Causal inference: the potential outcome framework

Machine Learning VS Causal Inference

Machine learning: Powerful predictive models that rely on correlations. A central goal is to understand what usually happens in a given situation: Given today's weather, what's the chance tomorrow's air pollution levels will be dangerously high?

Causal inference: We want to predict what would happen if we change the system: How does the answer to the above question change if we reduce the number of cars on the road?

Concepts of causality are fundamental for having action levers, making recommendations and answering the questions *"what would happen if"?*

Human like AI: reasonable decisions in never experienced situations.

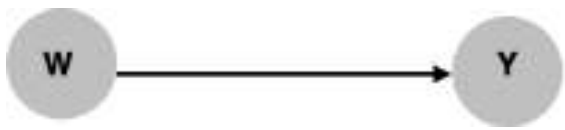
Long tradition in economics and epidemiology, public policies.

Causal discovery VS Causal Inference

Relevant questions about causation:

- Identifying causal direction between 2 variables (Work by Schölkopf, Janzing, Guyon, Peters, and others)
- Learning causal graph structure from data (Work by Bühlmann, Maathuis, Pearl, Meinshausen)

Here, we assume that W causes Y and we want to **estimate the effect** as accurately as possible (bias and variance).



$\Rightarrow$ **Statistical focus but using machine learning machinery** (computationally heavy tools or potentially heuristic approaches e.g., decision trees, neural networks, non-convex optimization)

Examples of causal questions

$\Rightarrow$ Effect of a policy/intervention/treatment W on an outcome Y

- Is there an effect of financial incentives on teacher performance (measured both directly by teacher absences and indirectly by educational output measures, such as average class test scores). [1]
- Does the students succeeded because of the new teacher?
  Had the students remained with the old teacher, they wouldn't have succeeded
- Do job training programs raise average future income?
- Is there an effect of of social pressure on voter participation?
  Neighbors mailing: the recent voting record of everyone on their households would be sent to all their neighbors
- What is the impact of the advertising campaign?
- What is the effect of social media on mental health?
- **What is the effect of hydrochloroquine on mortality?**

[1] Duflo et al. 2012.

PO framework (Neyman, 1923, Rubin, 1974)

Causal effect

- n iid samples $(X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
1.1	20	F	1	200
-6	45	F	0	10
0	15	M	1	150
	...		...	...
-2	52	M	0	100

PO framework (Neyman, 1923, Rubin, 1974)

Causal effect

- n iid samples $(X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
1.1	20	F	1	200
-6	45	F	0	10
0	15	M	1	150
	...		...	...
-2	52	M	0	100

Average treatment effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment.

ATE=0.05: mortality rate in the treated group is 5% points higher than in the control group. So, on average the treatment increases the risk of dying.

PO framework (Neyman, 1923, Rubin, 1974)

Causal effect

- n iid samples $(X_i, W_i, Y_i(1), Y_i(0)) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
1.1	20	F	1	200
-6	45	F	0	10
0	15	M	1	150
	...		...	...
-2	52	M	0	100

Average treatment effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment.

Rq: the post-intervention distribution $P(Y = y \mid do(W = w))$ can also be denoted in the PO framework as $P(Y(w) = y)$.

Randomized Controlled Trial (A/B testing)

Identifiability assumptions

- $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ (consistency)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ **(random treatment assignment)**
  Flip a coin to assign the treatment

We can check that τ
$$\begin{aligned} &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\ &= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] \\ &= \mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0] \end{aligned}$$

$\Rightarrow$ Although Δ_i never observe, τ is identifiable and can be estimated

Difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_0} \sum_{W_i=0} Y_i$$

$\hat{\tau}_{DM}$ unbiased and $\sqrt{n}$-consistent $\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V_{DM})$

Randomized Controlled Trial (A/B testing)

Identifiability assumptions

- $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ (consistency)
- $W_i \perp\!\!\!\perp \{Y_i(0), Y_i(1), X_i\}$ **(random treatment assignment)**
  Flip a coin to assign the treatment

We can check that τ
$$= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$
$$= \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0]$$
$$= \mathbb{E}[Y_i|W_i = 1] - \mathbb{E}[Y_i|W_i = 0]$$

$\Rightarrow$ Although Δ_i never observe, τ is identifiable and can be estimated

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_0} \sum_{W_i=0} Y_i$; ATE = mean(red)-mean(blue)

RCT - Experimental data

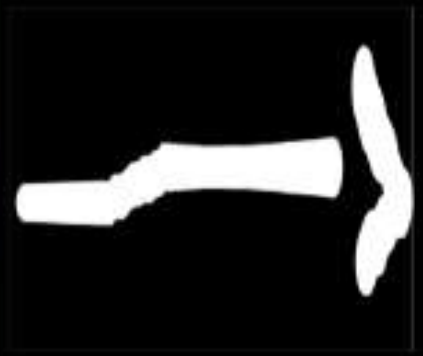
- **Gold standard** to assess the causal effect of an intervention or treatment on an outcome.
- The allocation of the treatment is under control. **The distribution of the covariates for treated and control patients is balanced** (as many young/old; diabetic/non diabetic, etc.) so that a simple difference in means estimator can be consistent. Control group looks like treatment group: difference in response likely due to treatment.

