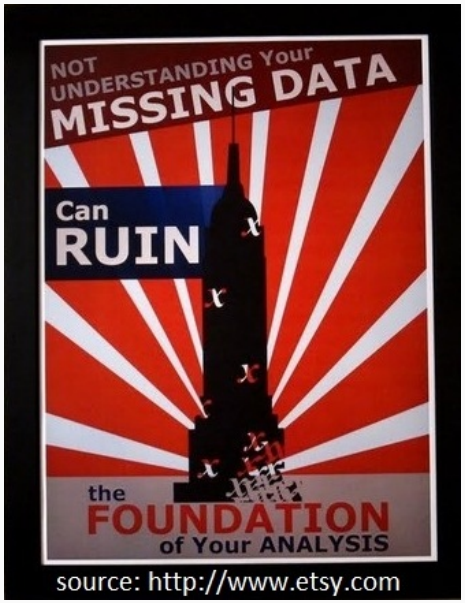
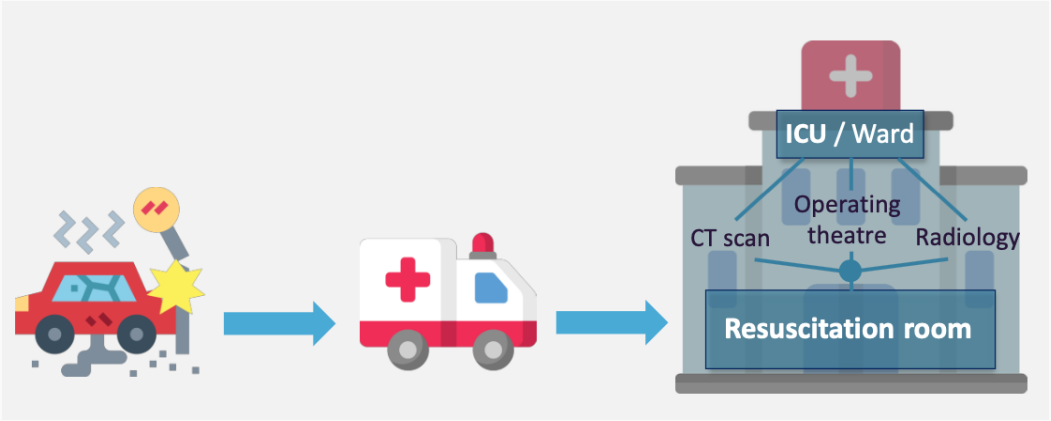
Causal effect on a target population: a sensitivity analysis to handle missing covariates





Julie Josse

My collaborators

Bénédicte Colnet (Inria), Erwan Scornet (X) and Gael Varoquaux (Inria)



Causal inference methods for combining randomized trials and observational studies: a review https://arxiv.org/abs/2011.08047

Causal effect on a target population: a sensitivity analysis to handle missing covariates https://arxiv.org/abs/2105.06435

Causal inference questions

Assume a policy/intervention/**treatment** W causes an **outcome** Y
Aim: **estimate the effect** as acurately as possible (bias & variance)

Medical collaborations:

- Traumabase: polytraumatized patients (car accident, fall, weapon). Suffer from hemorrhagic shock, head trauma
  $\Rightarrow$ Effect of **tranexamic acid** on **mortality** for brain trauma
- Gustave Roussy: define personalized optimal duration of adjuvant endocrine therapy in patients with early breast cancer

Causal inference questions

Assume a policy/intervention/**treatment** W causes an **outcome** Y
Aim: **estimate the effect** as acurately as possible (bias & variance)

Medical collaborations:

- Traumabase: polytraumatized patients (car accident, fall, weapon). Suffer from hemorrhagic shock, head trauma
  $\Rightarrow$ Effect of **tranexamic acid** on **mortality** for brain trauma
- Gustave Roussy: define personalized optimal duration of adjuvant endocrine therapy in patients with early breast cancer

Causal inference questions in many fields:

- Is there an effect of financial incentives on teacher performance (measured by teacher absences & class test scores)? (Duflo et al. 2012)
- What is the impact of the advertising campaign?

Potential Outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect for a binary treatment

▷ n i.i.d. obs ($\underbrace{X_i}_{\text{covariates}}$, $\overbrace{W_i}^{\text{treatment}}$, $\underbrace{Y_i(1), Y_i(0)}_{\textit{potential outcomes}}$) $\in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$

▷ Individual causal effect of the treatment: $\Delta_i = Y_i(1) - Y_i(0)$

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

Potential Outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect for a binary treatment

- n i.i.d. obs ($\underbrace{X_i}_{\text{covariates}}$, $\overbrace{W_i}^{\text{treatment}}$, $\underbrace{Y_i(1), Y_i(0)}_{\textit{potential outcomes}}$) $\in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i = Y_i(1) - Y_i(0)$

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): $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment

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

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

$\Rightarrow$ 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

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

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

Covariates			Treatment	Outcome
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

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

Data to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
  $\Rightarrow$ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  $\Rightarrow$ No personalized medicine
- **trial sample different from the population eligible for treatment**
  $\Rightarrow$ Low **external** validity

Data to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
  $\Rightarrow$ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  $\Rightarrow$ No personalized medicine
- **trial sample different from the population eligible for treatment**
  $\Rightarrow$ Low **external** validity

Observational data

- low cost
- large amounts of data (registries, biobanks, EHR, claims)
  $\Rightarrow$ patient's heterogeneity
- **representative of the target populations**
  $\Rightarrow$ High **external** validity

Data to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions of treated and control groups
  $\Rightarrow$ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  $\Rightarrow$ No personalized medicine
- **trial sample different from the population eligible for treatment**
  $\Rightarrow$ Low **external** validity

