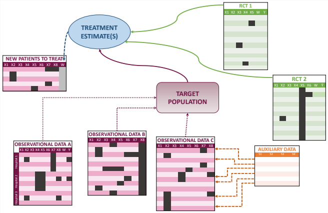
Risk Difference, Risk Ratio, Odds Ratio: Key Properties for Transportability and Federated Causal Inference

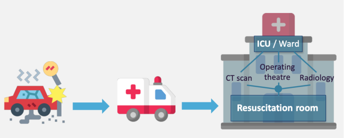
Julie Josse. Senior Researcher Inria 2020-; Prof. Ecole Polytechnique 2016-2020

Lead Inria-Inserm PreMeDICaL team: personalized medicine by data integration & causal learning



(Online) Decision support tool with quantified uncertainty

Ex: Traumatrix project[1]: Reducing under and over triage for improved resource allocation in trauma care



Major trauma: brain injuries or hemorrhagic shock from car accidents, falls, stab wounds, etc. ⇒ requires specialized care/resources in "trauma centers"

Many patients are misdirected: human/ economical costs

Clinical trial launched in 2025: real-time implementation of Machine Learning models in ambulance dispatch via a mobile data collection application

[1] www.traumabase.eu - https://www.traumatrix.fr/

Personalization of treatment recommendation

Ex: Estimating treatment effect from the Traumabase data

- 40000 trauma patients
- 300 heterogeneous features from pre-hospital and in-hospital settings
- 40 trauma centers, 4000 new patients per year

Center	Accident	Age	Sex	Weight	Lactacte	Blood Press.	TXA.	Y
Beaujon	fall	54	m	85	NA	180	treated	0
Pitie	gun	26	m	NA	NA	131	untreated	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NA	NA	107	untreated	0
HEGP	knife	16	m	98	2.5	118	treated	1
⋮								⋱

⇒ **Estimate causal effect** (with missing values[2]): Administration of the **treatment** *tranexamic acid (TXA)*, given within 3 hours of the accident, on the **outcome** (Y) *28 days in-hospital mortality* for trauma brain patients

[2] Mayer, I., Wager, S. & **J.J**. (2020). Doubly robust treatment effect estimation with incomplete confounders. *Annals Of Applied Statistics. (implemented in package `grf`).*

Causal inference: "what would happen if?"

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

- $\big(\underbrace{X}_{\text{covariates}} , \overbrace{W}^{\text{treatment}} , \underbrace{Y(1), Y(0)}_{\textit{potential outcomes}} \big) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$
- Individual **causal effect** of the binary treatment: $\Delta_i = Y_i(1) - Y_i(0)$

Problem: Δ_i never observed (only one outcome is observed per 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	**?**

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

ATE with **Risk Difference**: difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

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

Observational data

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in 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

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in 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 sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in treated and control groups
  ⇒ High **internal** validity

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

Observational data

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

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

Leverage both RCT and observational data

RCT

+ No confounding
− Trial sample different from the population eligible for treatment

(big) Observational data

− Confounding
+ Representative of the target population

We can use both to [3] ...

- ... validate observational methods, correct for confounding bias
- ... improve estimation of heterogeneous treatment effects
- ... **generalize the treatment effect to a target population** (data fusion, transportability, recovery from selection bias)[4],[5]

[3] Colnet, et al. **J.J.** (2022). Causal inf. for combining RCT & obs. studies. *Statistical Science*.
[4] Elias Bareinboim & Judea Pearl. (2016). Causal inference & the data-fusion problem. *PNAS*.
[5] Dahabreh, Haneuse, Robins, Robertson, Buchanan, Stuart, Hernan. (2021). Study Designs for Extending Causal Inferences From a RCT to a Target Population *American J. of Epidemiology*.

Leverage both RCT and observational data

RCT

+ No confounding
− Trial sample different from the population eligible for treatment

(big) Observational data

− Confounding
+ Representative of the target population

We can use both to [3] ...

▷ ... validate observational methods, correct for confounding bias

▷ ... improve estimation of heterogeneous treatment effects

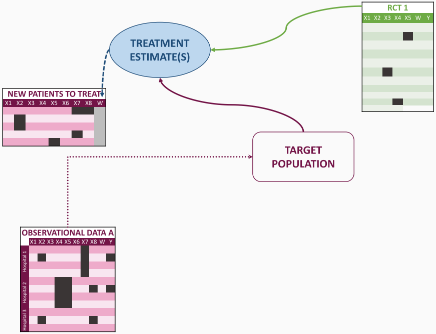
▷ ... **generalize the treatment effect to a target population** (data fusion, transportability, recovery from selection bias)[4],[5]

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

→ Reduce drug approval times and costs

[3] Colnet, et al. **J.J.** (2022). Causal inf. for combining RCT & obs. studies. *Statistical Science*.
[4] Elias Bareinboim & Judea Pearl. (2016). Causal inference & the data-fusion problem. *PNAS*.
[5] Dahabreh, Haneuse, Robins, Robertson, Buchanan, Stuart, Hernan. (2021). Study Designs for Extending Causal Inferences From a RCT to a Target Population *American J. of Epidemiology*.

Predicting treatment effects from 1 trial to another population



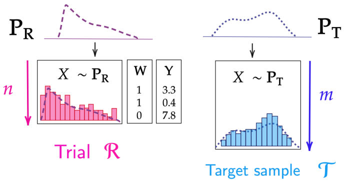
   

Bénédicte Colnet (Corps des Mines, French social security's direction), Imke Mayer (Owkin)
Erwan Scornet (X - Sorbonne Université), Gaël Varoquaux (Inria)

Generalization task from one RCT to a target population

Two data sources:

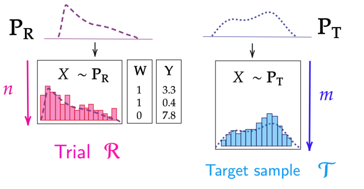
- ▷ A trial of size n with $p_{\mathrm{R}}(x)$ the probability of observing individual with $X = x$,
- ▷ A sample of the target population of interest – for e.g. a national cohort (resp. m and $p_{\mathrm{T}}(x)$).



Generalization task from one RCT to a target population

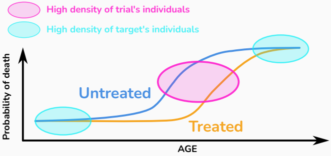
Two data sources:

- 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 $p_T(x)$).



Covariates distribution not the same in the RCT & target pop:

$$p_R(x) \neq p_T(x) \Rightarrow \underbrace{\tau_R := \mathbb{E}_R[Y(1) - Y(0)]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}_T[Y(1) - Y(0)] := \tau_T}_{\text{Target ATE}}$$



Assumptions for ATE identifiability in generalization

Overlap assumption[6]

$$\forall x \in \mathbb{X}, p_{\text{R}}(x) > 0 \text{ and } \operatorname{supp}(P_T(X)) \subset \operatorname{supp}(P_R(X))$$

The observational covariates support is included in the RCT's support. Every individual in the target population could have been selected into the trial

[6] If this is too strong, we could generalize on a different target population: the target population for which eligibility criteria of the trial are ensured

[7] Equivalent formulation with sampling mechanism S ($S = 1$ trial eligibility & willingness to participate) in non-nested design, $\{Y(1), Y(0)\} \perp\!\!\!\perp S \mid X$

Assumptions for ATE identifiability in generalization

Overlap assumption[6]

$$\forall x \in \mathbb{X}, p_{\mathrm{R}}(x) > 0 \text{ and } \operatorname{supp}(P_T(X)) \subset \operatorname{supp}(P_R(X))$$

The observational covariates support is included in the RCT's support. Every individual in the target population could have been selected into the trial

Transportability (Ignorability on trial participation)[7]

$$\forall w \in \{0,1\} \quad \mathbb{E}_{\mathrm{R}}[Y(w) \mid X] = \mathbb{E}_{\mathrm{T}}[Y(w) \mid X]$$

Corresponds to shifted prognostic variables

[6] If this is too strong, we could generalize on a different target population: the target population for which eligibility criteria of the trial are ensured