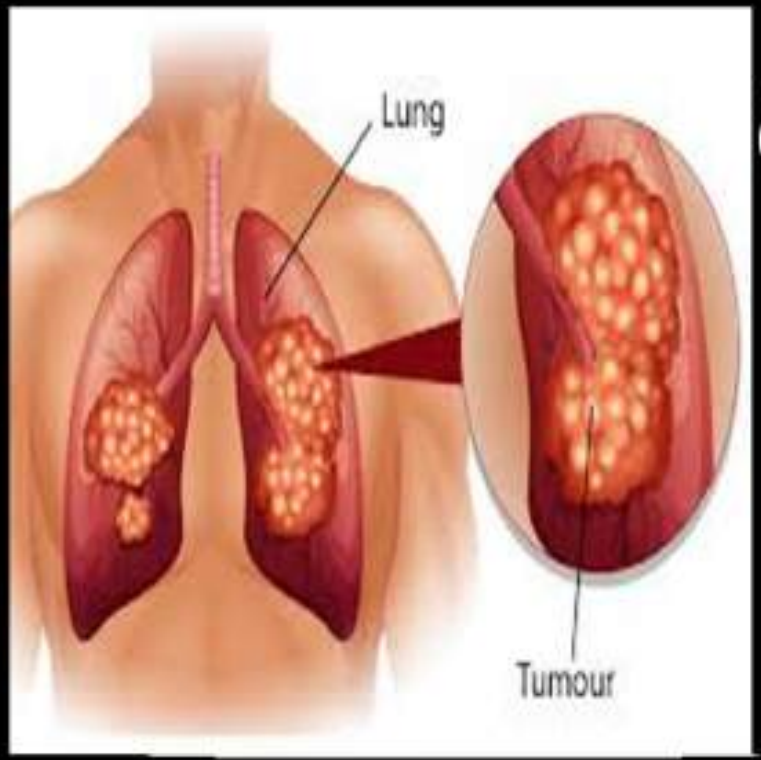
Drawbacks

- expensive, take a long time to set,
- small sample size, due to either recruitment difficulties or restrictive inclusion/exclusion criteria.
- narrowly-defined trial sample that is different from the population potentially eligible for the treatment

Lack of generalizability (**external validity**) to a target population. Study in one company/hospital/state/country could fail to generalize to others

- Not designed for personalized medicine




Lung
Tumour

?

Observational data

- Research: disease registries, epidemiological studies, biobanks/ data routinely collected via EHR, insurance claims, administrative data
- Less costly large sample representative of the target populations

Drawbacks: quality of these "big data"

- lack a controlled design opens the door to **confounding bias**.

Fear of lack of **internal validity**, impossibility of completely ruling out confounding bias.

Covid data

- 4780 patients (patients with at least one PCR-documented SARS-CoV-2 RNA from a nasopharyngeal sample)
- 119 continuous and categorical variables: **heterogeneous**
- 34 hospitals: **multilevel data**

Hospital	Treatment	Age	Sex	Weight	DDI	BP	dead28	...
Beaujon	HCQ	54	m	85	NA	180	yes	
Pitie	AZ	76	m	NA	NA	131	no	
Beaujon	HCQ+AZ	63	m	80	270	145	yes	
Pitie	HCQ	80	f	NA	NA	107	no	
HEGP	none	66	m	98	5890	118	no	
⋮								⋱

Covid data

- 4780 patients (patients with at least one PCR-documented SARS-CoV-2 RNA from a nasopharyngeal sample)
- 119 continuous and categorical variables: **heterogeneous**
- 34 hospitals: **multilevel data**

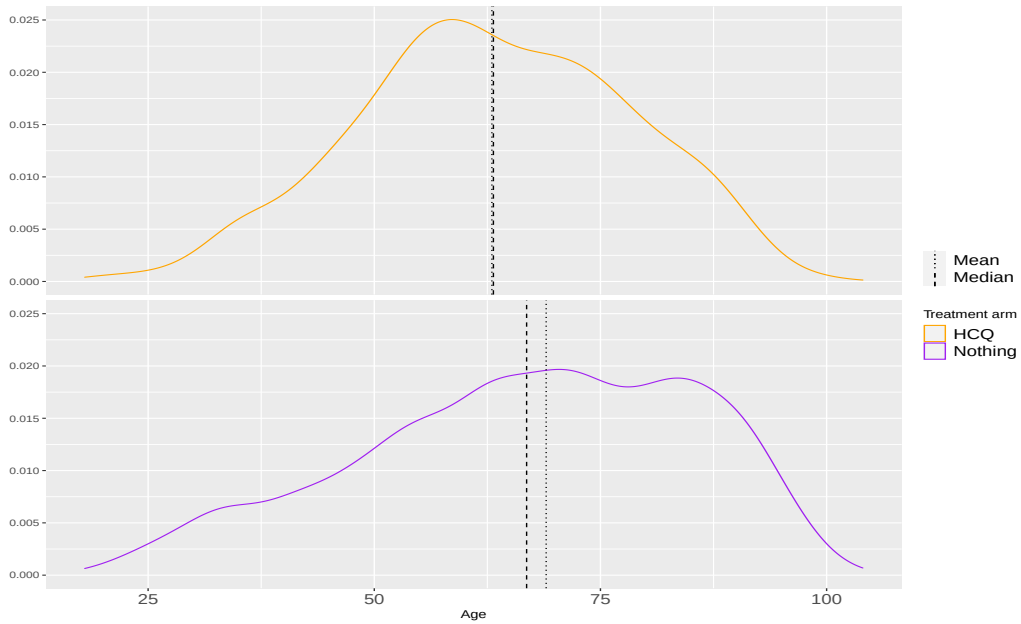
Hospital	Treatment	Age	Sex	Weight	DDI	BP	dead28	...
Beaujon	HCQ	54	m	85	NA	180	yes	
Pitie	AZ	76	m	NA	NA	131	no	
Beaujon	HCQ+AZ	63	m	80	270	145	yes	
Pitie	HCQ	80	f	NA	NA	107	no	
HEGP	none	66	m	98	5890	118	no	
⋮								⋱

