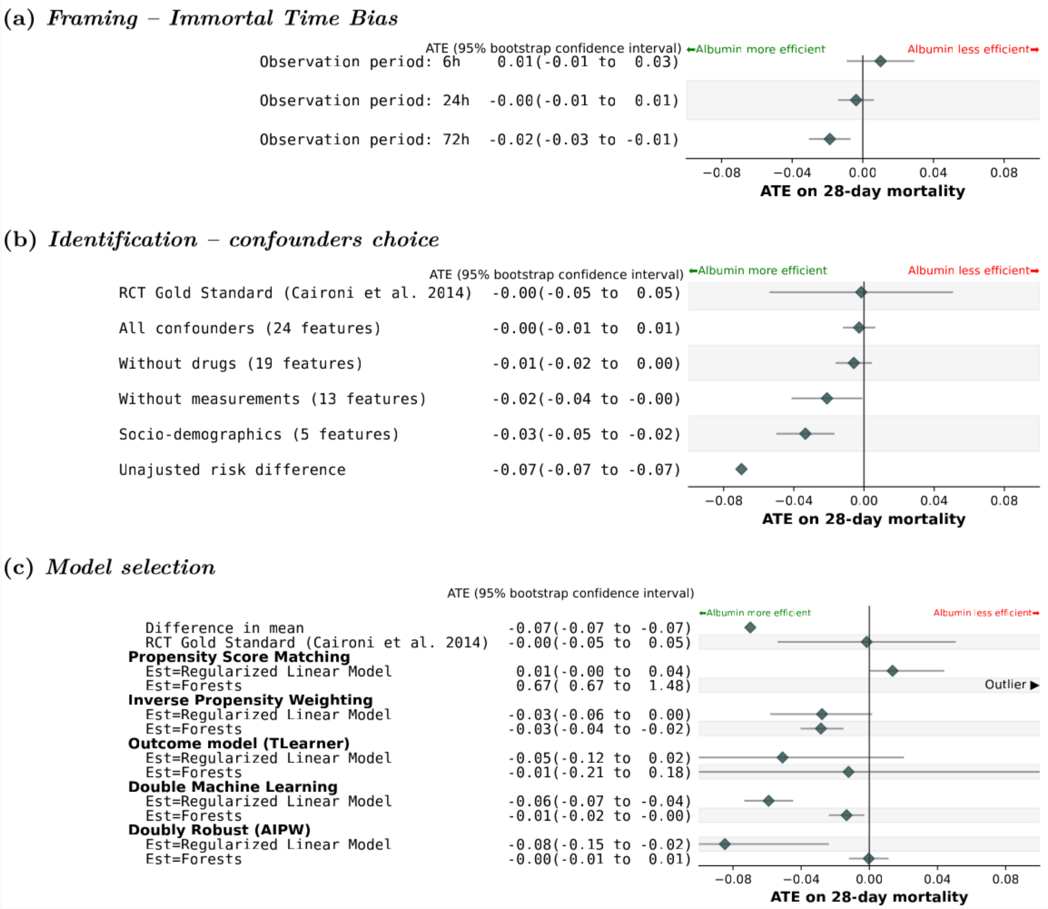


Results of an increasingly complete confounder set



- Missing imortant confounders lead to bias estimates.
- Missing less important confounders lead to less precise estimates: a perfect DAG might not be needed.

Full results of the sensitivity analysis



Practical session

To your notebooks! 👩‍💻

- url: https://straymat.github.io/causal-ml-course/practical_sessions.html

Bibliography

Abécassis, J., Dumas, E., Alberge, J., & Varoquaux, G. (2024). From prediction to prescription: Machine learning and Causal Inference.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. The Quarterly Journal of Economics, 106*(4), 979–1014.*

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly Journal of Economics, 114*(2), 533–575.*

Bellani, L., & Bia, M. (2019). The long-run effect of childhood poverty and the mediating role of education. Journal of the Royal Statistical Society Series A: Statistics in Society, 182*(1), 37–68.*

Bibliography

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. Arxiv Preprint Arxiv:2403.02467*. https://causalml-book.org/*

Colnet, B., Josse, J., Varoquaux, G., & Scornet, E. (2023). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?. Arxiv Preprint Arxiv:2303.16008.

Deschênes, O., & Greenstone, M. (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. American Economic Review, 97*(1), 354–385.*

Doutreligne, M., Struja, T., Abecassis, J., Morgand, C., Celi, L. A., & Varoquaux, G. (2025). Step-by-step causal analysis of EHRs to ground decision-making. PLOS Digital Health, 4*(2), e721.*

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. Journal of Econometrics, 221*(2), 644–654.*

Bibliography

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. Annual Review of Sociology, 40*(1), 31–53.*

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & Oregon Health Study Group, t. (2012). The Oregon health insurance experiment: evidence from the first year. The Quarterly Journal of Economics, 127*(3), 1057–1106.*

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. International Conference on Machine Learning, *1764–1772.*

Hernan, M., & Robins, J. (2020). Causal inference: What if. boca raton: Chapman & hill/crc. *https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/*

Hernández-Díaz, S., Schisterman, E. F., & Hernán, M. A. (2006). The birth weight "paradox" uncovered?. American Journal of Epidemiology, 164*(11), 1115–1120.*

Bibliography

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal machine learning: A survey and open problems. Arxiv Preprint Arxiv:2206.15475.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, *604–620.*

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych, 10*(1), 1–51.*

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82*(4), 669–688.*

Pearl, J., & others. (2000). Models, reasoning and inference. Cambridge, UK: Cambridgeuniversitypress, 19*(2), 3.*

Bibliography

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. ACP Journal Club, 123*(3), A12–3.*

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70*(1), 41–55.*

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66*(5), 688.*

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100*(469), 322–331.*

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36*(2), 111–133.*

VanderWeele, T. J. (2019). Principles of confounder selection. European Journal of Epidemiology, 34, *211–219.*

Bibliography

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage, 145, *166–179.*

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

Wager, S. (2024,). Causal inference: A statistical learning approach. *preparation.*