[7] Equivalent formulation with sampling mechanism S ($S = 1$ trial eligibility & willingness to participate) in non-nested design, $\{Y(1), Y(0)\} \perp\!\!\!\perp S \mid X$

Generalization of conditional outcome: identifiability

	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: $\tau_{\mathrm{T}} = \mathbb{E}_{\mathrm{T}}[Y_i(1) - Y_i(0)], \forall w \in \{0,1\}$

$$\begin{aligned}
\mathbb{E}_{\mathrm{T}}[Y(w)] &= \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{T}}[Y(w) \mid X]] \text{ Law of total expectation} \\
&= \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{R}}[Y(w) \mid X]] \text{ Ignorability} \\
&= \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{R}}[Y(w) \mid X = x, W = w]] \text{ Random treatment} \\
&= \mathbb{E}_{\mathrm{T}}\underbrace{[\mathbb{E}_{\mathrm{R}}[Y \mid X = x, W = w]]}_{\mu_w(x)} \text{ Consistency } Y = Y(1)W + (1-W)Y(0
\end{aligned}$$

Generalization of conditional outcome: identifiability

	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: $\tau_{\mathrm{T}} = \mathbb{E}_{\mathrm{T}}[Y_i(1) - Y_i(0)], \forall w \in \{0,1\}$

$$\begin{aligned}
\mathbb{E}_{\mathrm{T}}[Y(w)] &= \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{T}}[Y(w) \mid X]] \text{ Law of total expectation} \\
&= \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{R}}[Y(w) \mid X]] \text{ Ignorability} \\
&= \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{R}}[Y(w) \mid X = x, W = w]] \text{ Random treatment} \\
&= \mathbb{E}_{\mathrm{T}}\underbrace{[\mathbb{E}_{\mathrm{R}}[Y \mid X = x, W = w]]}_{\mu_w(x)} \text{ Consistency } Y = Y(1)W + (1-W)Y(0
\end{aligned}$$

Regression adjustment - plug-in gformula

$$\hat{\tau}_{g,n,m} = \frac{1}{m} \sum_{i \in \mathcal{T}} (\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i))$$

Plug-in gformula: difference between conditional mean

Plug-in gformula

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

$\mu_w(x) = \mathbb{E}_{\mathrm{R}}[Y \mid X = x, W = w]$

			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 ($\mathcal{R}$) for treated and for control to get $\widehat{\mu}_{1,n}$ & $\widehat{\mu}_{0,n}$

Plug-in gformula: difference between conditional mean

Plug-in gformula

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

$\mu_w(x) = \mathbb{E}_{\mathrm{R}}[Y \mid X = x, W = w]$

			Covariates			Treat	Outcomes	
	Set	S	X_1	X_2	X_3	W	$Y(0)$	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 ($\mathcal{R}$) 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 $\overline{\hat{\mu}_{1,n}(\cdot)}$ - $\overline{\hat{\mu}_{0,n}(\cdot)}$

Assumptions for ATE identifiability in generalization

Overlap assumption[8]

$$\forall x \in \mathbb{X}, p_{\text{R}}(x) > 0 \text{ and } \operatorname{supp}(P_T(X)) \subset \operatorname{supp}(P_R(X))$$

The observational covariate support is included in the RCT's support. Every individual in the target population could have been selected into the trial

Transportability of the conditional average treatment effect (CATE)[9]

$$\underbrace{\mathbb{E}_{\text{R}}[Y(1) - Y(0) \mid X]}_{\tau_R(X)} = \underbrace{\mathbb{E}_{\text{T}}[Y(1) - Y(0) \mid X]}_{\tau_{\text{T}}(X)}$$

Need to know which variables are shifted treatment effect modifiers

The treatment effect depends on covariates in the same way in the source (RCT) and target population

[8] If this is too strong, we could generalize on a different target population: the target population for which eligibility criteria of the trial are ensured

[9] Equivalent formulation (non-nested case) with sampling mechanism S: $(Y(1) - Y(0)) \perp\!\!\!\perp S \mid X$

Identifiability and estimation for generalization: weighting

Generalization of local effects (i.e. conditional effects/strata)

$$\begin{aligned}\tau_{\mathrm{T}} &= \mathbb{E}_{\mathrm{T}}[Y_i(1) - Y_i(0)] = \mathbb{E}_{\mathrm{T}}[\mathbb{E}_{\mathrm{T}}[Y_i(1) - Y_i(0)|X]] \\ &= \mathbb{E}_{\mathrm{T}}\left[\tau_{\mathrm{T}}(X)\right] = \mathbb{E}_{\mathrm{T}}\left[\tau_{\mathrm{R}}(X)\right] \text{ Transportability CATE} \\ &= \mathbb{E}_{\mathrm{R}}\left[\frac{p_{\mathrm{T}}(X)}{p_{\mathrm{R}}(X)}\tau_{\mathrm{R}}(X)\right] \text{ Overlap}\end{aligned}$$

IPSW: inverse propensity sampling weighting

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

π proba. of treatment assignment in trial

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

Re-weighting can be found in the 2000's (*standardization*)[10]

Idea of relying on an external representative sample to reweight is recent[11]

[10] Rothman & Greenland (1998). Modern Epidemiology.

[11] Cole & Stuart (2010). Generalizing from RCT to target ... *American Journal of Epidemiology*

Reweighting the RCT: reweight Horvitz-Thomson

$$\hat{\tau}_{\pi,n,m} = \frac{1}{n}\sum_{i\in\mathcal{R}} \quad \frac{p_T(X)}{p_R(X)} \quad Y_i\left(\frac{W_i}{\pi} - \frac{1-W_i}{1-\pi}\right)$$

- Estimate the **ratio of densities**[12]

▷ with parametric densities (i.e. Gaussian)

▷ with a parametric model for the ratio

▷ with logistic regression

$$r(X) := \frac{p_T(X)}{p_R(X)} = \frac{\mathbb{P}(X=x|S=0)}{\mathbb{P}(X=x|S=1)}$$

$$= \frac{\mathbb{P}(S=1)\mathbb{P}(S=0|X=x)}{\mathbb{P}(S=0)\mathbb{P}(S=1|X=x)}$$

$$\forall x \in \mathcal{X}, \quad \hat{r}(x) = \frac{n/(n+m)}{m/(n+m)}\frac{1-\hat{\sigma}(x,\boldsymbol{\beta}_{n+m})}{\hat{\sigma}(x,\boldsymbol{\beta}_{n+m})}$$

where $x \in \mathcal{X}$, $\sigma(x,\boldsymbol{\beta}) = \left(1+\exp(-x^\top\beta)\right)^{-1}$.

- Case with categorical features: finite sample & asymptotic analysis[13]

[12] Kanamori, et al. (2010). Theoretical analysis of density ratio estimation. *IEICE transactions.*

[13] Colnet, **J.J** (2022). Reweighting the RCT: finite sample analysis & variable selection. *JRSSA.*

Generalization from Crash 3 trial[14] to the Traumabase

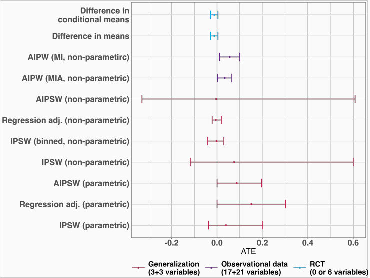
CRASH3

- Multi-centric RCT - 29 countries
- 9000 individuals - develp. countries
- Positive effect for moderately injured patients

Traumabase

- Observational sample
- 8200 patients with brain trauma
- Deleterious/No evidence for an effect of Tranexomic Acid

Comparison of trials, observational data, and generalization estimates



x-axis: Estimation of the Average Treatment Effect, Confidence intervals with bootstrap
y-axis: Estimation methods (estimation of nuisances: parametric: logistic regression - non parametric: forests)

[14](2019). Effects of tranexamic acid on death in patients with acute trauma. brain injury. *Lancet.*

Many medical and statistical challenges

- 1) Shifted effect modifiers not available in Traumabase[15]. Missing covariates in one/both sets: **sensitivity analysis**

	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		

[15] Colnet, **J.J**, et al. 2022. Generalizing a causal effect: sensitivity analysis. *J. of Causal Inference*.
[16] Mayer, **J.J.** 2021. Generalizing effects with incomplete covariates *Biometrical Journal.*
[17] Colnet, **J.J** et al. 2023. Reweighting the RCT for generalization: finite sample analysis and variable selection. *JRSSC.*
[18] Colnet, **J.J** et al. 2024. Risk-Ratio, Odds-ratio, wich causal measure is easier to generalize?