$\Rightarrow$ **Estimate causal effect**: Administration of the **treatment** "Hydroxychloroquine" on the **outcome** 28-day mortality.

Observational data: non random assignment

	survived	deceased	Proportion(survived | treatment)	Pr(deceased | treatment)
HCQ	497 (11.4%)	111 (2.6%)	0.817	0.183
HCQ+AZI	158 (3.6%)	54 (1.2%)	0.745	0.255
none	2699 (62.1%)	830 (19.1%)	0.765	0.235

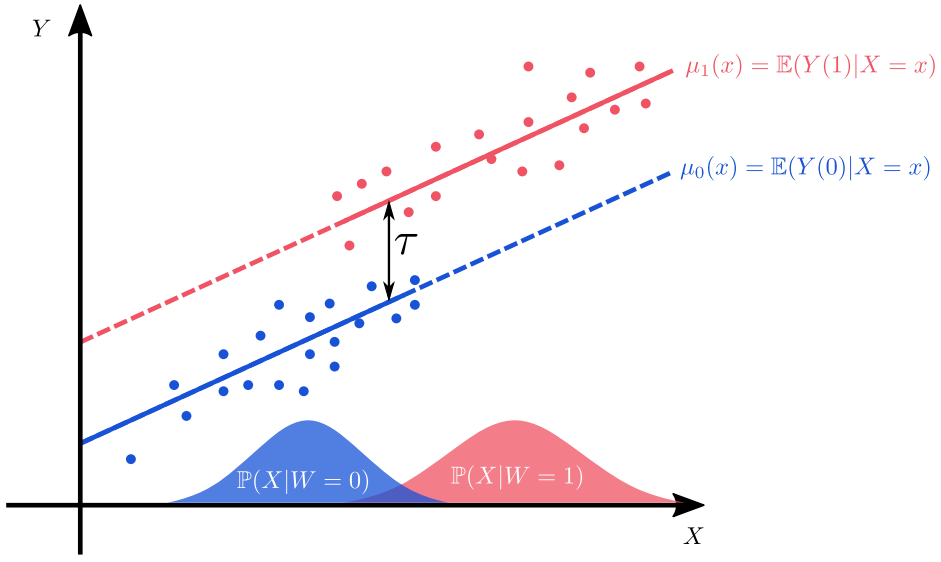
Mortality rate 22.9% - for HCQ 18.3% - non treated 23.5%: treatment helps?



Comparison of the distribution of Age between HCQ and non treated.

Younger patients (with lower risk of death) are more likely to be treated.
If control group does not look like treatment group, difference in response may be **confounded** by differences between the groups.

Observational data: non random assignment



$\Rightarrow$ Treatment assignment depends on covariates X, thus observed covariate distributions of treated and control are different.

A confounder is a third variable that is related to both the exposure of interest and the response.

Causal inference: control for confounding. Estimate causal relation between W and Y when the study is confounded due to the absence of randomization.

Observational data

True PO	
Y(0)	Y(1)
30	200
10	**100**
80	150
...	...
100	**57**

Experiment PO	
Y(0)	Y(1)
?	200
10	**?**
?	150
...	...
100	**?**

Treat.	Out.
W	Y
1	200
0	10
1	150
...	...
0	100

In RCT, the distribution of $Y|W = 0$ is the same as the distribution of $Y(0)$

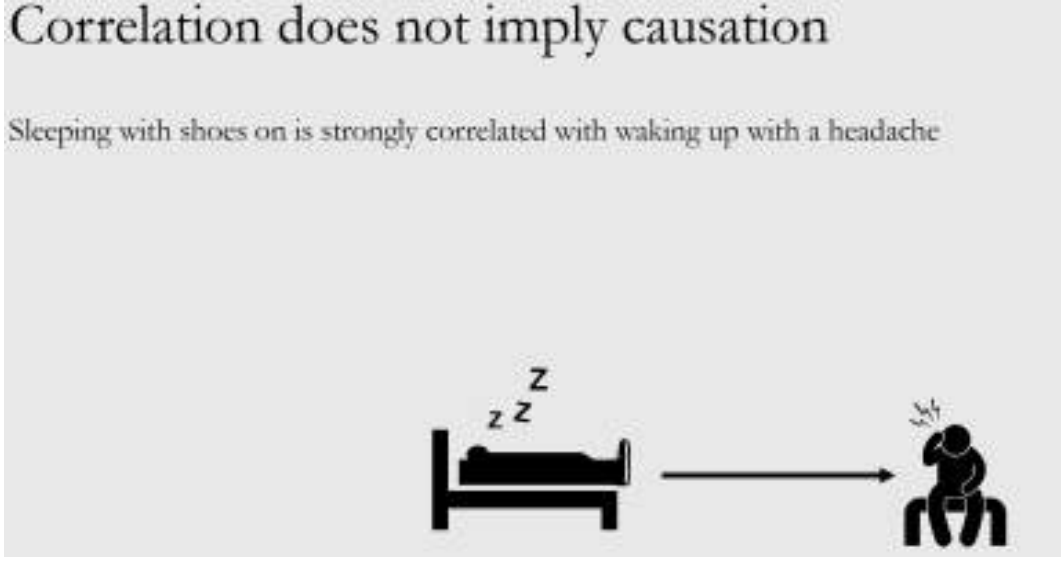
In observational data, the distribution of $Y|W = 0$ is different from the distribution of $Y(0)$

