Combining randomized and observational data

Toward new clinical evidence?

Bénédicte Colnet, Ph.D. student at Inria, Soda & PreMeDICaL teams

Tuesday, October 13th

Causal τ working group's seminars

My PhD advisors

Julie JOSSE
Senior Researcher
Inria

Missing values, causal inference

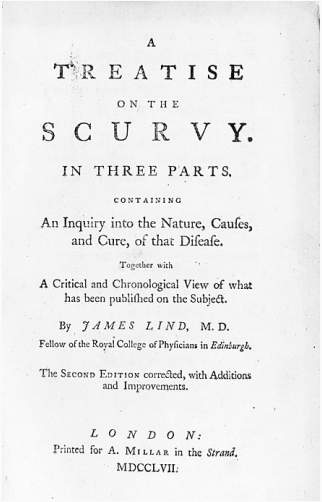
Erwan SCORNET
Associate professor
École Polytechnique

Random forests, missing values

Gaël VAROQUAUX
Research director
Inria

Co-founder of scikit-learn,
Machine-Learning

A rather old question



A

TREATISE

ON THE

SCURVY.

IN THREE PARTS.

CONTAINING

An Inquiry into the Nature, Cauſes, and Cure, of that Diſeaſe.

Together with

A Critical and Chronological View of what has been publiſhed on the Subject.

By *JAMES LIND*, M. D.

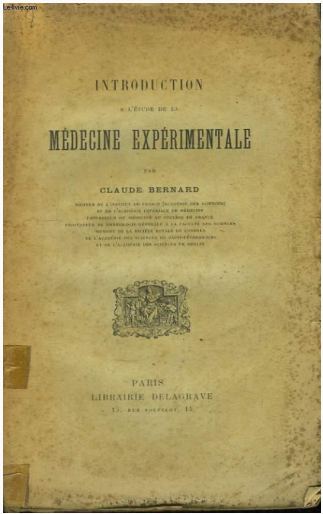
Fellow of the Royal College of Phyſicians in *Edinburgh*.

The SECOND EDITION corrected, with Additions and Improvements.

LONDON:

Printed for A. MILLAR in the *Strand*.

MDCCLVII.



Source of pictures: Wikipedia (left) & AbeBook (right)

Current practice: Randomized Controlled Trials (RCTs for short)

A longstanding presence of RCTs . . . now being **the** gold-standard

For e.g. in the 16^{th} century a cross-over trial has been documented about rhubarb's effect. **Source: The Conversation - Wellcome Collection, CC BY**

Drug Trials Snapshot	Active Ingredient	Date of FDA Approval	What is it Approved For
CABENUVA	cabotegravir and rilpivirine	January 20, 2021	Treatment of HIV-1 infection.
LUPKYNIS	voclosporin	January 22, 2021	Treatment of lupus nephritis
VERQUVO	vericiguat	January 19, 2021	Treatment of chronic heart failure
GEMTESA	vibegron	December 23, 2020	Treatment of symptoms of overactive bladder
EBANGA	ansuvimab-zykl	December 21, 2020	Treatment of Zaire ebolavirus infection
ORGOVYX	relugolix	December 18, 2020	Treatment of advanced prostate cancer

Recently approved drugs by the Food and Drug Administration (FDA), all with their corresponding RCT snapshot and information. **Source**: `www.fda.gov`

James Lind's experiment formalization: Who?

From left to right: Ronald Fisher, Jerzy Neyman, and Egon Pearson

Source: towardsdatascience website - What can an Octopus tell us?

James Lind's experiment formalization: What?

This slide is an introduction to the Potential Outcome framework.

Assume your goal is to measure the effect of a drug on an outcome.

James Lind's experiment formalization: What?

This slide is an introduction to the Potential Outcome framework.

Assume your goal is to measure the effect of a drug on an outcome.

Using the potential outcome framework (Neyman, 1923), we denote

- 💊 A the treatment,
- 🩺 X the covariates,
- 🌡️ Y the **observed** outcome.

For each individual i, consider each of the possible outcomes, as if we consider counterfactual worlds, $Y_i^{(1)}$ (treated), and $Y_i^{(0)}$ (untreated).

James Lind's experiment formalization: What?

This slide is an introduction to the Potential Outcome framework.

Assume your goal is to measure the effect of a drug on an outcome.

Using the potential outcome framework (Neyman, 1923), we denote

- 💊 A the treatment,
- 🩺 X the covariates,
- 🌡️ Y the **observed** outcome.

For each individual i, consider each of the possible outcomes, as if we consider counterfactual worlds, $Y_i^{(1)}$ (treated), and $Y_i^{(0)}$ (untreated).

$$Y_i^{(1)} \stackrel{?}{=} Y_i^{(0)}$$

Data at hand

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

Problem: Δ_i never observed (only observe one outcome/indiv). Causal inference as a missing value problem?

Covariates			Treatment	Outcome(s)		Observed outcome
X_1	X_2	X_3	A	$Y^{(0)}$	$Y^{(1)}$	Y
1.1	20	F	1	**NA**	T	T
-6	45	F	0	F	**NA**	F
0	15	M	1	**NA**	F	F
	...		...	...	...	...
-2	52	M	0	T	**NA**	T

Data at hand

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

Problem: Δ_i never observed (only observe one outcome/indiv). Causal inference as a missing value problem?