Many medical and statistical challenges

- 1) Shifted effect modifiers not available in Traumabase[15]. Missing covariates in one/both sets: **sensitivity analysis**

	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		

- 2) **Missing values:** Missing values (NA) in both RCT and Obs data[16]

[15] Colnet, **J.J**, et al. 2022. Generalizing a causal effect: sensitivity analysis. *J. of Causal Inference*.
[16] Mayer, **J.J.** 2021. Generalizing effects with incomplete covariates *Biometrical Journal*.
[17] Colnet, **J.J** et al. 2023. Reweighting the RCT for generalization: finite sample analysis and variable selection. *JRSSC*.
[18] Colnet, **J.J** et al. 2024. Risk-Ratio, Odds-ratio, wich causal measure is easier to generalize?

Many medical and statistical challenges

• 1) Shifted effect modifiers not available in Traumabase[15]. Missing covariates in one/both sets: **sensitivity analysis**

	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		

• 2) **Missing values:** Missing values (NA) in both RCT and Obs data[16]

• 3) Which covariates should be include? Would adding prognostic variables reduce the variance as in the classical case?[17]

[15] Colnet, **J.J**, et al. 2022. Generalizing a causal effect: sensitivity analysis. *J. of Causal Inference*.
[16] Mayer, **J.J.** 2021. Generalizing effects with incomplete covariates *Biometrical Journal.*
[17] Colnet, **J.J** et al. 2023. Reweighting the RCT for generalization: finite sample analysis and variable selection. *JRSSC.*
[18] Colnet, **J.J** et al. 2024. Risk-Ratio, Odds-ratio, wich causal measure is easier to generalize?

Many medical and statistical challenges

- 1) Shifted effect modifiers not available in Traumabase[15]. Missing covariates in one/both sets: **sensitivity analysis**

	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		

- 2) **Missing values:** Missing values (NA) in both RCT and Obs data[16]

- 3) Which covariates should be include? Would adding prognostic variables reduce the variance as in the classical case?[17]

- 4) Clinicians are more interested in the **risk ratio** than the risk difference[18]

[15] Colnet, **J.J**, et al. 2022. Generalizing a causal effect: sensitivity analysis. *J. of Causal Inference*.
[16] Mayer, **J.J.** 2021. Generalizing effects with incomplete covariates *Biometrical Journal*.
[17] Colnet, **J.J** et al. 2023. Reweighting the RCT for generalization: finite sample analysis and variable selection. *JRSSC*.
[18] Colnet, **J.J** et al. 2024. Risk-Ratio, Odds-ratio, wich causal measure is easier to generalize?

Comparing two average situations

Binary outcome: $\mathbb{P}[Y(w)=1]=\mathbb{E}[Y(w)]$ and $\mathbb{P}[Y(w)=0]=1-\mathbb{E}[Y(w)]$.

Absolute measures

$$\tau^{\mathrm{RD}} := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)], \qquad \tau^{\mathrm{NNT}} := (\tau^{\mathrm{RD}})^{-1}.$$

- Number Needed to Treat (NNT): how many individuals should be treated to observe one individual answering positively to treatment.

Relative measures

$$\tau^{\mathrm{RR}} := \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]}, \quad \tau^{\mathrm{SR}} := \frac{\mathbb{P}[Y(1)=0]}{\mathbb{P}[Y(0)=0]} = \frac{1-\mathbb{E}[Y(1)]}{1-\mathbb{E}[Y(0)]},$$

$$\tau^{\mathrm{OR}} := \frac{\mathbb{P}[Y(1)=1]}{\mathbb{P}[Y(1)=0]} \left(\frac{\mathbb{P}[Y(0)=1]}{\mathbb{P}[Y(0)=0]}\right)^{-1}$$

- A null effect now corresponds to a Risk Ratio of 1
- Survival Ratio (SR) corresponds to the RR with swapped labels Y
- RR is not symmetric to the choice of outcome 0 and 1 –e.g. counting the living or the dead while Odds Ratio (OR) is

Different treatment measures give different impressions

An example: Randomized Control Trial (RCT) from Cook and Sackett (1995)

- $Y = 1$ stroke in 5 years and $Y = 0$ no stroke
- W antihyperintensive therapy
- Feature X (blood pressure), $X = 1$ low baseline risk (15/1000 versus 2/10)

$$\mathbb{P}[Y(0) = 1 \mid X = 0] \geq \mathbb{P}[Y(0) = 1 \mid X = 1]$$

	τ_{RD}	τ_{RR}	τ_{SR}	τ_{NNT}	τ_{OR}
All (P_{R})	-0.0452	0.6	1.05	22	0.57
X = 1	−0.006	0.6	1.01	167	0.6
X = 0	−0.08	0.6	1.1	13	0.545

- RD: treatment reduces by 0.045 the probability to suffer from a stroke
- RR: the treated has 0.6 $\times$ the risk of having a stroke comp. with the control
- SR: increased chance of not having a stroke when treated (factor 1.05).
- NNT: one has to treat 22 people to prevent one additional stroke
- OR $\approx$ RR in a stratum where prevalence of the outcome is low

Different treatment measures give different impressions

An example: Randomized Control Trial (RCT) from Cook and Sackett (1995)

- $Y = 1$ stroke in 5 years and $Y = 0$ no stroke
- W antihyperintensive therapy
- Feature X (blood pressure), $X = 1$ low baseline risk (15/1000 versus 2/10)

$$\mathbb{P}[Y(0) = 1 \mid X = 0] \geq \mathbb{P}[Y(0) = 1 \mid X = 1]$$

	τ_{RD}	τ_{RR}	τ_{SR}	τ_{NNT}	τ_{OR}
All (P_R)	-0.0452	0.6	1.05	22	0.57
X = 1	−0.006	0.6	1.01	167	0.6
X = 0	−0.08	0.6	1.1	13	0.545

- RD is heterogeneous with X while RR is homogeneous with X
- Heterogeneity's property defined w.r.t. *(i)* covariates & *(ii)* a measure
- Impact of the baseline risk: with 3% baseline mortality reduced to 1% by treatment, RD shows a 0.02 drop, while RR shows controls have three times the risk: RD suggests a small effect; RR highlights a larger one

The age-old question of how to report effects


Source: Wikipedia

" We wish to decide whether we shall count the failures or the successes and whether we shall make relative or absolute comparisons"

— Mindel C. Sheps, New England Journal of Medicine, in 1958

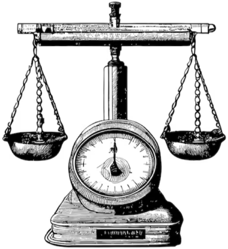
The choice of the measure is still actively discussed

e.g. Spiegelman and VanderWeele, 2017; Baker and Jackson, 2018; Feng et al., 2019; Doi et al., 2022; Xiao et al., 2021, 2022; Huitfeldt et al., 2021; Lapointe-Shaw et al., 2022; Liu et al., 2022 …

— CONSORT guidelines recommend to report all of them

Risk ratio, odds ratio, risk difference

Which causal measure is easier to generalize?