$P(Y|W = 0)$ is different from the distribution of $P(Y|do(W = 0))$

Assumption for ATE identifiability: Unconfoundness

Unconfoundness: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$. Measure all possible confounders. **ATE not identifiable without assumption: it is not a sample size problem, i.e., w/o it we cannot solve even with infinite amount of data.**

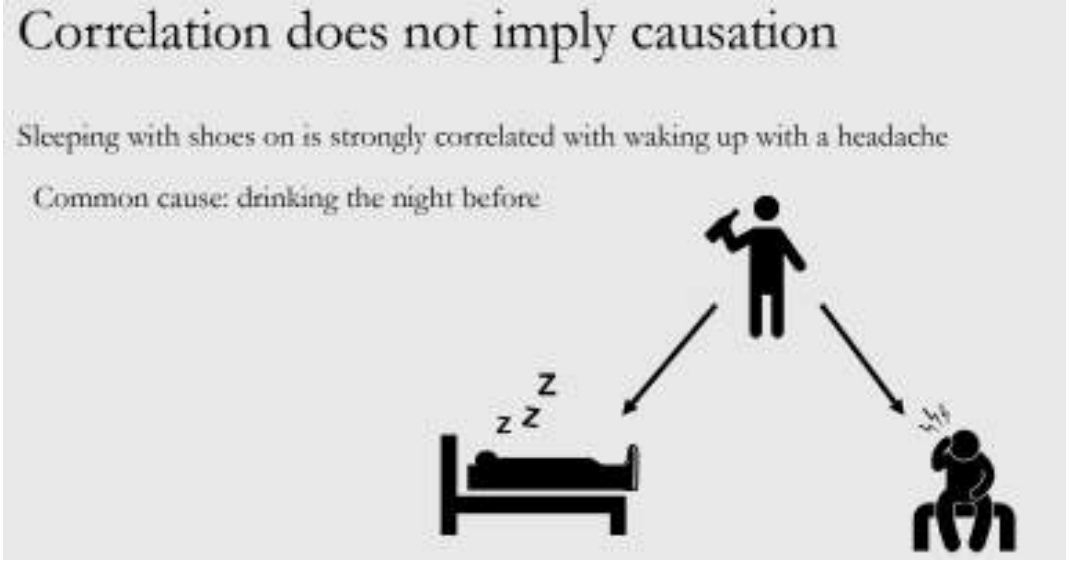
- Unobserved confounders makes it impossible to separate correlation and causality when correlated to both the outcome and the treatment.



Assumption for ATE identifiability: Unconfoundness

Unconfoundness: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$. Measure all possible confounders. **ATE not identifiable without assumption: it is not a sample size problem, i.e., w/o it we cannot solve even with infinite amount of data.**

- Unobserved confounders makes it impossible to separate correlation and causality when correlated to both the outcome and the treatment.

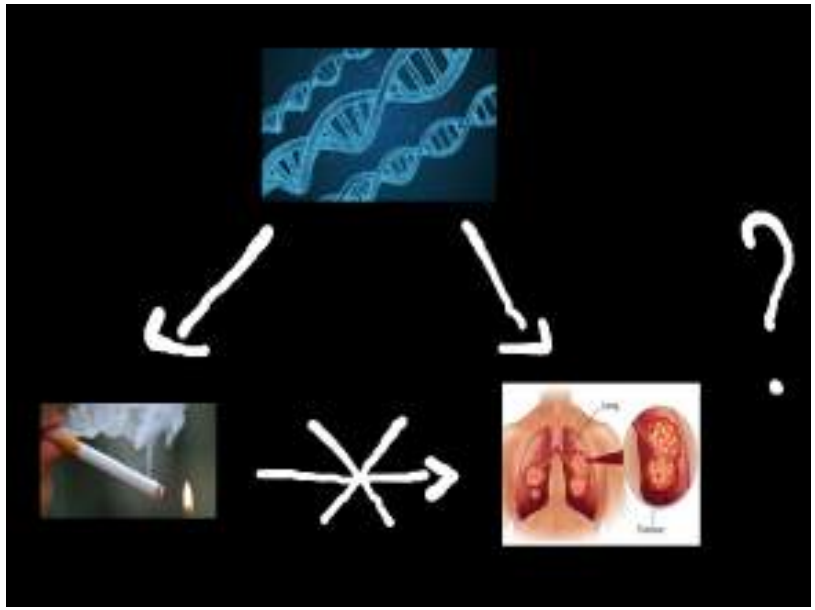


credit: Brady Neal

Assumption for ATE identifiability: Unconfoundness

Unconfoundness: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$. Measure all possible confounders. **ATE not identifiable without assumption: it is not a sample size problem, i.e., w/o it we cannot solve even with infinite amount of data.**

