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

Missing values, causal inference

Erwan Scornet

Associate professor

École Polytechnique

Random forests, missing values

Gaël Varoquaux

Research director

Inria

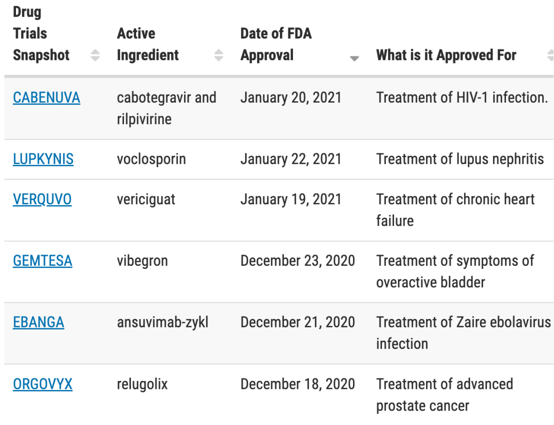
Co-founder of scikit-learn,

Machine-Learning

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`

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

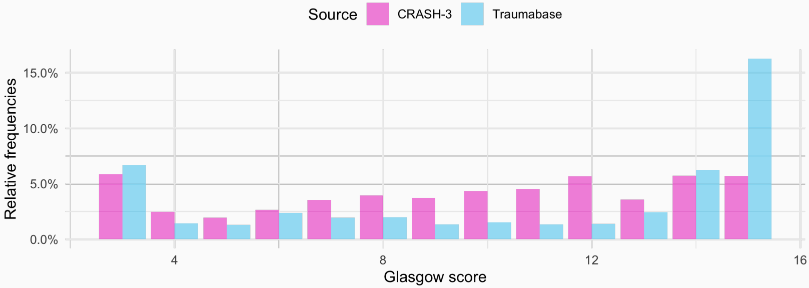
- Short timeframe,
- unrealistic real-world compliance,
- limited sample size,
- unrepresentative sample.

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] $Y_i^{(a)}$ 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 $p_{\mathrm{R}}(x)$ 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 $p_{\mathrm{T}}(x)$).

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

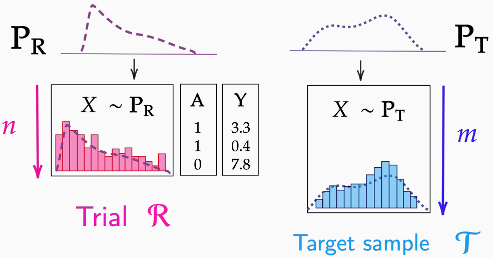
Introduction to the notations

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 $p_T(x)$).



A	Y
1	3.3
1	0.4
0	7.8

[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:

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

- 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,

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

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

Compute ATE averaging over the trial sample:

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

- 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?

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

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

Compute ATE averaging over the trial sample:

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

- 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,

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

Re-weighting the trial's data?

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

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

Generalization's *causal* assumptions.

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

$$w(X) := \frac{p_{\mathrm{T}}(X)}{p_{\mathrm{R}}(X)}.$$

Generalization's *causal* assumptions.

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

$$w(X) := \frac{p_{\mathrm{T}}(X)}{p_{\mathrm{R}}(X)}.$$

Which assumptions?

Transportability

$$\forall x \in \mathcal{X},\ \mathbb{P}_{\mathrm{R}}(Y^{(1)} - Y^{(0)} \mid X = x) = \mathbb{P}_{\mathrm{T}}(Y^{(1)} - Y^{(0)} \mid X = x).$$

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

Support inclusion

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

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 guarantees of IPSW with oracle weights

$$\hat{\tau}^{*}_{\pi,\text{T, R},n} = \frac{1}{n}\sum_{i\in\mathcal{R}} \overset{\text{True (or oracle) probabilities}}{\boxed{\frac{p_{\text{T}}(X_i)}{p_{\text{R}}(X_i)}}} \; Y_i \left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi}\right),$$

Theoretical guarantees of IPSW with oracle weights

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

Properties

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

where

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

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

How do we estimate weights in practice?

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

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

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

Theoretical guarantees of IPSW with completely estimated weights

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

Asymptotic properties

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

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

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

What if also estimating π?

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

Asymptotic properties

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

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

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: 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?

d

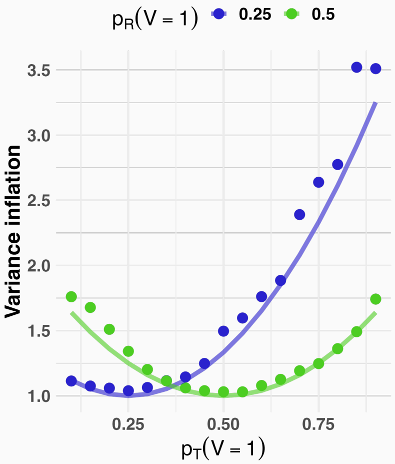
Impact of additional covariates: for the worse

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

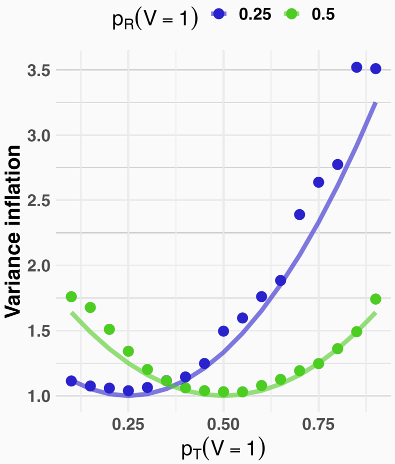
Impact of additional covariates: for the worse

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

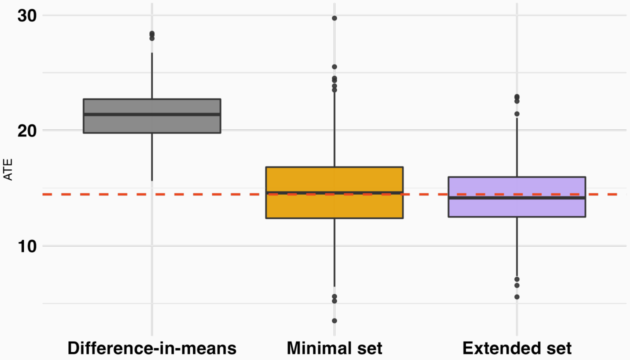
(i) Including non-necessary covariates can seriously damage precision!

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.

Semi-synthetic simulation

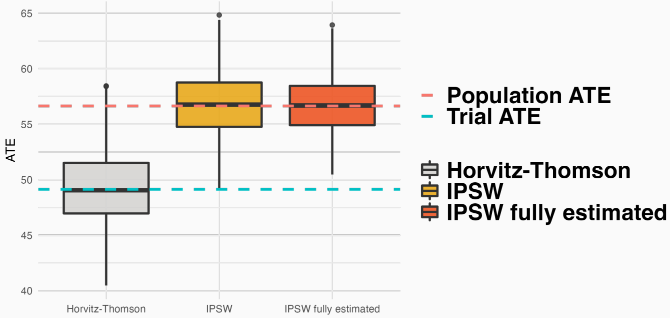
- All the results are illustrated on semi-synthetic simulations;
- Build from two large clinical data bases, reflecting a real-world situation
  - CRASH3 $\sim$ 9 000 individuals.
  - Traumabase $\sim$ 30 000 individuals.
- The outcome is the only synthetic part,

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

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!*

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

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

Estimated with $\mathcal{R}$

Asymptotic properties

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

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