Observational data

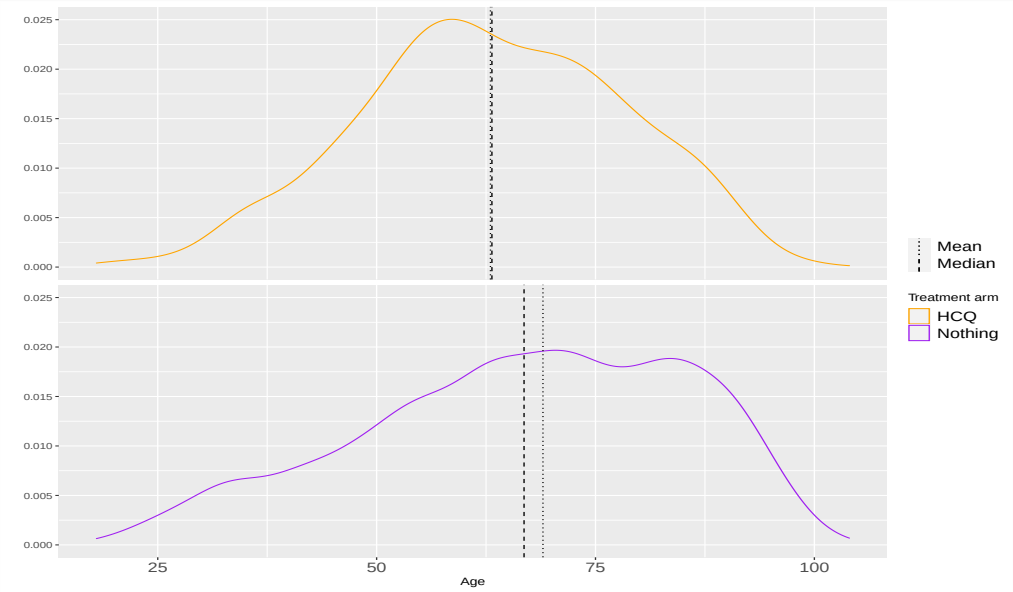
- "big data": low quality
- lack of a controlled design opens the door to **confounding bias**
  $\Rightarrow$ Low **internal** validity

- low cost
- large amounts of data (registries, biobanks, EHR, claims)
  $\Rightarrow$ patient's heterogeneity
- **representative of the target populations**
  $\Rightarrow$ High **external** validity

Confounding biais in observational data: covid ex.

	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.

Leverage both RCT and observational data

RCT

– Narrowly defined population
\+ High **internal** validity

Observational data

– **Confounding**
\+ High **external** validity

We could use both to [1] ...

- ▷ ...validate observational methods
- ▷ ...correct confounding bias
- ▷ ...improve estimation of heterogeneous treatment effects
- ▷ ...**generalize the Average Treatment Effect to a (broader) target population** (data fusion, transportability, data integration)[2]

[1] Colnet, J.J. et al. (2021). Causal inference methods for combining RCT and observational studies: a review. *In review in Statistical Science*.

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

Leverage both RCT and observational data

RCT

– Narrowly defined population

\+ High **internal** validity

Observational data

– **Confounding**

\+ High **external** validity

We could use both to [1] ...

▷ ...validate observational methods

▷ ...correct confounding bias

▷ ...improve estimation of heterogeneous treatment effects

▷ ...**generalize the Average Treatment Effect to a (broader) target population** (data fusion, transportability, data integration)[2]

> The FDA has greenlighted the usage of the drug palbociclib to men with breast cancer, though clinical trials were performed only on women

→ Reduce drug approval times and costs for patients who could benefit

---

[1] Colnet, J.J. et al. (2021). Causal inference methods for combining RCT and observational studies: a review. *In review in Statistical Science*.

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

Generalization task

	Set	S	X_1	X_2	X_3	W	$Y(0)$	Y(1)
1	$\mathcal{R}$	1	1.1	20	5.4	1	?	24.1
...	$\mathcal{R}$	1		...		...	...	
$n-1$	$\mathcal{R}$	1	-6	45	8.3	0	26.3	?
n	$\mathcal{R}$	1	0	15	6.2	1	?	23.5
$n+1$	$\mathcal{O}$	?(0)	-2	52	7.1	NA	NA	NA
$n+2$	$\mathcal{O}$	?(1)	-1	35	2.4	NA	NA	NA
...	$\mathcal{O}$	?(0)		...		NA	NA	NA
$n+m$	$\mathcal{O}$	?(1)	-2	22	3.4	NA	NA	NA

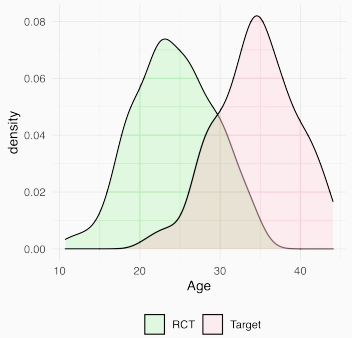
	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{R}$	1	1.1	20	5.4	1	24.1
...	$\mathcal{R}$	1		...		...	...
$n-1$	$\mathcal{R}$	1	-6	45	8.3	0	26.3
n	$\mathcal{R}$	1	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	NA	-2	52	7.1	NA	NA
$n+2$	$\mathcal{O}$	NA	-1	35	2.4	NA	NA
...	$\mathcal{O}$	NA		...		NA	NA
$n+m$	$\mathcal{O}$	NA	-2	22	3.4	NA	NA

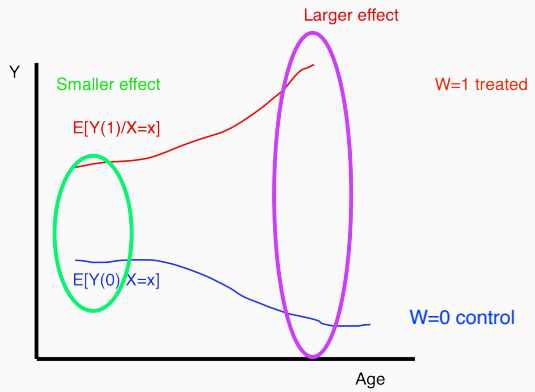
Data with observed treatment W and outcome Y only in the RCT.

- S indicator of eligibility for the trial (not observed in the observational data Set $\mathcal{O}$ but only in Set $\mathcal{R}$)
- Covariates distribution not the same in the RCT & target pop:

$$f_{X|S=1} \neq f_X$$

$$\Rightarrow \quad \underbrace{\tau_1 = \mathbb{E}[Y(1) - Y(0)|S=1]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}[Y(1) - Y(0)] = \tau}_{\text{Target ATE}}$$





Assumption for ATE identifiability

Ignorability assumption on trial participation

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

Trial eligibility S is random conditionally on covariates X

Sampling score - overlap assumption

$$\pi_S(x) = P(S_i = 1 \mid X_i = x) \quad \forall x \in \mathcal{X}$$

Assume overlap, i.e. $\pi_S(x) \geq c > 0, \quad \forall x \in \mathcal{X}$ and some constant c

ATE not identifiable without assumption: it is not a sample size problem!