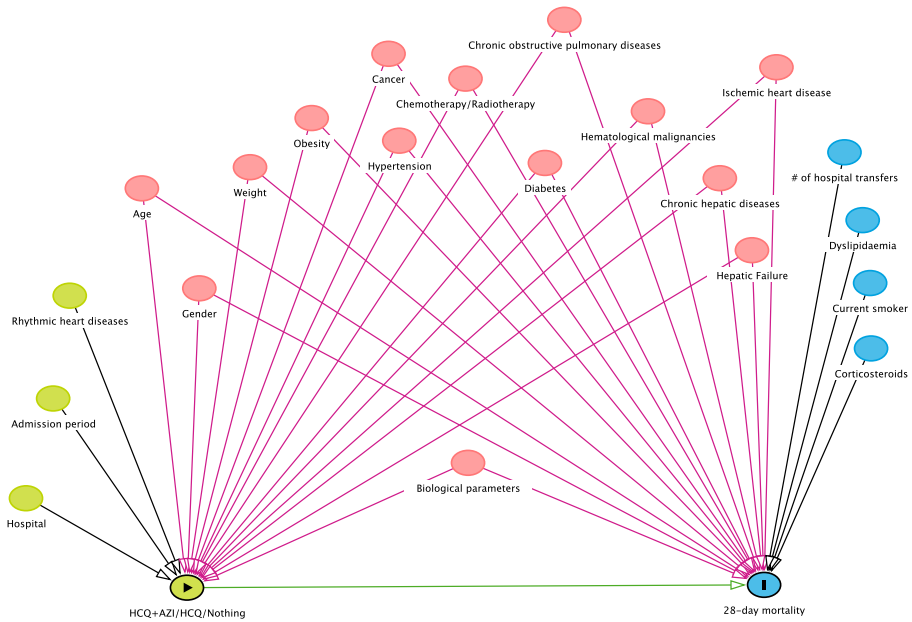
- Assumption not testable from the data.



Unmeasured confounding

Cochran, 1972: *observational studies require a good deal of humility because we can only claim to be groping toward the truth. So, even though we're studying the field of causal inference and we believe we'll do a better job of getting that causality, we're not going to know for sure whether we're there.*

Believe no unmeasured confounding holds? Use domain knowledge.



Some solutions for unmeasured confounding: **Instrumental variables** (a variable which affects treatment assignment but not the outcome); **Sensitivity analysis**, etc.

Assumption for ATE identifiability in observational data

Propensity score - overlap assumption

Propensity score: probability of treatment given observed covariates.

$$e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x) \quad \forall\, x \in \mathcal{X}.$$

We assume overlap, i.e. $\eta < e(x) < 1 - \eta, \quad \forall\, x \in \mathcal{X}$ and some $\eta > 0$

Left: Non smoker and never treated

Right: Smokers and all treated

If proba to be treated when smoker $e(x) = 1$, how to estimate the outcome for smokers when not treated $Y(0)$? How to extrapolate if total confusion?