Covariates			Treatment	Outcome(s)		Observed outcome
X_1	X_2	X_3	A	$Y^{(0)}$	$Y^{(1)}$	Y
1.1	20	F	1	**NA**	T	T
-6	45	F	0	F	**NA**	F
0	15	M	1	**NA**	F	F
	...		...	...	...	...
-2	52	M	0	T	**NA**	T

💡 Two sources of randomness in this data set:

- Treatment assignment allocation,
- Sampling individuals in a wider population.

Randomized Controlled Trial: an empirical trick to measure the causal effect

Statistical trick: Inference on potential outcomes' distributions.

$$\mathbb{E}\left[Y^{(1)}\right] \stackrel{?}{=} \mathbb{E}\left[Y^{(0)}\right].$$

Randomized Controlled Trial: an empirical trick to measure the causal effect

Statistical trick: Inference on potential outcomes' distributions.

$$\mathbb{E}\left[Y^{(1)}\right] \stackrel{?}{=} \mathbb{E}\left[Y^{(0)}\right].$$

More precisely people often target the so-called Average Treatment Effect (ATE),

$$\tau = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right].$$

Running a randomized controlled trial is a way to ensure,

Assumption - Treatment assignment exchangeability

$$\forall i, \quad Y_i^{(1)}, Y_i^{(0)} \perp\!\!\!\perp A_i,$$

💡 *Treated and control groups differ only with respect to treatment allocation.*

Another assumption we will assume today is the SUTVA assumption: no interference and consistency $Y_i(A_1, A_2, \ldots, A_n) = Y_i(A_i)$.

Statistical properties of the difference-in-means

Suppose we have access to n independent and identically distributed examples labeled $i = 1, \ldots, n$, a response $Y_i \in \mathcal{Y}$, and a binary treatment indicator $A_i \in \{0, 1\}$ assigned randomly.

Definition - Horvitz-Thomson

$$\hat{\tau}_{\mathrm{HT}} = \frac{1}{n_1} \sum_{A_i=1} Y_i - \frac{1}{n_0} \sum_{A_i=0} Y_i \quad \text{, where } n_a = |\{i : A_i = a\}| \,,$$

Proposition - Asymptotically normal estimator

The difference-in-means estimator is asymptotically normal,

$$\sqrt{n} \left(\hat{\tau}_{\mathrm{HT}} - \tau \right) \xrightarrow{d} \mathcal{N} \left(0, \sigma^2_{\mathrm{HT}} \right),$$

where

$$\sigma^2_{\mathrm{HT}} = \frac{\mathbb{E}\left[\left(Y^{(1)} \right)^2 \right]}{\pi} + \frac{\mathbb{E}\left[\left(Y^{(0)} \right)^2 \right]}{1 - \pi} - \tau^2.$$

Bonus: $\hat{\tau}_{\mathrm{HT}}$ is an unbiased estimator.

But, the limited scope of RCTs is increasingly under scrutiny

- Short timeframe,
- unrealistic real-world compliance,
- limited sample size,
- unrepresentative sample.

But, the limited scope of RCTs is increasingly under scrutiny

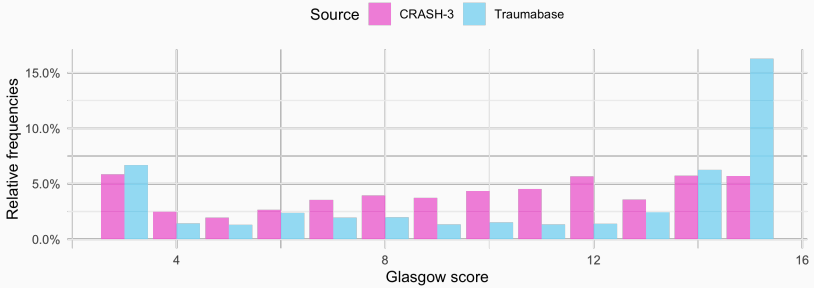
- Short timeframe,
- unrealistic real-world compliance,
- limited sample size,
- unrepresentative sample.

Can the result of a large international trial – assessing the efficacy of Tranexamic Acid (TXA) on brain-injured death (TBI) – be **generalized** to the French population?

But, the limited scope of RCTs is increasingly under scrutiny

- Short timeframe,
- unrealistic real-world compliance,
- limited sample size,
- unrepresentative sample.

Can the result of a large international trial – assessing the efficacy of Tranexamic Acid (TXA) on brain-injured death (TBI) – be **generalized** to the French population?



Source: CRASH3 data trial and Traumabase cohort data comparing patients suffering from Traumatic Brain Injuries, and in particular their Glasgow score (severity of the trauma).

Question from clinicians[a]

[a] www.traumabase.eu

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

Question from clinicians[a]

[a] www.traumabase.eu

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

Data sources and evidence at hand:

CRASH3

- Multi-centric RCT over 29 counties,
- $\sim$ 9 000 individuals,
- High **internal** validity
- Measured a positive effect of TXA on moderate injured patients

Traumabase

- Observational sample,
- $\sim$ 30 000 individuals,
- High **external** validity
- Observational analysis can not reject the null hypothesis of no effect (and pushing toward negative effect).