A desirable property: collapsibility

Collapsibility[19]: Population's effect is equal to a weighted sum of local effects (conditional effects)

Direct collapsibility - weights are equal to population's proportions

$$\tau = \mathbb{E}\left[\tau(X)\right]$$

- Risk Difference is directly collapsible

	τ_{RD}	τ_{RR}	τ_{SR}	τ_{NNT}	τ_{OR}
All (P_{R})	-0.0452	0.6	1.05	22	0.57
X = 1	-0.006	0.6	1.01	167	0.6
X = 0	-0.08	0.6	1.1	13	0.545

$$\tau_{\text{R}}^{\text{RD}} = p_{\text{R}}(X=1) \times \tau_{\text{R}}^{\text{RD}}(X=1) + p_{\text{R}}(X=0) \times \tau_{\text{R}}^{\text{RD}}(X=0)$$

$$-0.0452 = -0.47 \times 0.006 - 0.53 \times 0.08.$$

Useful for generalization! (replacing p_{R} by p_{T})

[19] Greenland (1987), Hernan et al. (2011), Huitfield et al. (2019), Didelez & Stensrud (2022), etc.

A desirable property: collapsibility

Collapsibility: Population's effect is equal to a weighted sum of local effects (conditional effects)

Collapsibility: weights depend on the baseline distribution $Y(0)$

$$\mathbb{E}\left[w(X, P(X, Y(0)))\, \tau(X)\right] = \tau \quad \text{with } w \geq 0,\ \mathbb{E}\left[w(X, P(X, Y(0)))\right] = 1$$

- Risk Ratio is collapsible:

$$\mathbb{E}\left[\tau_{RR}(X) \frac{\mathbb{E}\left[Y(0) \mid X\right]}{\mathbb{E}\left[Y(0)\right]}\right] = \tau_{RR}$$

- Estimation challenges: No methods or theoretical properties for RR in RCTs & observational data. In Boughdiri, et al (2024)[20] we propose: Weighting & outcome modeling estimators (asymptotic & finite-sample analyses) + Two doubly robust estimators via semi-parametric theory.

[20] Boughdiri, **J.J.**, Scornet. (2024). Estimating Risk Ratios in Causal Inference. *Submited*.

Summary of causal measure properties

Direct collapsibility

$$\mathbb{E}\left[\tau(X)\right] = \tau$$

Collapsibility: weights depend on the baseline distribution $Y(0)$

$$\mathbb{E}\left[w(X, P(X, Y(0)))\,\tau(X)\right] = \tau \quad \text{with } w \geq 0,\ \mathbb{E}\left[w(X, P(X, Y(0)))\right] = 1$$

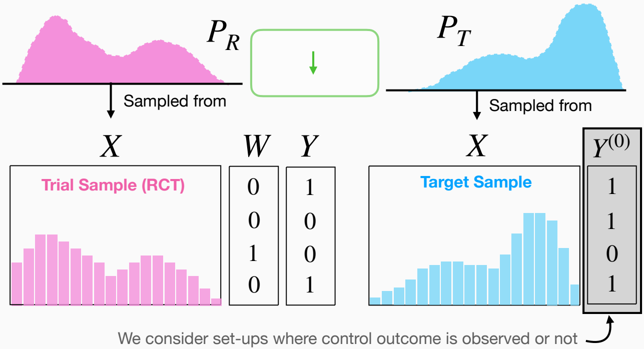
Logic respecting (Simpson paradox)

$$\tau \in \left[\min_x(\tau(x)), \max_x(\tau(x))\right].$$

Ex. OR: Overall population, $\tau_{\text{OR}} \approx 0.26$ $\tau_{\text{OR}|F=1} \approx 0.167$ and $\tau_{\text{OR}|F=0} \approx 0.166$

Measure	Dir. collapsible	Collapsible	Logic-respecting
Risk Difference	Yes	Yes	Yes
Number Needed to Treat	No	No	Yes
Risk Ratio	No	Yes	Yes
Survival Ratio	No	Yes	Yes
Odds Ratio	No	No	No

Back to generalizability from one RCT to a Target pop.



Back to generalizability from one RCT to a Target pop.

Generalizing	Conditional Outcome	Local effects/CATE
Assumption	$\mathbb{E}_{\mathrm{R}}[Y(w) \mid X] = \mathbb{E}_{\mathrm{T}}[Y(w) \mid X]$	$\tau_R(X) = \tau_{\mathrm{T}}(X)$
Variables	All shifted prognostic covariates	All shifted effect modifiers
Identification	$\mathbb{E}_{\mathrm{T}}\left[Y(w)\right] = \mathbb{E}_{\mathrm{T}}\left[\mathbb{E}_{\mathrm{R}}\left[Y(w) \mid X\right]\right]$	$\mathbb{E}_{\mathrm{R}}\left[\frac{p_{\mathrm{T}}(X)}{p_{\mathrm{R}}(X)}\, w_{\mathrm{T}}(Y(0), X)\, \tau_{\mathrm{R}}(X)\right]$
Estimation	Ex: Regression (G-formula)	Ex: Weighting

- Generalize local effects only for collapsible measures, need info. on $Y(0)$
- Generalizing conditional outcome require stronger assumptions
- Depending on the underlying DGP assumption & direction of the effects, some measure disantangle baseline risk from effect modifiers[21]

[21] Colnet, **J.J**, et al. (2024). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?

Generalization of a first moment population-level estimands

Let $P(Y(0), Y(1))$ the joint distribution of the potential outcomes.

- τ^P a 1st moment population-level[22] measure if $\exists\ \Phi : D_\Phi \to \mathbb{R}$, $D_\Phi \subset \mathbb{R}^2$

$$\Phi\left(\mathbb{E}_P[Y(1)], \mathbb{E}_P[Y(0)]\right) = \tau^P_\Phi$$

Measure	Effect Measure	Domain D_Φ
Risk Difference (RD)	$\Phi(x, y) = x - y$	$\mathbb{R}^2$
Risk Ratio (RR)	$\Phi(x, y) = \frac{x}{y}$	$\mathbb{R} \times \mathbb{R}^*$
Odds Ratio (OR)	$\Phi(x, y) = \frac{x}{1-x} \cdot \frac{1-y}{y}$	$\mathbb{R}/\{1\} \times \mathbb{R}^*$
NNT	$\Phi(x, y) = \frac{1}{x-y}$	$\{(x, y) \in \mathbb{R}^2 \vert x + y \neq 0\}$

- An individual-level measure depends on the joint distribution. Considered non identifiable but workarounds exist[23]. Ex: $\mathbb{E}\left[\frac{Y_i(1)}{Y_i(0)}\right] \neq \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Y_i(0)]}$

[22] Fay & Li. (2024). Causal interpretation of the hazard ratio in RCTs. *Clinical Trials*.
[23] Even, **J.J.** (2025). Rethinking the win ratio: causal framework for hierarchical outcome Analysis.

Generalization of first moment population-level estimands

Identifiability formulae

$$\mathbb{E}_{T}\left[Y(w)\right] = \mathbb{E}_{R}\left[\frac{p_T(X)}{p_R(X)} Y(w)\right]$$

Estimator: Oracle Re-weighted Horvitz-Thomson

$$\hat{\tau}_{\Phi}^{\pi,n,m,\sigma} = \Phi\left(\frac{1}{n}\sum_{i=1}^{n} r(X_i)\frac{Y_i W_i}{\pi}, \frac{1}{n}\sum_{i=1}^{n} r(X_i)\frac{Y_i(1-W_i)}{1-\pi}\right),$$

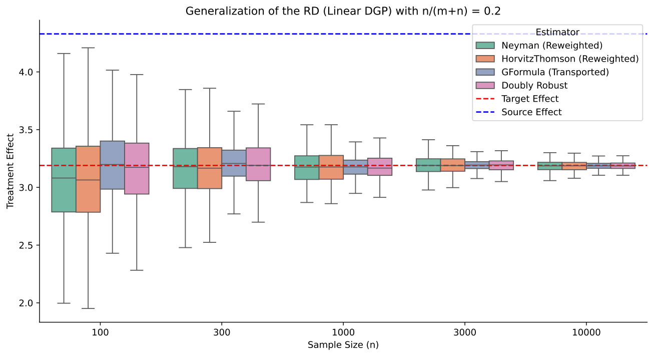
$$\sqrt{(n+m)}\left(\hat{\tau}_{\Phi}^{\pi,n,m,\sigma} - \tau_{\Phi}^{T}\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\Phi}^{\pi,\alpha,\sigma}\right)$$