Identification formulae: Regression formula (g-formula)

	Set	S	X_1	X_2	X_3	W	$Y(0)$	Y(1)
1	$\mathcal{R}$	1	1.1	20	5.4	1	?	24.1
...	$\mathcal{R}$	1		...		...	...	
$n-1$	$\mathcal{R}$	1	-6	45	8.3	0	26.3	?
n	$\mathcal{R}$	1	0	15	6.2	1	?	23.5
$n+1$	$\mathcal{O}$	?(0)	-2	52	7.1	NA	NA	NA
$n+2$	$\mathcal{O}$	?(1)	-1	35	2.4	NA	NA	NA
...	$\mathcal{O}$	?(0)		...		NA	NA	NA
$n+m$	$\mathcal{O}$	?(1)	-2	22	3.4	NA	NA	NA

	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{R}$	1	1.1	20	5.4	1	24.1
...	$\mathcal{R}$	1		...		...	...
$n-1$	$\mathcal{R}$	1	-6	45	8.3	0	26.3
n	$\mathcal{R}$	1	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	NA	-2	52	7.1	NA	NA
$n+2$	$\mathcal{O}$	NA	-1	35	2.4	NA	NA
...	$\mathcal{O}$	NA		...		NA	NA
$n+m$	$\mathcal{O}$	NA	-2	22	3.4	NA	NA

Data with observed treatment W and outcome Y only in the RCT.

Average Treatment Effect: $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$

$$
\begin{aligned}
\mathbb{E}[Y(w)] &= \mathbb{E}[\mathbb{E}[Y(w) \mid X]] && \text{Law of total expectation} \\
&= \mathbb{E}[\mathbb{E}[Y(w) \mid X, S=1]] && \text{Ignorability } \{Y(0), Y(1)\} \perp\!\!\!\perp S \mid X \\
&= \mathbb{E}[\mathbb{E}[Y(w) \mid X, S=1, W=w]] && \text{Random treatment} \\
&= \mathbb{E}[\mathbb{E}[Y \mid X, S=1, W=w]] && \text{Consistency } Y = Y(1)W + (1-W)Y(0)
\end{aligned}
$$

$\Rightarrow$ Transportability assumption: $\tau(x) = \tau_1(x)$

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

g-estimator: difference between conditional mean

g-estimator

$$\hat{\tau}_{g,n,m} = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i) \right),$$

$$\mu_{w,1}(x) = \mathbb{E}\left[Y \mid X = x, S = 1, W = w \right]$$

			Covariates			Treat	Outcomes
	Set	S	X_1	X_2	X_3	W	Y
1	$\mathcal{R}$	1	1.1	20	9.4	1	24.1
	$\mathcal{R}$	1	-6	45	8.3	0	26.3
n	$\mathcal{R}$	1	0	15	6.2	1	23.5
$n+1$	$\mathcal{O}$	?	-1	35	7.1		
$n+2$	$\mathcal{O}$	?	-2	52	2.4		
	$\mathcal{O}$	?		...			
$n+m$	$\mathcal{O}$	?	-2	22	3.4		

- Fit two models of the outcome (Y) on covariates (X) among trial participants ($S = 1$) for treated and for control to get $\widehat{\mu}_{1,n}$ & $\widehat{\mu}_{0,n}$

g-estimator: difference between conditional mean

g-estimator

$$\hat{\tau}_{g,n,m} = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i) \right),$$

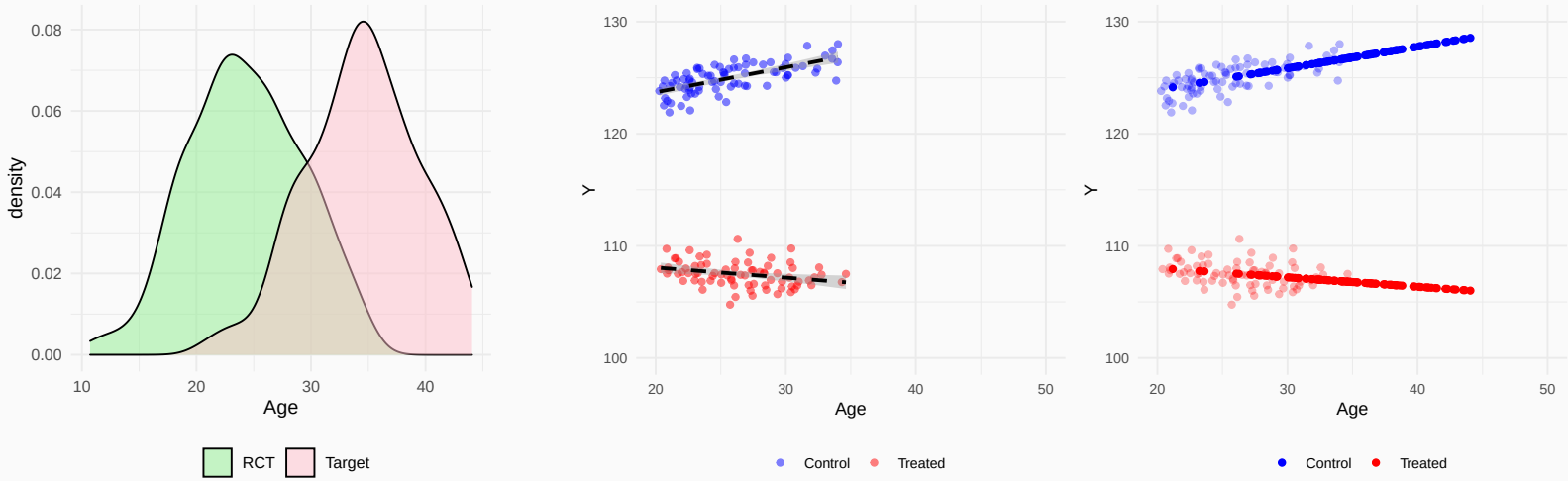
$$\mu_{w,1}(x) = \mathbb{E}\left[Y \mid X = x, S = 1, W = w \right]$$

	Set	S	Covariates X_1	X_2	X_3	Treat W	Outcomes $Y(0)$	$Y(1)$
1	$\mathcal{R}$	1	1.1	20	9.4	1		24.1
	$\mathcal{R}$	1	-6	45	8.3	0	26.3	
n	$\mathcal{R}$	1	0	15	6.2	1		23.5
$n+1$	$\mathcal{O}$	?	-1	35	7.1		$\hat{\mu}_0(X_{n+1})$	$\hat{\mu}_1(X_{n+1})$
$n+2$	$\mathcal{O}$	?	-2	52	2.4		$\hat{\mu}_0(X_{n+2})$	$\hat{\mu}_1(X_{n+2})$
	$\mathcal{O}$	?		...		...	...	...
$n+m$	$\mathcal{O}$	?	-2	22	3.4		$\hat{\mu}_0(X_{n+m})$	$\hat{\mu}_1(X_{n+m})$

- Fit two models of the outcome (Y) on covariates (X) among trial participants ($S = 1$) for treated and for control to get $\widehat{\mu}_{1,n}$ & $\widehat{\mu}_{0,n}$
- Apply these models to the covariates in the target pop, i.e., marginalize over the covariate distribution of the target pop, gives the expected outcomes
- Compute the differences between the expected outcomes on the target population $\overline{\hat{\mu}_{1,n}(\cdot)}$ - $\overline{\hat{\mu}_{0,n}(\cdot)}$

g-estimator: intuition

1) Covariate shift - 2) Estimation in RCT - 3) Extrapolation in Obs data



- ▷ 1) age distribution is different in the RCT (green) and in the target pop (pink), younger in the RCT
- ▷ 2) model outcome on age in the RCT for treated (red) and control (blue)
- ▷ 3) extrapolate the RCT model in the Obs data for older people

Inverse probability of sampling weighting (IPSW)

Y outcome, X covariates, W binary treatment, S eligibility in trial, $e_1(x)$ propensity score $P(W = 1 \mid X = x, S = 1)(= 0.5)$, set $\mathcal{R}$: RCT (size n); $\mathcal{O}$ Obs data (size m).