Question from clinicians[a]

[a] www.traumabase.eu

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

Data sources and evidence at hand:

CRASH3

- Multi-centric RCT over 29 counties,
- $\sim$ 9 000 individuals,
- High **internal** validity
- Measured a positive effect of TXA on moderate injured patients

Traumabase

- Observational sample,
- $\sim$ 30 000 individuals,
- High **external** validity
- Observational analysis can not reject the null hypothesis of no effect (and pushing toward negative effect).

Is there a paradox?

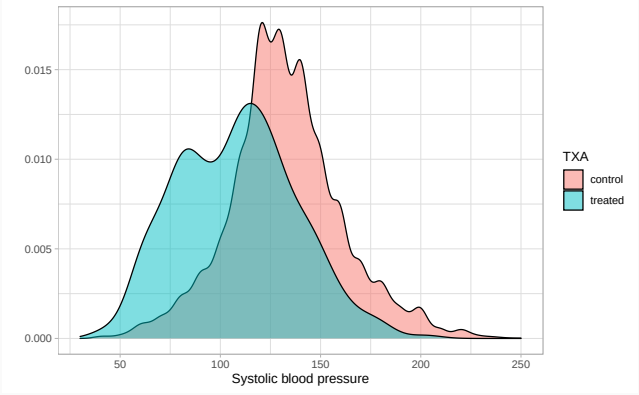
Effects of tranexamic acid on death, disability, vascular occlusive events and other morbidities in patients with acute traumatic brain injury (CRASH-3): a randomised, placebo-controlled trial

*The CRASH-3 trial collaborators**

Results Between July 20, 2012, and Jan 31, 2019, we randomly allocated 12 737 patients with TBI to receive tranexamic acid (6406 [50·3%] or placebo [6331 [49·7%], of whom 9202 (72·2%) patients were treated within 3 h of injury. Among patients treated within 3 h of injury, the risk of head injury-related death was 18·5% in the tranexamic acid group versus 19·8% in the placebo group (855 *vs* 892 events; risk ratio [RR] 0·94 [95% CI 0·86–1·02]). In the prespecified sensitivity analysis that excluded patients with a GCS score of 3 or bilateral unreactive pupils at baseline, the risk of head injury-related death was 12·5% in the tranexamic acid group versus 14·0% in the placebo group (485 *vs* 525 events; RR 0·89 [95% CI 0·80–1·00]). The risk of head injury-related death reduced with tranexamic acid in patients with mild-to-moderate head injury (RR 0·78 [95% CI 0·64–0·95]) but not in patients with severe head injury (0·99 [95% CI 0·91–1·07]; p value for heterogeneity 0·030). Early treatment was more effective than was later treatment in patients with mild and moderate head injury (p=0·005) but time to treatment had no obvious effect in patients with severe head injury (p=0·73). The risk of vascular occlusive events was similar in the tranexamic acid and placebo groups (RR 0·98 (0·74–1·28). The risk of seizures was also similar between groups (1·09 [95% CI 0·90–1·33]).

Observational data

Non-experimental studies – called Observational data – are often **confounded**, meaning that treated patients are not exactly like untreated ones.



In other words, the conditional independence does no longer hold, $\mathbb{E}[Y \mid A = a] \neq \mathbb{E}[Y^{(a)}]$.

Combining data for generalizability or transportability

Introduction to the notations

Using the potential outcome framework[1], we denote

- 💊 A the treatment,
- 🩺 X the covariates,
- 🌡️ Y the **observed** outcome.

[1] $Y_i^{(a)}$ is the potential outcome, would the individual i have received treatment a. (Neyman, 1923)

Introduction to the notations

Using the potential outcome framework[1], we denote

- 💊 A the treatment,
- 🩺 X the covariates,
- 🌡️ Y the **observed** outcome.

Two data sources:

- A trial of size n with $p_R(x)$ the probability of observing individual with $X = x$,
- A sample of the target population of interest – for e.g. a national cohort (resp. m and $p_T(x)$).

[1] $Y_i^{(a)}$ is the potential outcome, would the individual i have received treatment a. (Neyman, 1923)

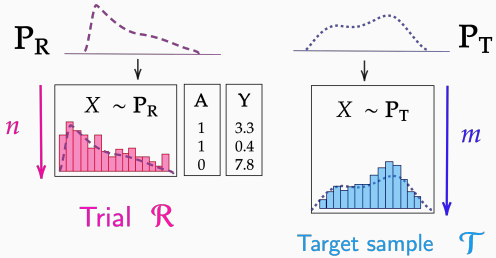
Introduction to the notations

Using the potential outcome framework[1], we denote

- 💊 A the treatment,
- 🩺 X the covariates,
- 🌡️ Y the **observed** outcome.

Two data sources:

- A trial of size n with $p_R(x)$ the probability of observing individual with $X = x$,
- A sample of the target population of interest – for e.g. a national cohort (resp. m and $p_T(x)$).



[1] $Y_i^{(a)}$ is the potential outcome, would the individual i have received treatment a. (Neyman, 1923)

What if we only use the trial to estimate the Average Treatment Effect (ATE)?

Compute ATE averaging over the trial sample:

$$\hat{\tau}_{\mathrm{HT},n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right),$$

- where π is the probability to receive treatment in the trial (usually 0.5),
- Unbiased and consistent estimator of the average effect of treatment on population P_{R}.