Measure	Variance
Risk Difference (RD)	$\frac{1}{\alpha}\left(\frac{\mathbb{E}_T\left[r(X)(Y^{(1)})^2\right]}{\pi} + \frac{\mathbb{E}_T\left[r(X)(Y^{(0)})^2\right]}{1-\pi} - (\tau_{RD}^{T})^2\right)$
Risk Ratio (RR)	$\frac{(\tau_{RR}^{T})^2}{\alpha}\left(\frac{\mathbb{E}_T\left[r(X)(Y^{(1)})^2\right]}{\pi\mathbb{E}_T\left[Y^{(1)}\right]^2} + \frac{\mathbb{E}_T\left[r(X)(Y^{(0)})^2\right]}{(1-\pi)\mathbb{E}_T\left[Y^{(0)}\right]^2}\right)$
Odds Ratio (OR)	$\frac{(\tau_{OR}^{T})^2}{\alpha}\left(\frac{\mathbb{E}_T\left[r(X)(Y^{(1)})^2\right]}{\pi(\mathbb{E}_T[Y^{(1)}])^2} + \frac{\mathbb{E}_T\left[r(X)(Y^{(0)})^2\right]}{(1-\pi)(\mathbb{E}_T[Y^{(0)}])^2} - 1\right)$

Generalization of RD, RR and OR

Under assumption $\mathbb{E}_{\mathrm{R}}[Y(w) \mid X] = \mathbb{E}_{\mathrm{T}}[Y(w) \mid X]$

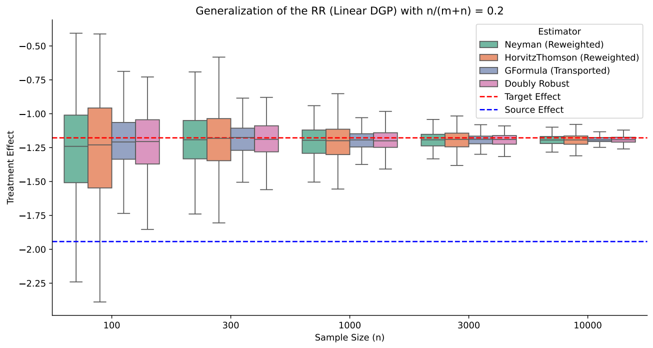
- $Y_{\mathrm{R}}(w) = c(w) + X_{\mathrm{R}}\beta(w) + \epsilon_{\mathrm{R}(w)}$
- $X_{\mathrm{R}} \sim \mathcal{N}(\mu_{\mathrm{R}}, \Sigma)$
- $X_{\mathrm{T}} \sim \mathcal{N}(\mu_{\mathrm{T}}, \Sigma)$



Generalization of RD, RR and OR

Under assumption $\mathbb{E}_{\text{R}}[Y(w) \mid X] = \mathbb{E}_{\text{T}}[Y(w) \mid X]$

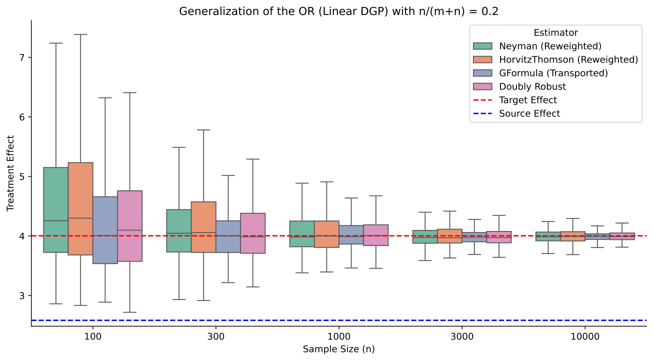
- $Y_{\text{R}}(w) = c(w) + X_{\text{R}}\beta(w) + \epsilon_{\text{R}(w)}$
- $X_{\text{R}} \sim \mathcal{N}(\mu_{\text{R}}, \Sigma)$
- $X_{\text{T}} \sim \mathcal{N}(\mu_{\text{T}}, \Sigma)$



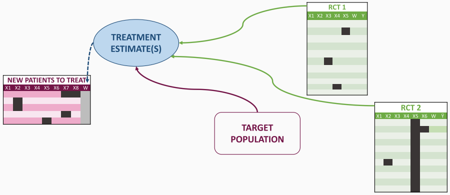
Generalization of RD, RR and OR

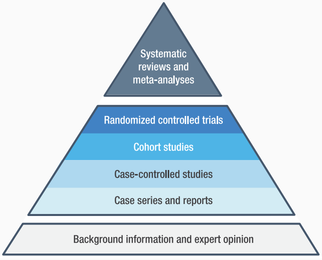
Under assumption $\mathbb{E}_{\mathrm{R}}[Y(w) \mid X] = \mathbb{E}_{\mathrm{T}}[Y(w) \mid X]$

- $Y_{\mathrm{R}}(w) = c(w) + X_{\mathrm{R}}\beta(w) + \epsilon_{\mathrm{R}(w)}$
- $X_{\mathrm{R}} \sim \mathcal{N}(\mu_{\mathrm{R}}, \Sigma)$
- $X_{\mathrm{T}} \sim \mathcal{N}(\mu_{\mathrm{T}}, \Sigma)$



From one to multiple Randomized Control Trials (RCTs)



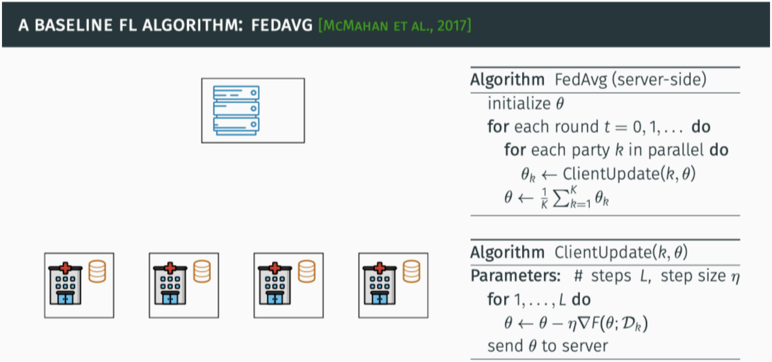


Meta-analysis (aggregating estimated effects from multiple studies) is at the top of the pyramid of evidence based medicine.

Meta-analysis still faces significant challenges:

- Be careful with aggregation of causal measures (Odds Ratio?)
- **Heterogeneity** across studies: sample size, population, center effects
- **Difficulty to share individual-level data**: **data silos** & **regulations**

Going beyond meta-analysis with federated causal inference[25]



Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

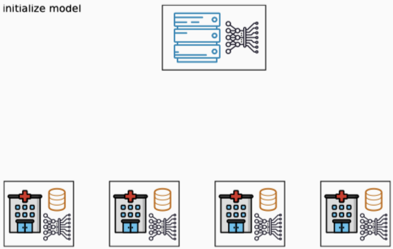
Going beyond meta-analysis on individual data[24]

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[25]

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

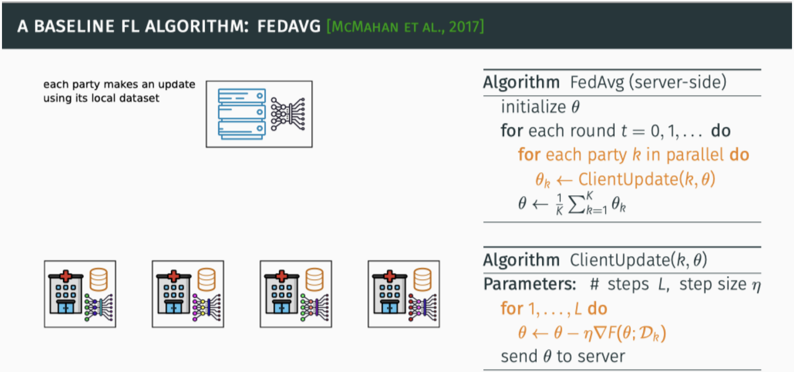
Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

Going beyond meta-analysis on individual data[24]

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[25]