IPSW estimator

$$\hat{\tau}_{IPSW,n,m} = \frac{1}{m} \sum_{i=1}^{n} \frac{Y_i}{\hat{\alpha}_{n,m}(X_i)} \left(\frac{W_i}{e_1(X_i)} - \frac{1 - W_i}{1 - e_1(X_i)} \right) \quad \text{with,}$$

$$\alpha(x) = \frac{P(i \in \mathcal{R} \mid \exists i \in \mathcal{R} \cup \mathcal{O}, X_i = x)}{P(i \in \mathcal{O} \mid \exists i \in \mathcal{R} \cup \mathcal{O}, X_i = x)} = \frac{n}{m} \times \frac{f_{X|S=1}(x)}{f_X(x)} = \frac{P(S = 1)}{P(S = 1 \mid X_i = x)}$$

- weighted difference of average Y between treated & control in trial
- weights: inverse of odd ratio of the indicatrix of being in RCT (account for the shift of the covariate distribution from RCT to target pop.)

Ex: if proba to be in trial when old is small, then up-weight old in trial
$\Rightarrow$ Balance the differences between the two groups RCT and Obs data

IPSW: intuition

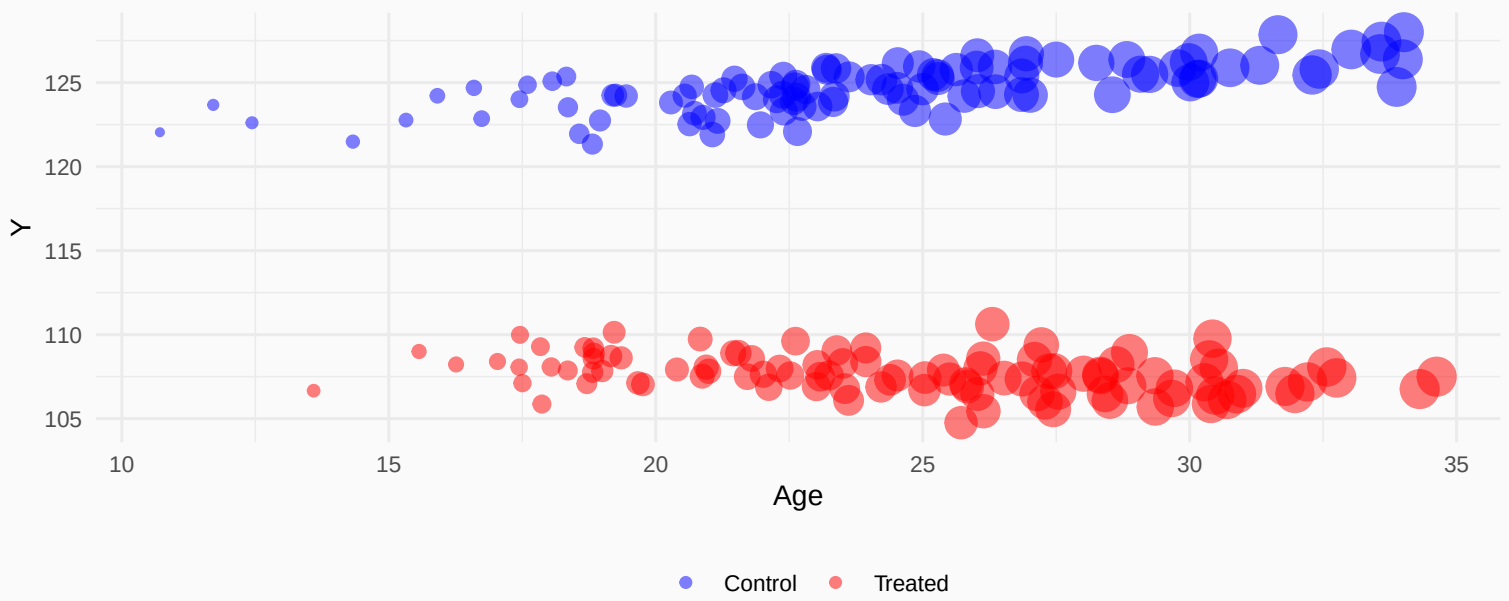


Figure 1: Representation of the outcome with respect to age in the trial sample for treated (red) and control (blue). Size of the dot corresponds to the weight.

$\Rightarrow$ Reweights the RCT sample so that it looks like the target population distribution: Older persons are upweighted because the target population is older than the RCT one.

Estimators of the average treatment effect for generalization

g-estimator: model conditional outcomes & extrapolate to target

Model Outcome on Covariates $\mu_{(w,1)}(w) = \mathbb{E}[Y_i(w) \mid X_i = x, S_i = 1]$

Assumption: consistency of surface responses' estimation

- (H1-g) For $w \in \{0,1\}$, $\mathbb{E}[\hat{\mu}_{w,n}(X) \mid \mathcal{D}_n] \xrightarrow{a.s.} \mathbb{E}[\mu_w(X)]$ when $n \to \infty$,
- (H2-g) For $w \in \{0,1\}$, there exist C_1, N_1 so that for all $n \geqslant N_1$, almost surely, $\mathbb{E}[\hat{\mu}^2_{w,n}(X) \mid \mathcal{D}_n] \leqslant C_1$. $\mathcal{D}_n$ the RCT sample.