What if we only use the trial to estimate the Average Treatment Effect (ATE)?

Compute ATE averaging over the trial sample:

$$\hat{\tau}_{\text{HT},n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right),$$

- where π is the probability to receive treatment in the trial (usually 0.5),
- Unbiased and consistent estimator of the average effect of treatment on population P_{R}.

But, because distributions are different between the trial and the target population,

$$p_{\text{R}}(x) \neq p_{\text{T}}(x) \Rightarrow \quad \underbrace{\tau_{\text{R}} := \mathbb{E}_{\text{R}}[Y^{(1)} - Y^{(0)}]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}_{\text{T}}[Y^{(1)} - Y^{(0)}] := \tau}_{\text{Target ATE}}$$

What if we only use the trial to estimate the Average Treatment Effect (ATE)?

Compute ATE averaging over the trial sample:

$$\hat{\tau}_{\mathrm{HT},n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right),$$

- where π is the probability to receive treatment in the trial (usually 0.5),
- Unbiased and consistent estimator of the average effect of treatment on population P_{R}.

But, because distributions are different between the trial and the target population,

$$p_{\mathrm{R}}(x) \neq p_{\mathrm{T}}(x) \Rightarrow \quad \underbrace{\tau_{\mathrm{R}} := \mathbb{E}_{\mathrm{R}}[Y^{(1)} - Y^{(0)}]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}_{\mathrm{T}}[Y^{(1)} - Y^{(0)}] := \tau}_{\text{Target ATE}}$$

Re-weighting the trial's data?

$$\hat{\tau}_{\mathrm{IPSW}} := \frac{1}{n} \sum_{i \in \mathcal{R}} w(X_i) \underbrace{\left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right)}_{\text{Horvitz-Thomson.}}$$

What if we only use the trial to estimate the Average Treatment Effect (ATE)?

Compute ATE averaging over the trial sample:

$$\hat{\tau}_{\text{HT},n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right),$$

- where π is the probability to receive treatment in the trial (usually 0.5),
- Unbiased and consistent estimator of the average effect of treatment on population P_{R}.

But, because distributions are different between the trial and the target population,

$$p_{\text{R}}(x) \neq p_{\text{T}}(x) \Rightarrow \quad \underbrace{\tau_{\text{R}} := \mathbb{E}_{\text{R}}[Y^{(1)} - Y^{(0)}]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}_{\text{T}}[Y^{(1)} - Y^{(0)}] := \tau}_{\text{Target ATE}}$$

Re-weighting the trial's data?

$$\hat{\tau}_{\text{IPSW}} := \frac{1}{n} \sum_{i \in \mathcal{R}} w(X_i) \underbrace{\left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right)}_{\text{Horvitz-Thomson.}}$$

$\Longrightarrow$ *Inverse Propensity Sampling Weighting* (IPSW) - Stuart et al. 2010.

What if we only use the trial to estimate the Average Treatment Effect (ATE)?

Compute ATE averaging over the trial sample:

$$\hat{\tau}_{\text{HT},n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right),$$

- where π is the probability to receive treatment in the trial (usually 0.5),
- Unbiased and consistent estimator of the average effect of treatment on population P_{R}.

But, because distributions are different between the trial and the target population,

$$p_{\text{R}}(x) \neq p_{\text{T}}(x) \Rightarrow \quad \underbrace{\tau_{\text{R}} := \mathbb{E}_{\text{R}}[Y^{(1)} - Y^{(0)}]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}_{\text{T}}[Y^{(1)} - Y^{(0)}] := \tau}_{\text{Target ATE}}$$

Re-weighting the trial's data?

$$\hat{\tau}_{\text{IPSW}} := \frac{1}{n} \sum_{i \in \mathcal{R}} w(X_i) \underbrace{\left(\frac{Y_i A_i}{\pi} - \frac{Y_i (1 - A_i)}{1 - \pi} \right)}_{\text{Horvitz-Thomson.}}$$

$\Longrightarrow$ *Inverse Propensity Sampling Weighting* (IPSW) - Stuart et al. 2010.
📖 *Other methods exist.*

Intuitions for the methods 🖊

Generalization's *causal* assumptions.

Re-weight, so that the trial follows the target sample's distribution,

$$w(X) := \frac{p_T(X)}{p_R(X)}.$$

Generalization's *causal* assumptions.

Re-weight, so that the trial follows the target sample's distribution,

$$w(X) := \frac{p_T(X)}{p_R(X)}.$$

Which assumptions?

Transportability

$$\forall x \in X,\ \mathbb{P}_R(Y^{(1)} - Y^{(0)} \mid X = x) = \mathbb{P}_T(Y^{(1)} - Y^{(0)} \mid X = x).$$

i.e. Needed covariates to re-weight correspond to shifted treatment effect modifier covariates (along the absolute scale).

Support inclusion

$$\mathsf{supp}(P_T(X)) \subset \mathsf{supp}(P_R(X))$$

i.e. Each individuals in the target population has to be represented in the trial.

State-of-the-art and open practical questions

State-of-the-art

- Re-weighting can be found back in the early 2000's;
  $\Longrightarrow$ see books in epidemiology, under the name *standardization*
- But the idea of relying on an external representative sample is recent;
  $\Longrightarrow$ in particular seminal articles can be found in the early 2010's[2] and is getting more and more popular[3]
- Since, other approaches than IPSW have been proposed
  $\Longrightarrow$ outcome-modeling (G-formula), balancing, doubly-robust approaches, . . .