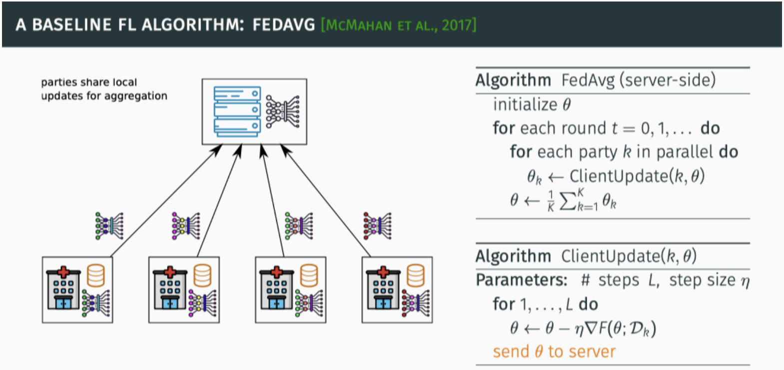
Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

Going beyond meta-analysis on individual data[24]

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[25]



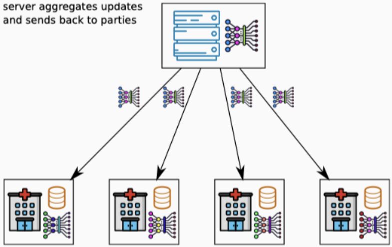
Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

Going beyond meta-analysis on individual data[24]

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*
[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[25]

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \dots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \dots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

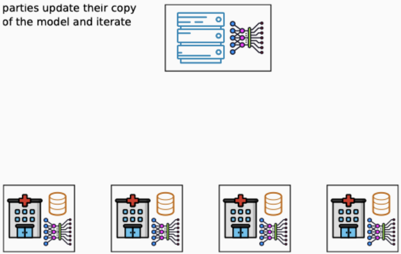
Going beyond meta-analysis on individual data[24]

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[25]

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \dots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \dots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

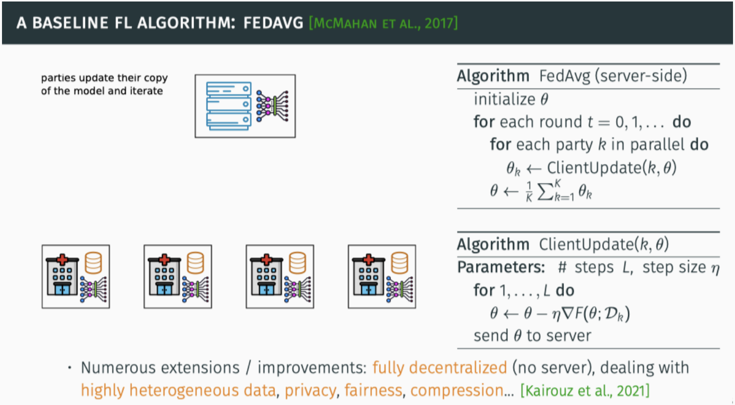
Going beyond meta-analysis on individual data[24]

---

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Going beyond meta-analysis with federated causal inference[25]



Bridging causal inference and federated learning to improve treatment effect estimation from **decentralized data sources**

Going beyond meta-analysis on individual data[24]

[24] Morris, T. et al. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Stat. Med.*

[25] Khellaf R, Bellet, A. & **J.J**. (2025). Multi-study ATE estimation beyond meta-analysis. *AISTAT*.

Federated Averaging (FedAvg) for Linear Regression

Linear Regression

$Y = X\beta + \varepsilon$. Estimate β by minimizing the MSE:

$$\text{argmin}_{\beta} \ell(\beta; X_i, Y_i) \text{ with } \ell(\beta; X_i, Y_i) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

Gradient Descent (GD)

1. Initialize β_0 with zeros
2. Update $\beta_{t+1} := \beta_t - \eta\nabla\ell(\beta_t)$, $\nabla\ell(\beta_t) = -\frac{2}{n}\sum_{i=1}^{n} X_i'(Y_i - X_i\beta)$
3. Repeat for L steps until convergence

Choices: learning rate η & L to get $\hat{\beta}_{\mathrm{GD}} \approx \hat{\beta}_{\mathrm{OLS}}$, equality as $L \to \infty$.

$\eta < \frac{2}{L}$, L the smoothness const., the highest eigenvalue $\lambda_{\max}$ of $X^{\top}X$

Federated Averaging (FedAvg) for Linear Regression

Linear Regression

$Y = X\beta + \varepsilon$. Estimate $\hat{\beta}^{\text{FedAvg}}$ by minimizing:

$$\text{argmin}_\beta \sum_{k=1}^{K} \frac{n_k}{n} \ell_k(\beta) \text{ with } \ell_k(\beta) = \frac{1}{n_k} \sum_{i=1}^{n_k} (Y_i^k - X_i^k \beta)^2$$

Federated Learning **extends GD to a distributed setting**

1. Initialize on central server β_0 with zeros (globally shared)
2. For each **communication round** $t = 1, \ldots, T$:
   - Each site $k = 1, \ldots, K$ performs $L = 1$ gradient step on its data:

   $\beta_{t+1}^k := \beta_t^k - \eta \nabla \ell_k(\beta_t^k)$ with $\nabla \ell_k(\beta_t^k) = -\frac{2}{n_k} \sum_{i=1}^{n_k} X_i'^k (Y_i^k - X_i^k \beta)$

   - Parameters sent to the server for aggregation: $\beta_{t+1} := \frac{1}{K} \sum_{k=1}^{K} \beta_{t+1}^k$

Choices: learning rate η, communication T & L.

$T = 1$ & $L \to \infty$: One-shot federated learning, meta analysis on β

Our setting: decentralized heterogeneous RCTs

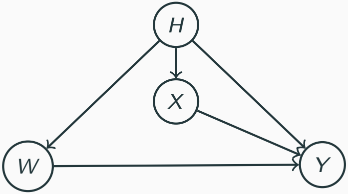
We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

Our setting: decentralized heterogeneous RCTs

We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

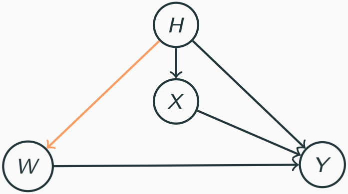


Our setting: decentralized heterogeneous RCTs

We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

Heterogeneity in **treatment allocation**

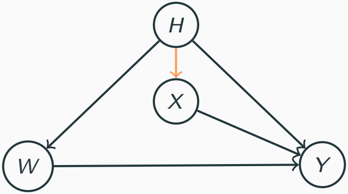


Our setting: decentralized heterogeneous RCTs

We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

Heterogeneity in **covariates distribution**

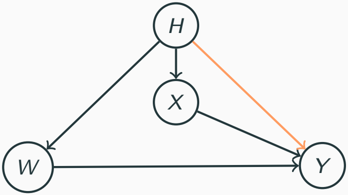


Our setting: decentralized heterogeneous RCTs

We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

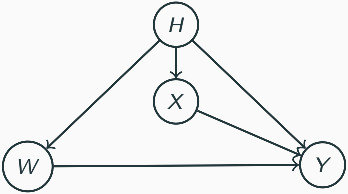
Presence of
center effects



Our setting: decentralized heterogeneous RCTs

We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2



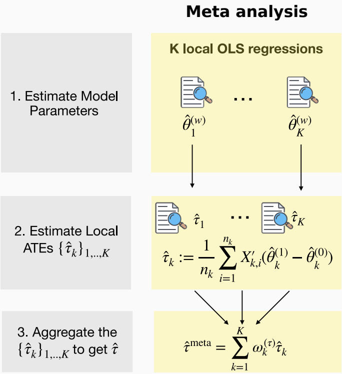
How to estimate τ without pooling together individual-level data?