Theorem: g-estimator consistency $|\hat{\tau}_{g,n,m} - \tau| \xrightarrow{a.s.} 0$, when $n, m \to \infty$

IPSW: weight RCT sample so that it ressembles the target pop

Model Set on Covariates $P(i \in \mathcal{R} \mid \exists i \in \mathcal{R} \cup \mathcal{O}, X_i = x)$, estimate odd ratio α

Assumption: consistency of weights' estimation

- (H1-IPSW) $\sup_{x \in \mathcal{X}} |\frac{n}{m\hat{\alpha}_{n,m}(x)} - \frac{f_X(x)}{f_{X|S=1}(x)}| = \epsilon_{n,m} \xrightarrow{a.s.} 0$, when $n, m \to \infty$
- (H2-IPSW) Y is square-integrable

Theorem: IPSW consistency $|\hat{\tau}_{IPSW,n,m} - \tau| \xrightarrow{a.s.} 0$, when $n, m \to \infty$

Doubly robust estimator: combine the previous ideas

Augmented IPSW

$$\hat{\tau}_{AIPSW,n,m} = \frac{1}{n}\sum_{i=1}^{n} \frac{n}{m\,\hat{\alpha}_{n,m}(X_i)} \left[\frac{W_i\,(Y_i - \hat{\mu}_{1,n}(X_i))}{e_1(X_i)} - \frac{(1-W_i)\,(Y_i - \hat{\mu}_{0,n}(X_i))}{1-e_1(X_i)} \right] + \frac{1}{m}\sum_{i=n+1}^{m+n} (\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i))\,.$$

is consistent if $\hat{\mu}_{w,1}(X)$ $(w = 0, 1)$ or $\hat{\alpha}_{n,m}(X)$ are consistent.

Possibility to use **any (machine learning) procedure** such as **random forests**, deep nets, etc. to estimate $\hat{\alpha}_{n,m}(X)$ and $\hat{\mu}_{(w,1)}(x)$ without harming the interpretability of the causal effect estimation.

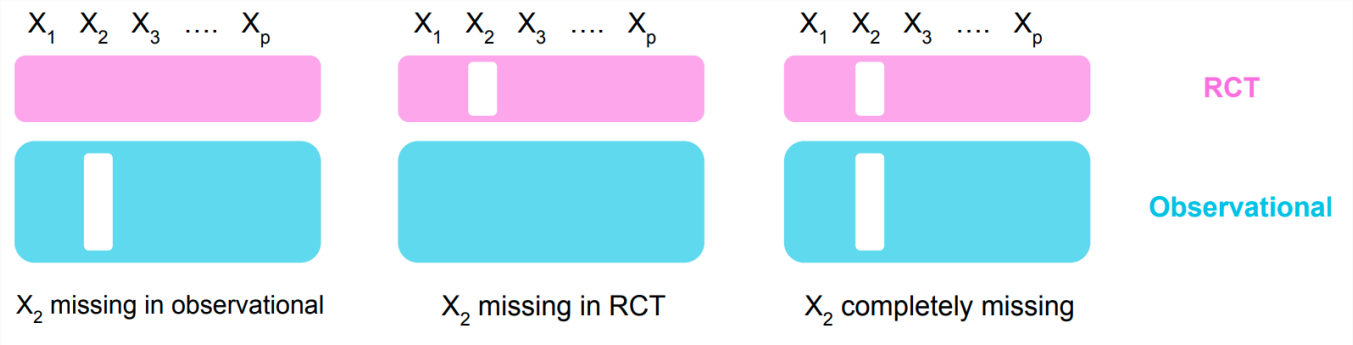
Property (to be proved)

If $\mathbb{E}[(\hat{\mu}_{w,n}(X) - \mu_w(X))^2]\mathbb{E}[(\alpha_{n,m}(X)) - \alpha(X))^2] = o\left(\frac{1}{n}\right)$, then $\sqrt{n}\,(\hat{\tau}_{AIPSW,n,m} - \tau) \xrightarrow[n,m\to\infty]{d} \mathcal{N}(0, V^*)$, V^* semiparametric efficient variance [3].

[3] Chernozukov, Duflot, et al (2018), *Double/debiased machine learning for treatment and structural parameters. Econometrics journal*

Sensitivity analysis

What if a covariate is missing?



In practice the common subset of covariates is used

- X_{mis} totally missing or partially missing covariate
- X_{obs} covariates observed in both data sets
- $X = X_{\text{mis}} \cup X_{obs}$

$\Rightarrow$ Breaks the identifiability assumption $\{Y(1), Y(0)\} \not\perp\!\!\!\perp S \mid X_{obs}$

Is there a way to assess how dramatic the situation is?

Solutions: **sensitivity analysis** (Rosenbaum & Rubin (1983), Imbens (2003) [4], Franks et. al. (2019), Veitch & Zaveri (2020), etc.)

[4] Sensitivity to Exogeneity Assumptions in Program Evaluation, *The American Economic Review*

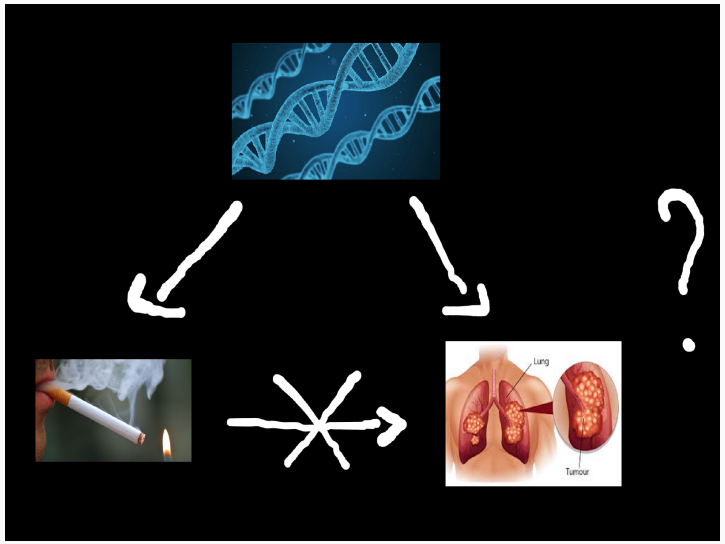
Unobserved confounders: major issue with observational data

Unconfoundness identifiability assumption

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$$

$\Rightarrow$ Measure all possible confounders X (drives treatment W & outcome (Y); ex. age in covid data)

- ▷ Unobserved confounders: impossible to separate correlation & causality
- ▷ Assumption not testable from the data



Smoking and lung cancer - Cornfield, 1956

- If people have gene B: their disease rate is r_1. If not, disease rate is r_2 (we suppose a lower prevalence).
- Instead of B, we observe A, smoking status. Suppose that, $p(B \mid A) = p_1$ and $p(B \mid \bar{A}) = p_2$, and the presence of B is correlated with A, so $p_1 > p_2$.
- In practice, when observing A, we observe an apparent rate of disease denoted R_A: $p_1 r_1 + (1 - p_1) r_2 = R_A$.
- Because $R_A > R_{\bar{A}}$, and doing a bit of computation gives ...

$$\frac{p_1}{p_2} = \frac{R_A}{R_{\bar{A}}} + \frac{r_2}{p_2 r_1} \left(\frac{R_A}{R_{\bar{A}}} (1 - p_2) - (1 - p_1) \right).$$

- Because $p_1 > p_2$ and $R_A > R_{\bar{A}}$, the third term is positive, therefore, $\frac{R_A}{R_{\bar{A}}} < \frac{p_1}{p_2}$.

If cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer (i.e. $\frac{R_A}{R_{\bar{A}}} = 9$), and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X producers among cigarette smokers must be at least 9 times greater than nonsmokers (i.e. $\frac{p_1}{p_2} > 9$).

Semi-parametric model

Assumption on the generative model

Consider that the potential outcomes are generated according to:

$$Y(W) = \mu(W, X) + \varepsilon_W,$$

for any function $\mu \in \mathbf{L}^2(\{0,1\} \times \mathcal{X} \to \mathbb{R})$ and such that $\mathbb{E}[\varepsilon_W \mid X] = 0$.

With binary treatment W, there exists a function $g : \mathcal{X} \to \mathbb{R}$ such that

$$Y(W) = g(X) + W\,\tau(X) + \varepsilon_W, \quad \text{where } \tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$$

Linear Conditional Average Treatment Effect (CATE)

We suppose there exist $\delta \in \mathbb{R}^d$, and $\sigma \in \mathbb{R}^+$ such that:

$$Y = g(X) + W\langle X, \delta\rangle + \varepsilon, \qquad \text{where } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Semi-parametric model

Assumption on the generative model

Consider that the potential outcomes are generated according to:

$$Y(W) = \mu(W, X) + \varepsilon_W,$$

for any function $\mu \in \mathbf{L}^2(\{0,1\} \times \mathcal{X} \to \mathbb{R})$ and such that $\mathbb{E}[\varepsilon_W \mid X] = 0$.

With binary treatment W, there exists a function $g : \mathcal{X} \to \mathbb{R}$ such that

$$Y(W) = g(X) + W\,\tau(X) + \varepsilon_W, \quad \text{where } \tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$$

Linear Conditional Average Treatment Effect (CATE)

We suppose there exist $\delta \in \mathbb{R}^d$, and $\sigma \in \mathbb{R}^+$ such that:

$$Y = g(X) + W\langle \delta_{obs}, X_{obs}\rangle + W\,\delta_{mis}\,X_{mis} + \varepsilon, \qquad \text{where } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Asymptotic bias when considering the common set of variables

Assumption on covariates

$X \sim \mathcal{N}(\mu, \Sigma)$

$X \mid S = 1 \sim \mathcal{N}(\mu_{RCT}, \Sigma)$ (transportability of Σ)

$X = X_{\text{mis}} \cup X_{obs}$

Theorem

Assume partially linear model & assumption on covariates. Let B be:

$$B = \sum_{j \in mis} \underbrace{\delta_j}_{X_{mis}\text{'s strength}} \left(\underbrace{\mathbb{E}[X_j] - \mathbb{E}[X_j \mid S = 1]}_{\text{Shift of } X_{mis}} - \Sigma_{j,obs} \Sigma^{-1}_{obs,obs} (\mathbb{E}[X_{obs}] - \mathbb{E}[X_{obs} \mid S = 1]) \right),$$

with $\Sigma_{obs,obs}$ submatrix of Σ of observed index of rows and columns.

- ▷ Granting consistency of the surface response, $\tau - \lim_{n,m \to \infty} \mathbb{E}[\hat{\tau}_{g,n,m,obs}] = B$
- ▷ Granting consistency of IPSW, $\tau - \lim_{n,m \to \infty} \mathbb{E}[\hat{\tau}_{IPSW,n,m,obs}] = B$

Large distributional shift & treatment modifying strength → large bias

Sensitivity analysis for a completely missing covariate

	Set	S	Covariates X_1	X_2	X_3	Treat W	Outcomes Y
1	$\mathcal{R}$	1	1.1	20	NA	1	24.1
	$\mathcal{R}$	1	-6	45	NA	0	26.3
n	$\mathcal{R}$	1	0	15	NA	1	23.5
$n+1$	$\mathcal{O}$	?	-1	35	NA		
$n+2$	$\mathcal{O}$	?	-2	52	NA		
	$\mathcal{O}$	?		...			
$n+m$	$\mathcal{O}$	?	-2	22	NA		

$$B = \underbrace{\delta_{mis}}_{X_{mis}\text{'s strength}} \quad \underbrace{(\mathbb{E}[X_{mis}] - \mathbb{E}[X_{mis} \mid S=1]}_{\text{Shift of } X_{mis}\text{: } \Delta_m} - \underbrace{\Sigma_{mis,obs}\Sigma^{-1}_{obs,obs}(\mathbb{E}[X_{obs}] - \mathbb{E}[X_{obs} \mid S=1]))}_{\text{Can not be estimated from the data}}$$

Bias under an additional independence assumption

$X_{mis} \perp\!\!\!\perp X_{obs}$ and $X_{obs} \perp\!\!\!\perp X_{mis} \mid S = 1$

Asymptotic bias: $\tau - \lim_{n,m\to\infty} \mathbb{E}[\hat{\tau}_{G,n,m,obs}] = \delta_{mis}\,\Delta_m$

- ▷ Define range for plausible δ_{mis} values
- ▷ Define range for plausible Δ_m values
- ▷ Compute all possible bias $\delta_{mis}\Delta_m$ and return Austen plot

Imbens (2003) for observational data: 2 sensitivity parameters (strength on W & Y, linear models)

Austen plot for a completely missing covariate

What would be the bias of $\hat{\tau}_{n,m,obs}$ when we totally missed X_3?

Sensitivity analysis: translates sensitivity parameter(s) (treatment effect modifier's strength δ_{miss} & covariate shift's strength Δ_m) into range bias

Austen plot: shows how strong an unobserved key covariate would need to be to induce a bias that would force to reconsider the conclusions (bias above a certain threshold, $\hat{\tau}_{n,m,obs} - \hat{\tau}_1 =\sim 6$ in blue)

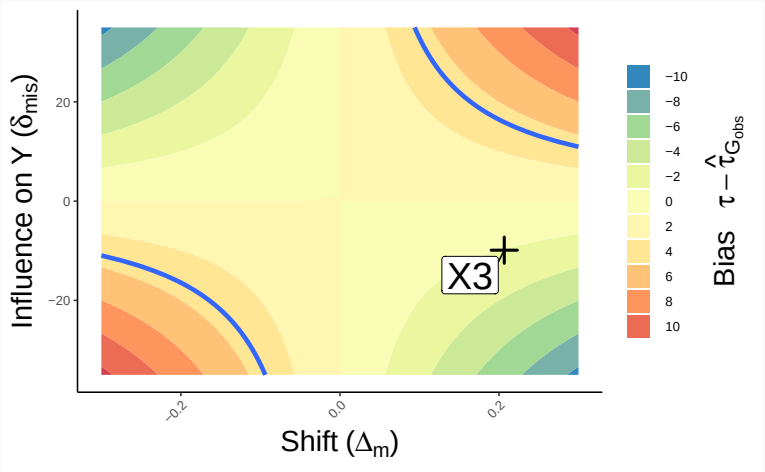


Figure 2: Austen plots: Heatmap showing the landscape and sign of the bias.