[2] Stephen R. Cole, Elizabeth A. Stuart. (2010) Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial, *American Journal of Epidemiology*

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

State-of-the-art and open practical questions

State-of-the-art

- Re-weighting can be found back in the early 2000's;
  $\Longrightarrow$ see books in epidemiology, under the name *standardization*
- But the idea of relying on an external representative sample is recent;
  $\Longrightarrow$ in particular seminal articles can be found in the early 2010's[2] and is getting more and more popular[3]
- Since, other approaches than IPSW have been proposed
  $\Longrightarrow$ outcome-modeling (G-formula), balancing, doubly-robust approaches, . . .

In practice, open questions remain

- What is the impact of the two data sources' sizes n and m?
- Which covariates should we use?

[2] Stephen R. Cole, Elizabeth A. Stuart. (2010) Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial, *American Journal of Epidemiology*

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

State-of-the-art and open practical questions

State-of-the-art

- Re-weighting can be found back in the early 2000's;
  $\Longrightarrow$ see books in epidemiology, under the name *standardization*
- But the idea of relying on an external representative sample is recent;
  $\Longrightarrow$ in particular seminal articles can be found in the early 2010's[2] and is getting more and more popular[3]
- Since, other approaches than IPSW have been proposed
  $\Longrightarrow$ outcome-modeling (G-formula), balancing, doubly-robust approaches, . . .

In practice, open questions remain

- What is the impact of the two data sources' sizes n and m?
- Which covariates should we use?

For the rest of the work, we assume X is composed of categorical covariates

$\Longrightarrow$ for e.g. gender, smoking status, Glasgow score, insurance status, . . .

[2] Stephen R. Cole, Elizabeth A. Stuart. (2010) Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial, *American Journal of Epidemiology*

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

Theoretical study of IPSW

Theoretical guarantees of IPSW with oracle weights

True (or oracle) probabilities

$$\hat{\tau}^*_{\pi,\mathrm{T},\mathrm{R},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_{\mathrm{T}}(X_i)}{p_{\mathrm{R}}(X_i)} \; Y_i\left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right),$$

Theoretical guarantees of IPSW with oracle weights

True (or oracle) probabilities

$$\hat{\tau}^*_{\pi,\mathrm{T},\mathrm{R},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_{\mathrm{T}}(X_i)}{p_{\mathrm{R}}(X_i)} Y_i \left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right),$$

Properties

$$\mathbb{E}\left[\hat{\tau}^*_{\pi,\mathrm{T},\mathrm{R},n}\right] = \tau, \text{ and } \mathbf{Var}\left[\hat{\tau}^*_{\pi,\mathrm{T},\mathrm{R},n}\right] = \frac{V_{\text{oracle}}}{n},$$

where

$$V_{\text{oracle}} := \mathbf{Var}_{\mathrm{R}}\left[\frac{p_{\mathrm{T}}(X)}{p_{\mathrm{R}}(X)}\tau(X)\right] + \mathbb{E}_{\mathrm{R}}\left[\left(\frac{p_{\mathrm{T}}(X)}{p_{\mathrm{R}}(X)}\right)^2 V_{\mathrm{HT}}(X)\right].$$

$\tau(x)$ being the effect of treatment on strata $X = x$.

How do we estimate weights in practice?

$$\hat{\tau}^*_{\pi,\mathrm{T},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_{\mathrm{T}}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \; Y_i\left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right),$$

Estimated with $\mathcal{R}$

Estimation is intuitive, and corresponds to how many times the specific combinaison of category x appears in the trial, that is

$$\hat{p}_{\mathrm{R},n}(x) := \frac{1}{n}\sum_{i\in\mathcal{R}} 1_{X_i = x}$$

How do we estimate weights in practice?

$$\hat{\tau}^*_{\pi,\mathrm{T},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_\mathrm{T}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \; Y_i \left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right), \quad \text{with } \hat{p}_{\mathrm{R},n}(x) := \frac{1}{n}\sum_{i\in\mathcal{R}} 1_{X_i=x}.$$

Finite-sample properties - semi oracle estimate

$$\mathbb{E}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] - \tau = -\sum_{x\in X} p_\mathrm{T}(x)\left(1-p_\mathrm{R}(x)\right)^n \tau(x)$$

$$\text{and} \quad \mathrm{Var}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] \leq \frac{2V_{SO}}{n+1} + \left(1 - \min_{x\in\mathbb{X}} p_\mathrm{R}(x)\right)^n \mathbb{E}_\mathrm{T}\left[\tau(X)^2\right],$$

where

$$V_{SO} := \mathbb{E}_\mathrm{R}\left[\left(\frac{p_\mathrm{T}(X)}{p_\mathrm{R}(X)}\right)^2 V_{HT}(X)\right] = V_O - \mathrm{Var}_\mathrm{R}\left[\frac{p_\mathrm{T}(X)}{p_\mathrm{R}(X)}\tau(X)\right].$$

How do we estimate weights in practice?