Three types of federated estimators - collapsible measures (RD)

Ex: linear outcome model for all studies $\forall k$: $Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}$

Baseline: estimator $\hat{\tau}_{\text{pool}} = \frac{1}{n}\sum_{i=1}^{n} X_i'(\hat{\theta}_{\text{pool}}^{(1)} - \hat{\theta}_{\text{pool}}^{(0)})$ on pooled data

$\hat{\theta}_{\text{pool}}^{(w)} = (\hat{c}_{\text{pool}}^{(w)}, \hat{\beta}_{\text{pool}}^{(w)}) = \left(X'^{(w)\top} X'^{(w)}\right)^{-1} X'^{(w)\top} Y^{(w)}$ with $X'^{(w)} = [1, X^{(w)}]$



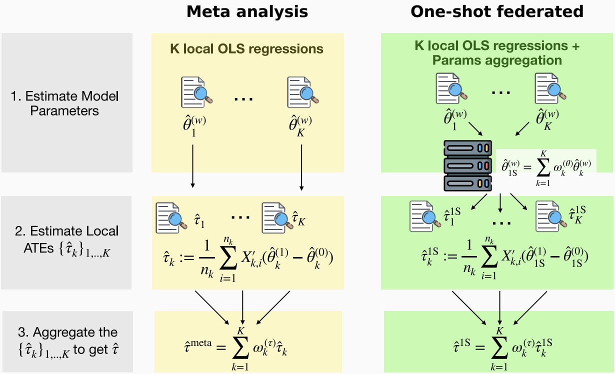
Aggregation w_k: sample size weights (SW) or inverse variance weights (IVW)

Three types of federated estimators - collapsible measures (RD)

Ex: linear outcome model for all studies $\forall k$: $Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}$

Baseline: estimator $\hat{\tau}_{\text{pool}} = \frac{1}{n}\sum_{i=1}^{n} X_i'(\hat{\theta}_{\text{pool}}^{(1)} - \hat{\theta}_{\text{pool}}^{(0)})$ on pooled data

$\hat{\theta}_{\text{pool}}^{(w)} = (\hat{c}_{\text{pool}}^{(w)}, \hat{\beta}_{\text{pool}}^{(w)}) = \left(X'^{(w)\top} X'^{(w)}\right)^{-1} X'^{(w)\top} Y^{(w)}$ with $X'^{(w)} = [1, X^{(w)}]$



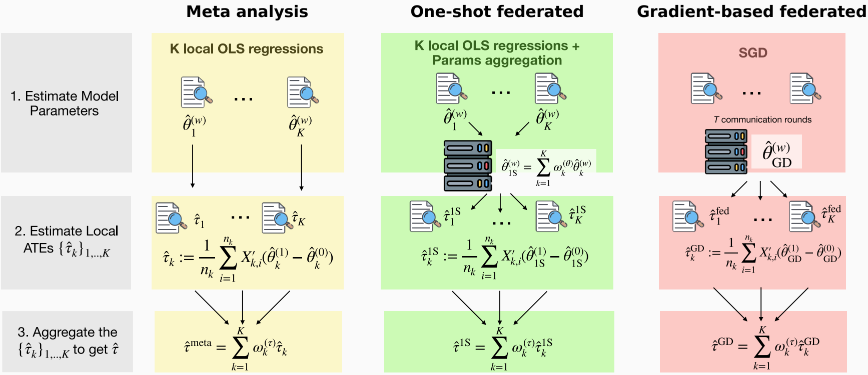
Aggregation w_k: sample size weights (SW) or inverse variance weights (IVW)

Three types of federated estimators - collapsible measures (RD)

Ex: linear outcome model for all studies $\forall k$: $Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}$

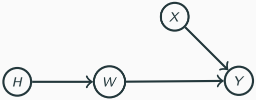
Baseline: estimator $\hat{\tau}_{\mathrm{pool}} = \frac{1}{n}\sum_{i=1}^{n} X_i'(\hat{\theta}_{\mathrm{pool}}^{(1)} - \hat{\theta}_{\mathrm{pool}}^{(0)})$ on pooled data

$\hat{\theta}_{\mathrm{pool}}^{(w)} = (\hat{c}_{\mathrm{pool}}^{(w)}, \hat{\beta}_{\mathrm{pool}}^{(w)}) = \left({X'^{(w)}}^{\top} X'^{(w)}\right)^{-1} {X'^{(w)}}^{\top} Y^{(w)}$ with $X'^{(w)} = [1, X^{(w)}]$



Aggregation w_k: sample size weights (SW) or inverse variance weights (IVW)

Statistical perf. & communication costs



Heterogeneity: Source membership H only affects treatment allocation: $W_{k,i} \sim \mathcal{B}(p_k)$

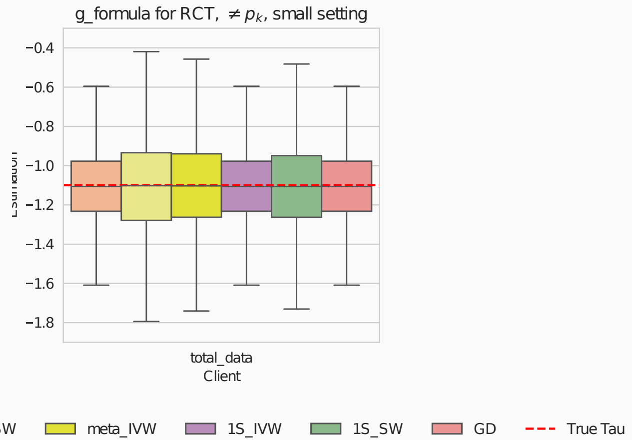
Unbiased estimators but different asymptotic variance & communication costs:

Estimator	$\mathbb{V}^\infty$	Com. rounds	Com. cost
$\hat{\tau}_{\text{Meta-SW}}$	$\frac{\sigma^2}{n}\sum_{k=1}^{K}\frac{\rho_k}{p_k(1-p_k)}+\frac{1}{n}\lVert\beta^{(1)}-\beta^{(0)}\rVert_\Sigma^2$	1	$O(1)$
$\hat{\tau}_{\text{Meta-IVW}}$	$\left(\sum_{k=1}^{K}(\sigma^2\frac{n\rho_k}{p_k(1-p_k)}+\frac{1}{n_k}\lVert\beta^{(1)}-\beta^{(0)}\rVert_\Sigma^2)^{-1}\right)^{-1}$	1	$O(1)$
$\hat{\tau}_{\text{1S-SW}}$	V_{pool}	2	$O(d)$
$\hat{\tau}_{\text{1S-IVW}}$	V_{pool}	2	$O(d^2)$
$\hat{\tau}_{\text{GD}}$	V_{pool}	$T+1$	$O(Td)$
$\hat{\tau}_{\text{pool}}$	$V_{\text{pool}}=\frac{\sigma^2}{n}\frac{1}{p(1-p)}+\frac{1}{n}\lVert\beta^{(1)}-\beta^{(0)}\rVert_\Sigma^2$	—	—

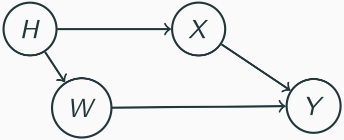
with $\rho_k=\mathbb{P}(H=k)$ and $p=\sum_{k=1}^{K}\frac{n_k}{n}p_k$

Numerical illustration

- $K = 5$ studies, $d = 10$ variables, $n_k = 5d$ observation/study
- Treatment allocation $p_1 = p_2 = p_3 = 0.9$, $p_4 = p_5 = 0.1$



Heterogeneity in covariates distributions



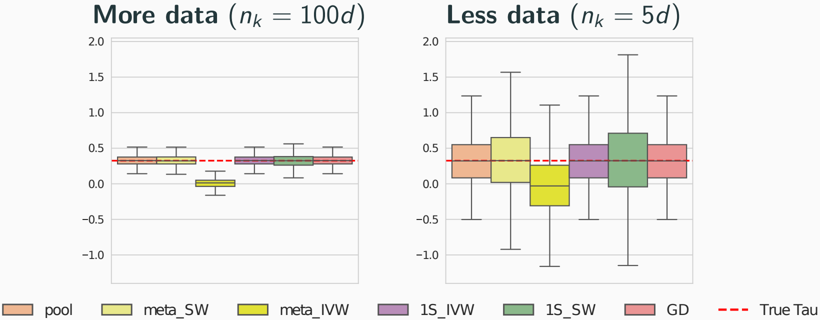
- **Distributional shift** across sources: $H \not\perp\!\!\!\perp X \implies \tau_k \neq \tau_{k'}$
- Global ATE is given by $\tau = \sum_{k=1}^{K} \rho_k \tau_k$ with $\rho_k = \mathbb{P}(H = k)$

