Combining randomized and observational data

Toward new clinical evidence?

Bénédicte Colnet, PhD student at Inria (Soda & PreMedICaL teams)
Monday, September 19th

9th International Meeting on Statistical Methods in Biopharmacy, Paris, 2022
My PhD advisors

Julie JOSSE  
Senior Researcher  
Inria  

Erwan SCORNET  
Associate professor  
École Polytechnique  

Gaël VAROQUAUX  
Research director  
Inria  

Missing values, causal inference  
Random forests, missing values  

Most of today's content comes from an ongoing work:

*Reweighting the RCT for generalization: finite sample analysis and variable selection*, Colnet et al. (2022).
A longstanding presence of RCTs ... now being the gold-standard

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

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

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

Using the potential outcome framework\(^1\), we denote

- A the treatment,
- \(X\) the covariates,
- \(Y\) the observed outcome.

\(^1\) 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

- \(\mathcal{A}\) the treatment,
- \(\mathcal{X}\) the covariates,
- \(\mathcal{Y}\) 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:

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

- where \( \pi \) 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:

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

- where \(\pi\) 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,

\[ p_R(x) \neq p_T(x) \Rightarrow \tau_R := \mathbb{E}_R[Y^{(1)} - Y^{(0)}] \neq \mathbb{E}_T[Y^{(1)} - Y^{(0)}] := \tau \]

ATE in the RCT

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

Compute ATE averaging over the trial sample:

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

- where \( \pi \) 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,

\[
p_R(x) \neq p_T(x) \Rightarrow \tau_R := E_R[Y^{(1)} - Y^{(0)}] \neq E_T[Y^{(1)} - Y^{(0)}] := \tau
\]

Re-weighting the trial’s data?

\[
\hat{\tau}_{IPSW} := \frac{1}{n} \sum_{i \in R} w(X_i) \left( \frac{Y_iA_i}{\pi} - \frac{Y_i(1 - A_i)}{1 - \pi} \right)
\]

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

Compute ATE averaging over the trial sample:

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

• where $\pi$ 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,

$$p_R(x) \neq p_T(x) \Rightarrow \tau_R := \mathbb{E}_R[Y^{(1)} - Y^{(0)}] \neq \mathbb{E}_T[Y^{(1)} - Y^{(0)}] := \tau$$

Re-weighting the trial’s data?

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

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

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

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

**Support inclusion**

\[
supp(P_T(X)) \subset supp(P_R(X))
\]

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;
  \[\Rightarrow\] see books in epidemiology, under the name *standardization*

- But the idea of relying on an external representative sample is recent;
  \[\Rightarrow\] 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
  \[\Rightarrow\] outcome-modeling (G-formula), balancing, doubly-robust approaches, . . .

---


State-of-the-art and open practical questions

State-of-the-art

- Re-weighting can be found back in the early 2000’s; 
  ⇒ see books in epidemiology, under the name standardization
- But the idea of relying on an external representative sample is recent; 
  ⇒ 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 
  ⇒ 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?

---


\(^3\)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; ⇒ see books in epidemiology, under the name standardization
- But the idea of relying on an external representative sample is recent; ⇒ 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 ⇒ 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 ⇒ for e.g. gender, smoking status, Glasgow score, insurance status, . . .

---


Theoretical guarantees of IPSW with oracle weights

True (or oracle) probabilities

\[
\hat{\tau}^{*}_{\pi,T,R,n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{p_T(X_i)}{p_R(X_i)} Y_i \left( \frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right),
\]
Theoretical guarantees of IPSW with oracle weights

True (or oracle) probabilities

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

Properties

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

where

$$V_{\text{oracle}} := \text{Var}_R \left[ \frac{p_T(X)}{p_R(X)} \tau(X) \right] + \mathbb{E}_R \left[ \left( \frac{p_T(X)}{p_R(X)} \right)^2 V_{HT}(X) \right].$$

$$\tau(x)$$ being the effect of treatment on strata $$X = x.$$
How do we estimate weights in practice?

\[
\hat{r}_{\pi, \tau, n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{p_t(X_i)}{\hat{p}_{\mathcal{R}, n}(X_i)} Y_i \left( \frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right),
\]

Estimated with \( \mathcal{R} \)

Estimation is intuitive, and corresponds to how many times the specific combination of category \( x \) appears in the trial, that is

\[
\hat{p}_{\mathcal{R}, n}(x) := \frac{1}{n} \sum_{i \in \mathcal{R}} 1_{X_i = x}
\]
Theoretical guarantees of IPSW with completely estimated weights

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

Asymptotic properties

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

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

Variance depends on the size of the two data sets, \( n \) and \( m \)
What if also estimating $\pi$?

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

Asymptotic properties

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

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

where 

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

Variance is smaller if also estimating $\pi$ with the data

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

---

(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?
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.
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,

\[ Y := f(GCS, \text{Gender}) + A\tau(TTT, \text{Blood Pressure}) + \epsilon_{TTT}, \]
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,

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

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

⇒ Physicians and epidemiologists have an important role to play in selecting a limited number of covariates when generalizing trial’s findings!
Theoretical guarantees of IPSW with semi-oracle (= so) weights

\[ \hat{\tau}_{\pi, T, n} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{p_T(X_i)}{\hat{p}_R(X_i)} Y_i \left( \frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right), \]

Estimated with \( \mathcal{R} \)

Asymptotic properties

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

Estimating \( p_R(x) \) is more efficient than taking the oracle probability (counter-intuitive!)