Sensitivity analysis when missing in the RCT

	Set	S	Covariates X_1	X_2	X_3	Treat W	Outcomes Y
1	$\mathcal{R}$	1	1.1	20	NA	1	24.1
	$\mathcal{R}$	1	-6	45	NA	0	26.3
n	$\mathcal{R}$	1	0	15	NA	1	23.5
$n+1$	$\mathcal{O}$	?	-1	35	7.1		
$n+2$	$\mathcal{O}$	?	-2	52	2.4		
	$\mathcal{O}$	?		...			
$n+m$	$\mathcal{O}$	?	-2	22	3.4		

$$B = \underbrace{\delta_{mis}}_{X_{mis}\text{'s strength}} \Big(\underbrace{\mathbb{E}[X_{mis}] - \mathbb{E}[X_{mis} \mid S=1]}_{\text{Shift of } X_{mis}\text{: } \Delta_m} - \underbrace{\Sigma_{mis,obs}\Sigma_{obs,obs}^{-1}(\mathbb{E}[X_{obs}] - \mathbb{E}[X_{obs} \mid S=1])}_{\text{Can be estimated from the data}} \Big)$$

No additional assumption on independence

- ▷ Define range for plausible δ_{mis} and Δ_m values
- ▷ Estimate $\Sigma_{obs,obs}$, $\Sigma_{mis,obs}$, and $\mathbb{E}[X_{obs}]$ on the observational dataset
- ▷ Estimate $\mathbb{E}[X_{obs} \mid S=1]$ on the RCT dataset
- ▷ Compute all possible bias B for range of δ_{mis} & Δ_m and return austeen plot

Hints for δ_{miss}: impute the missing variable and estimate δ_{miss}

Sensitivity analysis when missing in the observational data[7]

	Set	S	Covariates X_1	X_2	X_3	Treat W	Outcomes Y
1	$\mathcal{R}$	1	1.1	20	34	1	24.1
	$\mathcal{R}$	1	-6	45	12	0	26.3
n	$\mathcal{R}$	1	0	15	10	1	23.5
$n+1$	$\mathcal{O}$	?	-1	35	NA		
$n+2$	$\mathcal{O}$	?	-2	52	NA		
	$\mathcal{O}$	?		...			
$n+m$	$\mathcal{O}$	?	-2	22	NA		

$$\begin{aligned}\tau &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[g(X) + W\langle X, \delta\rangle \mid W = 1] - \mathbb{E}[g(X) + W\langle X, \delta\rangle \mid W = 0] \\ &= \langle \delta, \mathbb{E}[X]\rangle = \langle \delta_{obs}, \mathbb{E}[X_{obs}]\rangle + \langle \delta_{mis}, \underbrace{\mathbb{E}[X_{mis}]}_{\text{Unknown}}\rangle\end{aligned}$$

No assumption

- ▷ Define range for plausible $\mathbb{E}[X_{mis}]$ values
- ▷ Estimate δ with Robinson procedure (residuals on residuals) on the RCT [5] [6]
- ▷ Estimate $\mathbb{E}[X_{obs}]$ on the observational dataset
- ▷ Compute all possible bias for range of $\mathbb{E}[X_{mis}]$ and return austen plot

[5] Robinson, P. 1988, Root- N-Consistent Semiparametric Regression, *Econometrica*

[6] Nie, X & Wager, S. 2020, Quasi-Oracle Estimation of Heterogeneous Treatment, *Biometrika*

[7] Nguyen, et al. (2018), Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial, *PLOS ONE*

Student/teacher achievement ratio (STAR) data

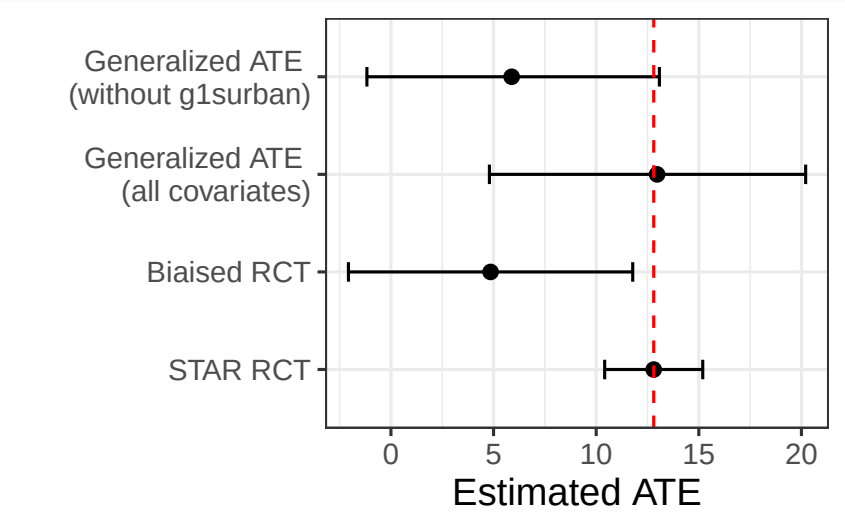
- RCT with 5000 students: effect of class size on grades
- Difference-in-means estimator (taken as true): $\tau = +12.80$ points on grades

Semi-synthetic simulation

- generate biased RCT based on `g1surban` ($n = 560$) $\hat{\tau}_1 = 4.85$
- generate a representative sample observational data ($m = 500$)
- generalize from RCT to obs. data with g-estimator & linear/random forests

$$\hat{\tau}_{g,n,m} = \frac{1}{m} \sum_{i=n+1}^{n+m} \left(\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i) \right),$$