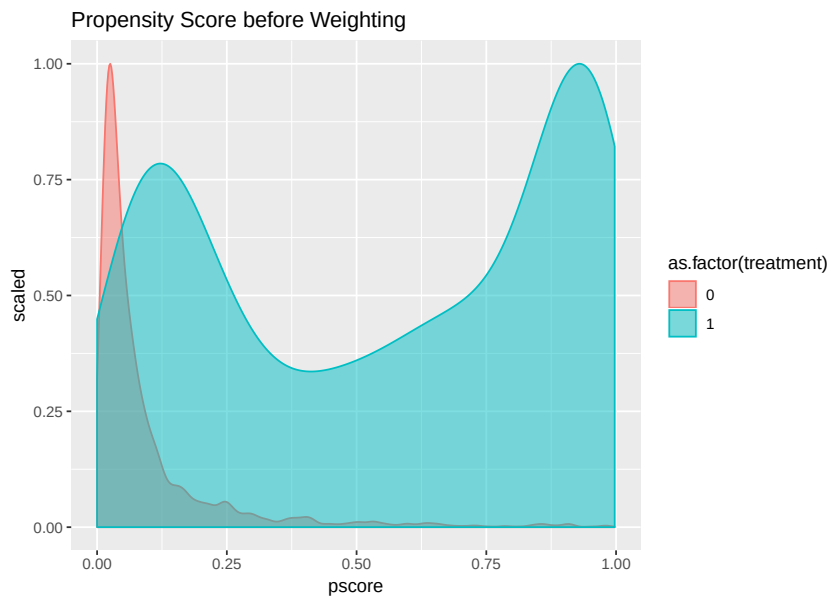
Common support

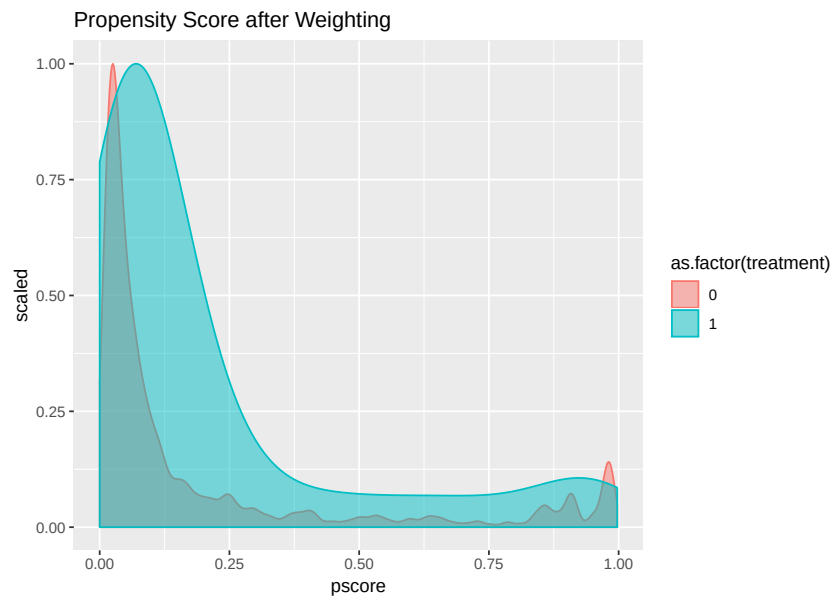
Did not receive job training

Received job training

Solutions to estimate ATE with observational data

- **Matching**: pair each treated (resp. untreated) patient with one or more similar untreated (resp. treated) patient
- **Regression adjustment** (difference between conditional expectation)
- **Inverse-propensity weighting**: to adjust for biases in the treatment assignment. Weighting groups so that control look like treated in terms of distribution of X



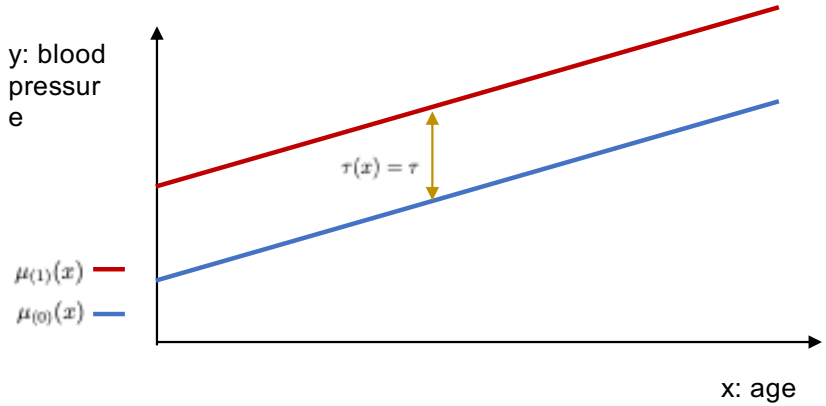


- **Double robust methods** for model misspecifications: covariate balancing propensity score, augmented IPW. (Robins *et al.*, 1994)

Regression adjustment

$\mu_{(w)}(x) \triangleq \mathbb{E}[Y(w)|X = x]$

OLS model $w \in \{0, 1\}$
$Y_i(w) = c_{(w)} + X_i\beta_{(w)} + \varepsilon_i(w)$



Identifiability (using $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$)

$$\begin{aligned}
\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i] = \mathbb{E}[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, |X_i = x]] \text{(uncounfoud)} \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i] - \mathbb{E}[Y_i|W_i = 0, X_i]] \text{(consistency)}
\end{aligned}$$

$\mathbb{E}[Y_i|W_i = 1, X_i]$ can be estimated from data but $\mathbb{E}[Y_i(1)|X_i]$ not.

Backdoor crit: $P(Y|do(w)) = \sum_x P(Y = y|W = w, X = x)P(X = x)$

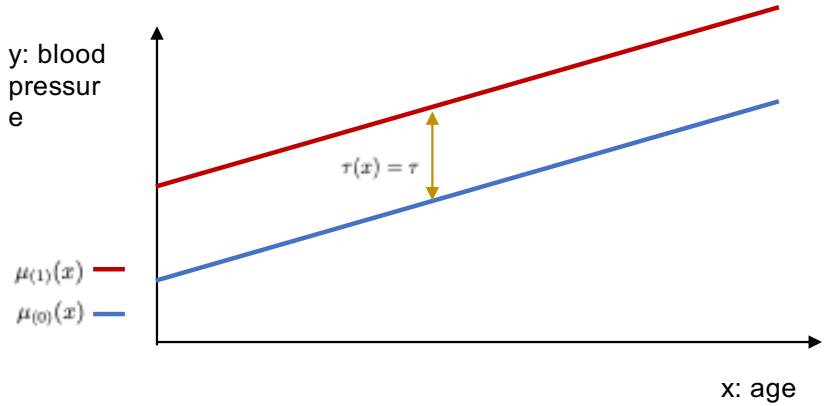
$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) = \frac{1}{n}\sum_{i=1}^{n}(\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}) - (\hat{c}_{(0)} + X_i\hat{\beta}_{(0)})$$