$$\hat{\tau}^*_{\pi,\mathrm{T},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_\mathrm{T}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \; Y_i\left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right), \quad \text{with } \hat{p}_{\mathrm{R},n}(x) := \frac{1}{n}\sum_{i\in\mathcal{R}} 1_{X_i=x}.$$

Finite-sample properties - semi oracle estimate

$$\mathbb{E}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] - \tau = -\sum_{x\in X} p_\mathrm{T}(x)\left(1-p_\mathrm{R}(x)\right)^n \tau(x)$$

$$\text{and} \quad \mathrm{Var}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] \leq \frac{2V_{SO}}{n+1} + \left(1 - \min_{x\in\mathbb{X}} p_\mathrm{R}(x)\right)^n \mathbb{E}_\mathrm{T}\left[\tau(X)^2\right],$$

where

$$V_{SO} := \mathbb{E}_\mathrm{R}\left[\left(\frac{p_\mathrm{T}(X)}{p_\mathrm{R}(X)}\right)^2 V_{HT}(X)\right] = V_O - \mathrm{Var}_\mathrm{R}\left[\frac{p_\mathrm{T}(X)}{p_\mathrm{R}(X)}\tau(X)\right].$$

- Positive but exponentially small bias compared to the oracle estimate
  → undercoverage of some categories in the trial

How do we estimate weights in practice?

$$\hat{\tau}^*_{\pi,\mathrm{T},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_\mathrm{T}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \; Y_i\left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right), \quad \text{with } \hat{p}_{\mathrm{R},n}(x) := \frac{1}{n}\sum_{i\in\mathcal{R}} 1_{X_i=x}.$$

Finite-sample properties - semi oracle estimate

$$\mathbb{E}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] - \tau = -\sum_{x\in X} p_\mathrm{T}(x)\left(1-p_\mathrm{R}(x)\right)^n \tau(x)$$

$$\text{and} \quad \mathrm{Var}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] \leq \frac{2V_{SO}}{n+1} + \left(1 - \min_{x\in\mathbb{X}} p_\mathrm{R}(x)\right)^n \mathbb{E}_\mathrm{T}\left[\tau(X)^2\right],$$

where

$$V_{SO} := \mathbb{E}_\mathrm{R}\left[\left(\frac{p_\mathrm{T}(X)}{p_\mathrm{R}(X)}\right)^2 V_{\mathrm{HT}}(X)\right] = V_O - \mathrm{Var}_\mathrm{R}\left[\frac{p_\mathrm{T}(X)}{p_\mathrm{R}(X)}\tau(X)\right].$$

- Positive but exponentially small bias compared to the oracle estimate
  → undercoverage of some categories in the trial
- Smaller variance than the oracle estimate
  → Estimating a denominator can be more efficient than using the oracle.

Theoretical guarantees of IPSW with completely estimated weights

Estimated with $\mathcal{T}$

$$\hat{\tau}_{\pi,n,m} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{\hat{p}_{\mathrm{T},m}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \, Y_i \left(\frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right),$$

Estimated with $\mathcal{R}$

Asymptotic properties

Letting $\lim_{n,m \to \infty} m/n = \lambda \in [0, \infty]$,

$$\lim_{n,m \to \infty} \min(n, m) \operatorname{Var}\left[\hat{\tau}_{\pi,n,m}\right] = \min(1, \lambda) \left(\frac{\operatorname{Var}\left[\tau(X)\right]}{\lambda} + V_{so} \right).$$

Variance depends on the size of the two data sets, n and m

What if also estimating π?

$$\hat{\tau}^*_{n,m} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{\hat{p}_{\mathrm{T},m}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \; Y_i \left(\frac{Y_i A_i}{\hat{\pi}_n(X)} - \frac{Y_i(1-A_i)}{1-\hat{\pi}_n(X)} \right),$$

Asymptotic properties

Letting $\lim_{n,m \to \infty} m/n = \lambda \in [0, \infty]$,

$$\lim_{n,m \to \infty} \min(n,m) \operatorname{Var}[\hat{\tau}_{n,m}] = \min(1,\lambda) \left(\frac{\operatorname{Var}[\tau(X)]}{\lambda} + \tilde{V}_{so} \right),$$

where

$$\tilde{V}_{so} \leq V_{so}.$$

Variance is smaller if also estimating π with the data

💡 This phenomenon is the same as the Difference-in-Means having better precision than the Horvitz-Thomson on a trial.

Impact of additional covariates

Impact of additional covariates: for the worse

Covariates needed to generalize are,

- Treatment effect modifier
  a covariate along which the treatment effect is modulated;
- Shifted
  not the same proportion in each population.

Impact of additional covariates: for the worse

Covariates needed to generalize are,

- Treatment effect modifier
  a covariate along which the treatment effect is modulated;
- Shifted
  not the same proportion in each population.

But in practice,
one may be tempted to add as many covariates as possible:

- It does prevent to miss important ones;

Impact of additional covariates: for the worse

Covariates needed to generalize are,

- Treatment effect modifier
  a covariate along which the treatment effect is modulated;
- Shifted
  not the same proportion in each population.

But in practice,
one may be tempted to add as many covariates as possible:

- It does prevent to miss important ones;
- But what happen if gender is added, but is only shifted?