with $\mu_{w,1}(x) = \mathbb{E}\left[Y \mid X = x, S = 1, W = w\right]$

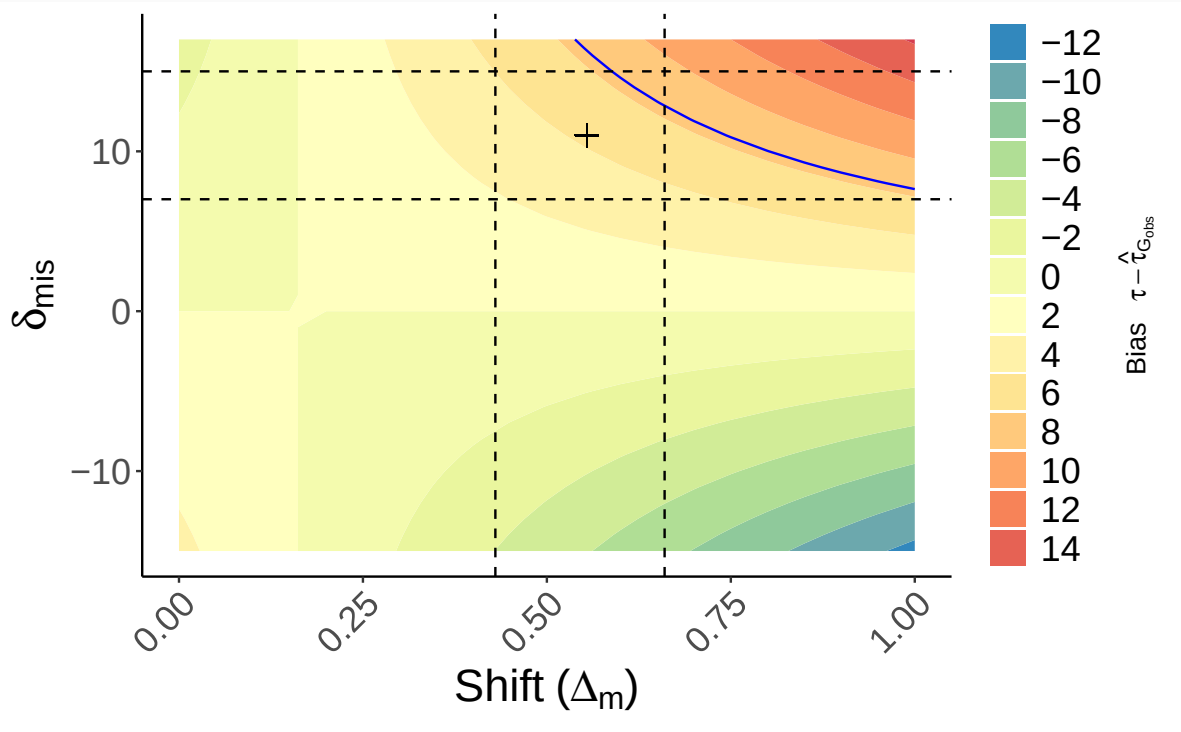


Sensitivity analysis on STAR

- Suppose `g1surban` observed in obs. data but not in RCT
- Ommiting this variable: bias (≈ 7) when generalizing the treatment effect

▷ Δ_m given by domain expert (interpretable: shift in children proportion leaving in suburbs versus city center)

▷ For δ_{mis} we impute the RCT, and model the outcome as a function of observed and imputed covariates



Motivating example

Question from physicians

Can we estimate the average effect of Tranexamic Acid (TXA) on brain-injured death on the French population in trauma centers?

CRASH3

- Multi-centric RCT over 29 countries
- No effect of TXA with difference in means (-0.3 with [95% CI -0.8 0.2])

Traumabase

- Representative sample
- 8200 patients with TBI

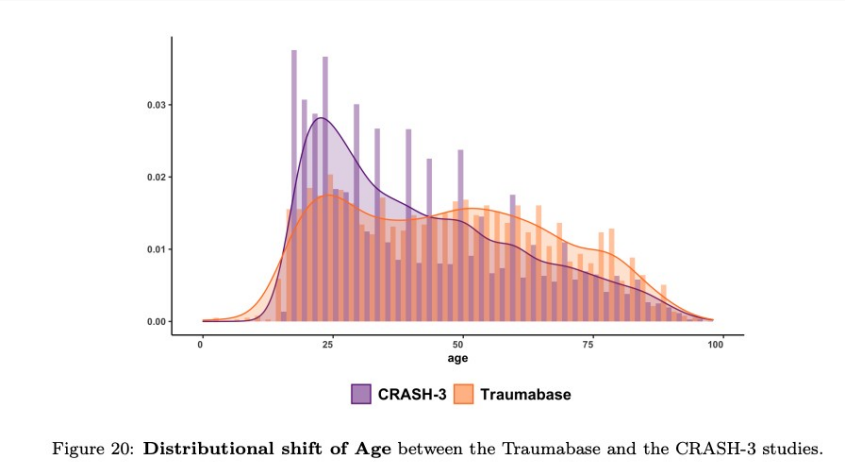


Figure 20: **Distributional shift of Age** between the Traumabase and the CRASH-3 studies.

Motivating example

Question from physicians

Can we estimate the average effect of Tranexamic Acid (TXA) on brain-injured death on the French population in trauma centers?

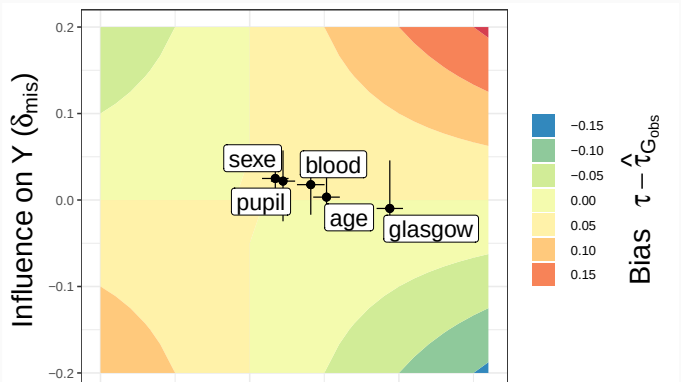
CRASH3

- Multi-centric RCT over 29 countries
- No effect of TXA with difference in means (-0.3 with [95% CI -0.8 0.2])

Traumabase

- Representative sample
- 8200 patients with TBI

ATE = -0.035, 95% CI [-0.38 0.28] when generalizing with g-estimator.
Treatment effect modifiers "time to treatment" is missing in Traumabase



Conclusion

Conclusion

Contributions

- Consistency of generalization estimators
- **Sensitivity analysis for generalization task**: **range of bias due to (partially) missing covariates**
- Semi-parametric model, no model on the sampling selection (S)
- Bias when imputing missing covariates & using a proxy

Conclusion

Contributions

- Consistency of generalization estimators
- **Sensitivity analysis for generalization task**: **range of bias due to (partially) missing covariates**
- Semi-parametric model, no model on the sampling selection (S)
- Bias when imputing missing covariates & using a proxy

Ongoing work

- ▷ Application on many RCTs and obs data
- ▷ $\{Y(0), Y(1)\} \perp\!\!\!\perp S \mid X$: covariates treatment effect modifier & shifted
  - Variance of the estimators when additional variables related to Y ?
  - Rate of convergence?

Limits: Difficult to give a priori values for sensitivity parameters,
Semi-parametric model (linear CATE), shift on means,
Austen plot for one missing variable

(Partially) missing covariates & sporadically missing covariates