$\Rightarrow$ Consistent if $\hat{\mu}_{(w)}$ consistent

Regression adjustment

$\mu_{(w)}(x) \triangleq \mathbb{E}[Y(w)|X = x]$

OLS model $w \in \{0, 1\}$
$Y_i(w) = c_{(w)} + X_i\beta_{(w)} + \varepsilon_i(w)$



Identifiability (using $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$)

$$\begin{aligned}
\tau &= \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0)|X_i] = \mathbb{E}[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, |X_i = x]] \text{(uncounfoud)} \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i] - \mathbb{E}[Y_i|W_i = 0, X_i]] \text{(consistency)}
\end{aligned}$$

$\mathbb{E}[Y_i|W_i = 1, X_i]$ can be estimated from data but $\mathbb{E}[Y_i(1)|X_i]$ not.

SCM: $\mathbb{E}[Y|do(w)] = \sum_x P(X = x)\mathbb{E}[Y = y|W = w, X = x)]$

$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) = \frac{1}{n}\sum_{i=1}^{n}(\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}) - (\hat{c}_{(0)} + X_i\hat{\beta}_{(0)})$$

$\Rightarrow$ Consistent if $\hat{\mu}_{(w)}$ consistent

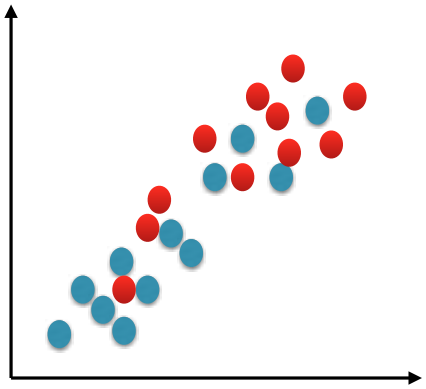
Inverse-propensity weighting estimation of ATE

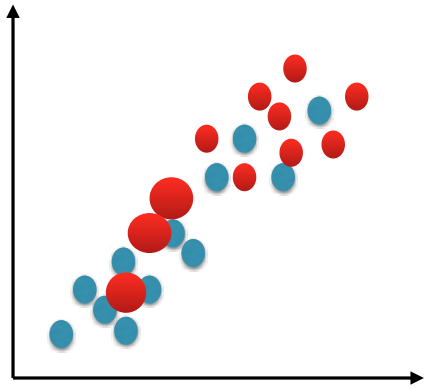
Average treatment effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$
Propensity score (proba to be treated given covariates):
$e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$

IPW estimator (Horvitz-Thomson, survey)

$$\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right)$$

$\Rightarrow$ Balance the differences between the two groups
$\Rightarrow$ Consistent estimator of τ when $\hat{e}(\cdot)$ consistent (logistic regression).
$\Rightarrow$ High variance (divide by probability)





Doubly robust ATE estimation

Define $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \,|\, X_i = x]$ and $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$.

Augmented IPW - Double Robust (DR)

$$\hat{\tau}_{AIPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

- $\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i) Y_i}{1-\hat{e}(X_i)} \right)$: Treatment assignment $\sim$ covariates
- $\hat{\tau}_{OLS} \triangleq \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$: Outcome $\sim$ covariates

$\Rightarrow$ Both sensitive to misspecification. DR: combine ols + ipw of residuals

Doubly robust ATE estimation

Define $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \,|\, X_i = x]$ and $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$.

Augmented IPW - Double Robust (DR)

$$\hat{\tau}_{AIPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

- $\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i) Y_i}{1 - \hat{e}(X_i)} \right)$: Treatment assignment $\sim$ covariates
- $\hat{\tau}_{OLS} \triangleq \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$: Outcome $\sim$ covariates

$\Rightarrow$ Both sensitive to misspecification. DR: combine ols + ipw of residuals

Rationale: makes group similar before extrapolation

$$\sum_{i:W_i=1} (\widetilde{\hat{\mu}}_{(0)}(X_i) - \mu_{(0)}(X_i)) = \underbrace{(\overline{X}_1 - \hat{\gamma}^T \overline{X}_0)}_{\textit{covariate balancing}} \underbrace{(\hat{\beta}^{(0)} - \beta^{(0)})}_{\textit{extrapolation}} + \text{noise term}$$

where $\hat{\gamma} = (1 - \hat{e}(X_j))^{-1}$

Doubly robust ATE estimation

Model Treatment on Covariates $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$
Model Outcome on Covariates $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \,|\, X_i = x]$

Augmented IPW - Double Robust (DR)

$$\hat{\tau}_{AIPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.

Possibility to use **any (machine learning) procedure** such as **random forests**, deep nets, etc. to estimate $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ without harming the interpretability of the causal effect estimation.
Implemented in the R package grf.