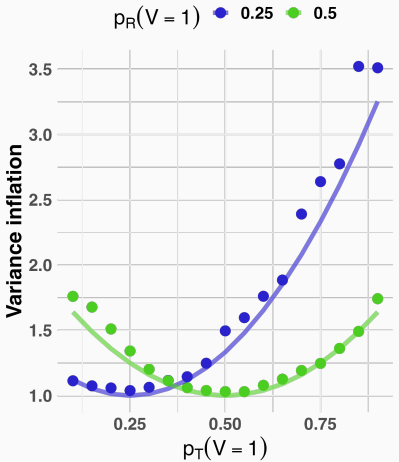
Impact of additional covariates: for the worse

Covariates needed to generalize are,

- Treatment effect modifier
  a covariate along which the treatment effect is modulated;
- Shifted
  not the same proportion in each population.

But in practice,
one may be tempted to add as many covariates as possible:

- It does prevent to miss important ones;
- But what happen if gender is added, but is only shifted?



Plot showing the impact of adding a non-necessary covariates V when generalizing. Plain lines are the theory, and dots the simulations

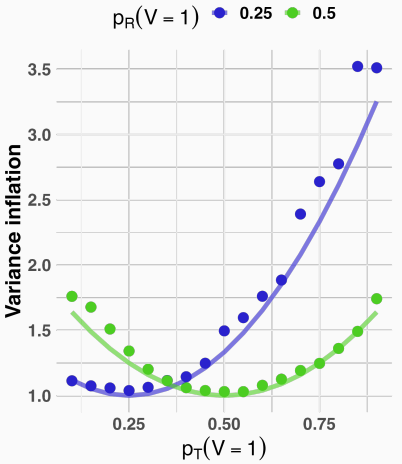
Impact of additional covariates: for the worse

Covariates needed to generalize are,

- Treatment effect modifier
  a covariate along which the treatment effect is modulated;
- Shifted
  not the same proportion in each population.

But in practice,
one may be tempted to add as many covariates as possible:

- It does prevent to miss important ones;
- But what happen if gender is added, but is only shifted?



Plot showing the impact of adding a non-necessary covariates V when generalizing. Plain lines are the theory, and dots the simulations

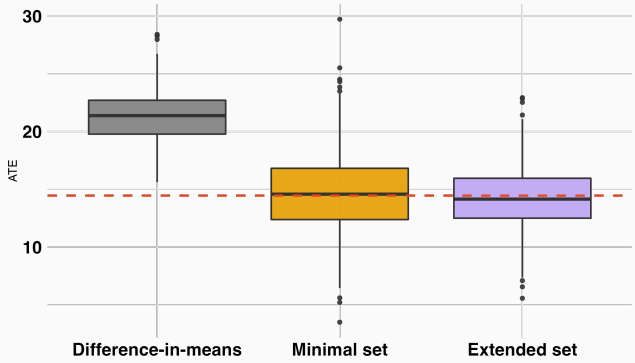
(i) Including non-necessary covariates can seriously damage precision!

Impact of additional covariates: for the worse, and the better

What happen if a non-shifted covariate, known to be treatment effect modifier, is added?

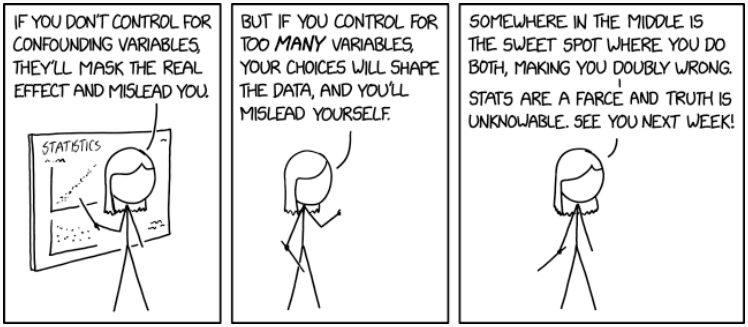
Impact of additional covariates: for the worse, and the better

What happen if a non-shifted covariate, known to be treatment effect modifier, is added?



(ii) Adding a non-shifted, but treatment effect modifiers covariate, in the adjustment set improves precision.

N.B.: To find *X really* is a tricky task!



Semi-synthetic simulation

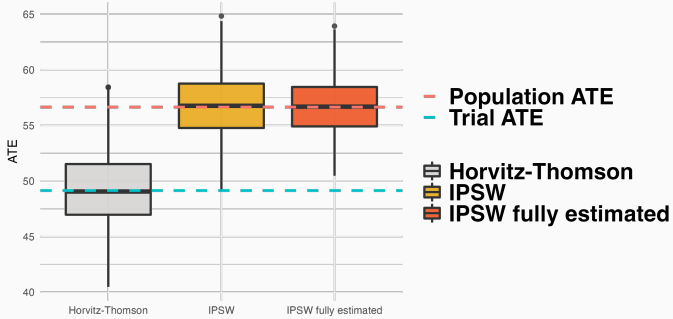
- All the results are illustrated on semi-synthetic simulations;
- Build from two large clinical data bases, reflecting a real-world situation
  - CRASH3 $\sim$ 9 000 individuals.
  - Traumabase $\sim$ 30 000 individuals.
- The outcome is the only synthetic part,

$$Y := f(\texttt{GCS}, \texttt{Gender}) + A\,\tau(\texttt{TTT}, \texttt{Blood Pressure}) + \epsilon_{\texttt{TTT}},$$

Semi-synthetic simulation

- All the results are illustrated on semi-synthetic simulations;
- Build from two large clinical data bases, reflecting a real-world situation
  - CRASH3 $\sim$ 9 000 individuals.
  - Traumabase $\sim$ 30 000 individuals.
- The outcome is the only synthetic part,

$$Y := f(\texttt{GCS}, \texttt{Gender}) + A\,\tau(\texttt{TTT}, \texttt{Blood Pressure}) + \epsilon_{\texttt{TTT}},$$



More in the main paper,

- Different asymptotic regimes,
- The re-weighted trial has not necessarily larger variance,
- Effect of adding non-necessary covariates.