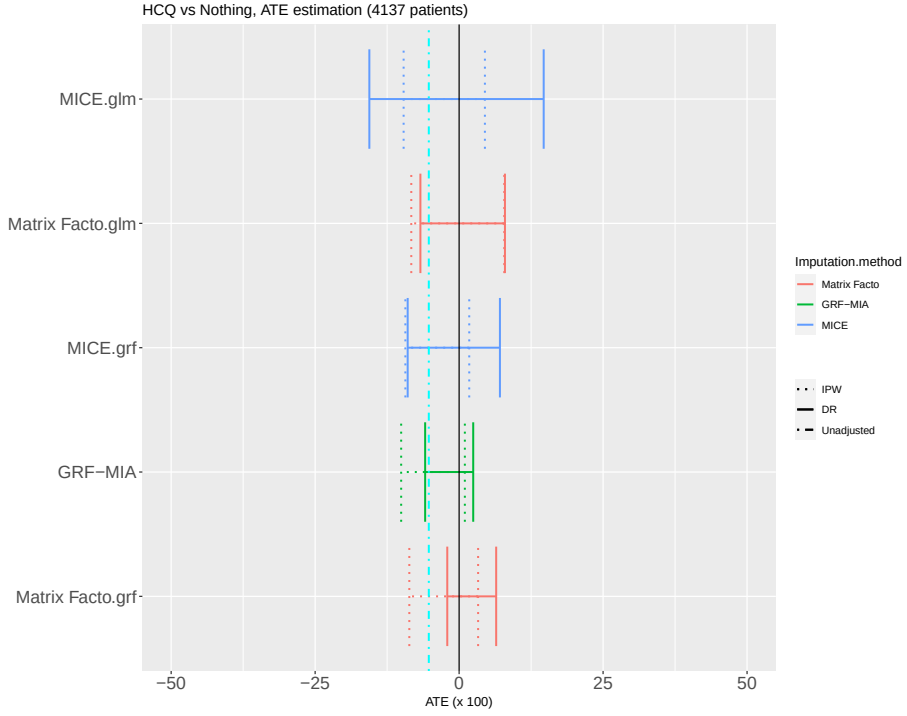
Properties - Double Machine Learning (chernozhukov, et al. 2018)

If $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$ converge at the rate $n^{1/4}$ then

$\sqrt{n}\,(\hat{\tau}_{DR} - \tau) \xrightarrow[n \to \infty]{d} \mathcal{N}(0, V^*)$, V^* semiparametric efficient variance.

Results for Covid Patients

33 covariates, 26 confounders. 4137 patients.
ATE estimations ($\times 100$): effect of Hydroxychloroquine on 28day mortality



(y-axis: estimation approach, solid: **Doubly Robust AIPW**, dotted: **IPW**),
(x-axis: ATE estimation with CI)

The obtained value corresponds to the **difference in percentage points between mortality rates in treatment and control**.
Light Blue: unadjusted (-5.3)
Be careful with missing values...

Conclusion and perspectives

- Causation different from Prediction
- Causality with A/B test
- Causality without A/B test but with additional assumptions **Powerful ML for causal inference**, but strong impact of assumptions
- Two steps: identification and estimation

Heterogeneous treatment effects

SCM VS PO

Supervised learning is not enough

Dataset of 10,000,000 patients – Medications, blood tests, past diagnoses, doctors's notes, demographics, genetic testing

Patient Anna comes in with hypertension.
Anamnesis: Asian, 54, history of diabetes, blood pressure 150/95, ...
Which medication will better lower her blood pressure:

- *Calcium channel blocker (A)?*
- *ACE inhibitor (B)?*

I have data from 10,000,000 other patients. Surely that can help...
Not necessarily! Indeed, this is not a classic supervised learning problem.
Our model was optimized to predict outcome, not to differentiate the influence of A vs. B
What if our high-dimensional model threw away the feature of medication A/B?
Most of the **features except the treatment** can be seen as **nuisance parameters**

Leverage strength of RCT and observational data

RCT: high internal validity, low external validity.
Obs: high external validity, issue with confounding.
$\Rightarrow$ Using both?

• Generalize the treatment effect on a target patient population. The FDA has greenlighted the usage of the drug Ibrance to men withbreast cancer, though clinical trials were performed only on women. Reduce the time to approve a drug for patients who could benefit from it.

• Validate observational methods.

• Improve estimation of heterogeneous treatment effects/ Correcting counfounding biais. RCTs are known to be under-powered for heterogeneous treatment effect. The COVID-19 health crisis is an example of a case where a very rapid response is needed. In the beginning, there are far more observational data than clinical trials.

Notations with the two data sources

- $(X, Y(0), Y(1), W, S)$ drawn from a distribution P
- S binary indicator for the inclusion in the RCT ($S = 1$) or not ($S = 0$).
- RCT drawn from $P(X, Y(0), Y(1), W, S \mid S = 1)$
- Observational drawn from $P(X, Y(0), Y(1), W, S)$

			Cov.			Treat	Out(s)	
		S	X_1	X_2	X_3	W	$Y(0)$	$Y(1)$
rct	1	1	1.1	20	F	1	NA	1
rct		1	-6	45	F	0	1	NA
rct	n	1	0	15	M	1	NA	0
Obs	$n+1$	0		...		...	...	...
Obs		1	-2	52	M	0	1	NA
Obs		0	-1	35	M	1	NA	1
Obs	$n+m$	0	-2	22	M	0	0	NA

Samples in the RCT and observational data do not follow the same distribution (link with **covariate shift problem**).

$\Rightarrow$ Bridging the findings from an RCT to the target population and combining both sources: **generalizability**, **transportability**, **data fusion**[2], **data integration**.

[2] Causal inference & the data-fusion problem. (2016) Elias Bareinboim & Judea Pearl