Conclusion

Main idea:

- RCTs are, and will remain, cornerstones of modern-based medicine,
- But they have limits, such as a lack of representativeness,
- So-called real-world data can help strengthen clinical evidence.

Conclusion

Main idea:

- RCTs are, and will remain, cornerstones of modern-based medicine,
- But they have limits, such as a lack of representativeness,
- So-called real-world data can help strengthen clinical evidence.

For this to happen:

- We need to build new methods . . .
- . . . along with a clear understanding of the assumptions and their statistical properties.

Conclusion

Main idea:

- RCTs are, and will remain, cornerstones of modern-based medicine,
- But they have limits, such as a lack of representativeness,
- So-called real-world data can help strengthen clinical evidence.

For this to happen:

- We need to build new methods . . .
- . . . along with a clear understanding of the assumptions and their statistical properties.

In this talk:

- New theoretical properties for an intuitive method i.e. trial re-weighting
- Alongside with clear and important guidelines for users about **covariate selection**.

  $\Longrightarrow$ *Physicians and epidemiologists have an important role to play in selecting a limited number of covariates when generalizing trial's findings!*

Thank you very much for your attention!! 🌹

Appendix

Theoretical guarantees of IPSW with semi-oracle (= so) weights

$$\hat{\tau}^*_{\pi,\mathrm{T},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{p_{\mathrm{T}}(X_i)}{\hat{p}_{\mathrm{R},n}(X_i)} \; Y_i\left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right),$$

Estimated with $\mathcal{R}$

Asymptotic properties

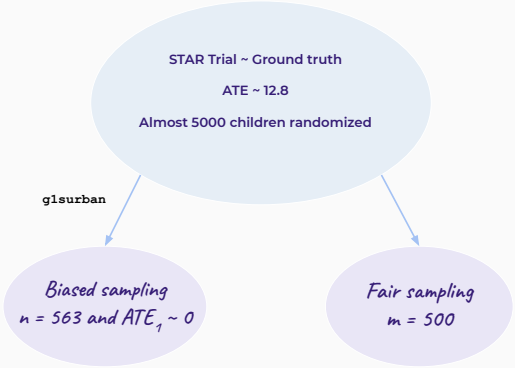
$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] = \tau, \quad \text{and} \quad \lim_{n\to\infty} n\,\mathrm{Var}\left[\hat{\tau}^*_{\pi,\mathrm{T},n}\right] = V_{\mathrm{so}} \leq V_{\mathrm{oracle}}$$

🤔 Estimating $p_{\mathrm{R}}(x)$ is more efficient than taking the oracle probability (counter-intuitive!)

Semi synthetic simulation

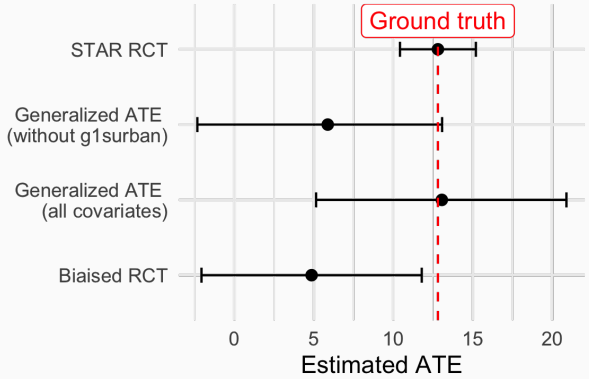
Using the data from the Tennessee Student/Teacher Achievement Ratio (STAR) study (**?**).

We generate a biased RCT sample based on covariate `g1surban` and a representative sample.



Semi synthetic simulation - Generalization with missing covariate

Bias induced is around 7 points when omitting `g1surban`.

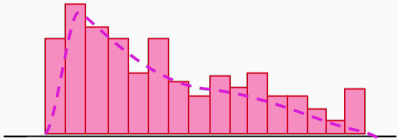


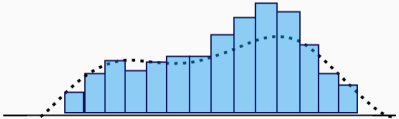
Can the sensitivity analysis estimates the bias when `g1surban` is missing in the observational data but not the RCT?

Context

Consider that a policy maker has at hand:

- an already conducted trial about a treatment or policy ($\rightarrow \hat{\tau}_1$),
- and a sample of the target population of interest ($\hat{\tau}?$).





Context

Consider that a policy maker has at hand:

- an already conducted trial about a treatment or policy ($\rightarrow \hat{\tau}_1$),
- and a sample of the target population of interest ($\hat{\tau}?$).