Summary of results

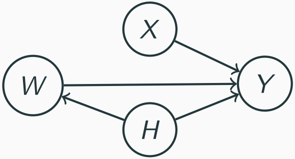
- $\hat{\tau}_{\mathrm{meta-IVW}}$ is biased because inverse variance weights give biased estimates of the ρ_k
- $\mathbb{V}^\infty(\hat{\tau}_{\mathrm{pool}}) = \mathbb{V}^\infty(\hat{\tau}_{\mathrm{GD}}) = \mathbb{V}^\infty(\hat{\tau}_{\mathrm{1S-IVW}}) \leq \mathbb{V}^\infty(\hat{\tau}_{\mathrm{meta-SW}})$
- $\hat{\tau}_{\mathrm{1S-SW}}$ is robust to heterogeneous covariances $\{\Sigma_k\}_k$ but has larger variance for different means $\{\mu_k\}_k$

Numerical illustration

$$X_k \sim \mathcal{N}(\mu_k, \Sigma_k)$$



Heterogeneity from Center Effects



- Studies may have varying practices or organizational contexts
- Model: **fixed effect of the source H onto the outcome Y**:

$$Y_{k,i}^{(w)} = c^{(w)} + h_k + X_{k,i}\beta^{(w)} + \varepsilon_i(w)$$

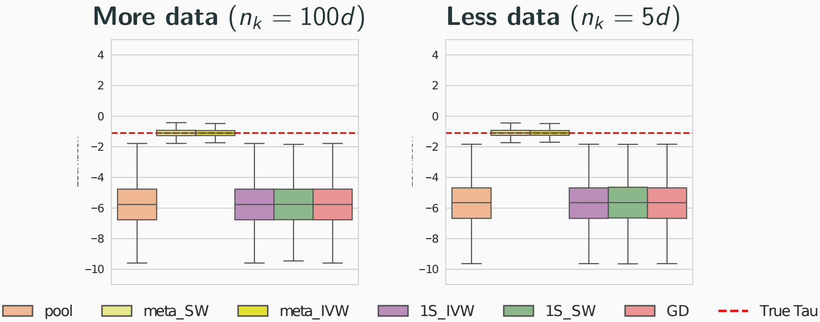
Note: CATEs $\mathbb{E}[Y(1) - Y(0)|X, H]$ are the same/sources

- Caution: H is now a confounder

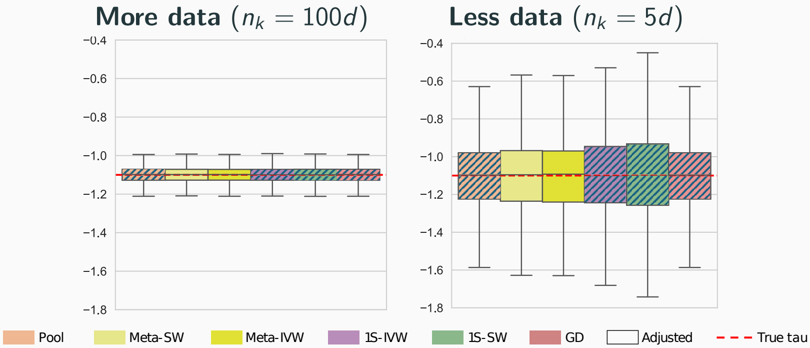
Summary of results

- $\hat{\tau}_{\mathrm{meta-SW}}$ and $\hat{\tau}_{\mathrm{meta-IVW}}$ naturally account for the center effects
- Other federated estimators are **biased** and need to be **adjusted**. GD estimators: add H as an additional covariate

Numerical illustration



Numerical illustration



Summary: decision diagram for practitioners

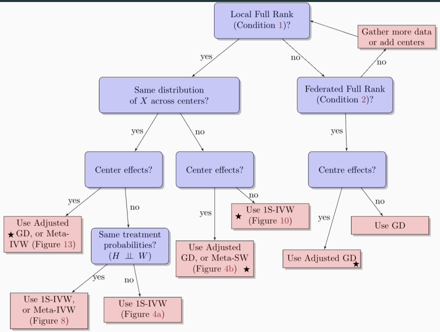


Figure 6: Decision Diagram for Practitionners. The sign ★ denotes scenarios where the DM estimator is biased.

Federated RCTs: Guidelines Meta & GD predilection regimes

- ▷ Small **sample size**: Gradient Descent: other need $n_k^{(w)} \geq d$ for k, w
- ▷ **Heterogeneity**: Shift across sources ($\hat{\tau}_{\text{meta}-\text{IVW}}$ biased); different baseline outcomes ($\hat{\tau}_{\text{meta}}$ handles center effects, $\hat{\tau}_{\text{GD}}$ needs adjustment/prior knowledge on the model)
- ▷ **Non collapsibility**: GD (step 2: estimate local $\mathbb{E}[Y(w)]$)

Federated Causal Inference/Generalization

Similarity between both problems

- FL: Target population defined as a mixture of K sites
-
-

Federated IPW for observational data

$$e(X_i) = \sum_{k=1}^{K} \mathbb{P}(H_i = k \cap W_i = 1 \mid X_i).$$

$$e(X_i) = \sum_{k=1}^{K} \underbrace{\rho_k \frac{\mathbb{P}(X_i \mid H_i = k)}{\mathbb{P}(X_i)}}_{\text{density weights}} e_k(X_i).$$

Federated Causal Inference/Generalization

Similarity between both problems

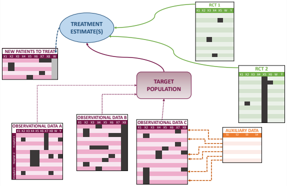
- FL: Target population defined as a mixture of K sites
- Same assumptions of transportability
- Same dichotomy of approaches between collapsible/non collapsible measures

Federated IPW for observational data

$$e(X_i) = \sum_{k=1}^{K} \mathbb{P}(H_i = k \cap W_i = 1 \mid X_i).$$

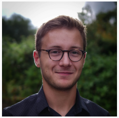
$$e(X_i) = \sum_{k=1}^{K} \underbrace{\rho_k \frac{\mathbb{P}(X_i \mid H_i = k)}{\mathbb{P}(X_i)}}_{\text{density weights}} e_k(X_i).$$

Future work: Multiple RCTs & Multiple Observational data



black correspond to sporadically & systematic missing covariates

- Implementation of a package
- Extension to population-level measures and individual-level ones[26]
- Complex outcome/treatment/features distributions, survival, time
- Federated Random Forests
- Provide robust privacy guarantees (differential privacy)

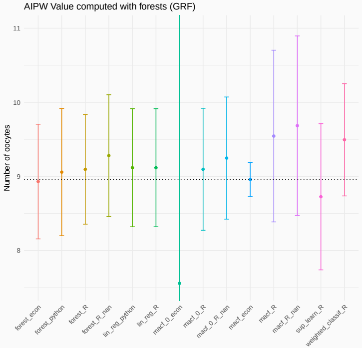
Clément Berenfeld, Ahmed Boudghiri, Rémi Khellaf, Aurelien Bellet, Erwan Scornet (Sorbone)

[26] Even, **J.J.** (2025). Rethinking the win ratio: causal framework for hierarchical outcome Analysis

Policy learning for personalized treatment

- **Policy estimation**
  - Counterfactual outcome estimation/CATE: T-learners, R-learner, X-learner, DR-learner, Causal Forest, etc.
  - Direct treatment rule estimation approach: Single stage outcome weighted learning, weighted classification
- **Policy evaluation**: Substitution estimator, AIPW, TMLE value

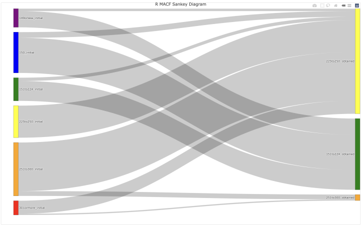
⇒ + Choice of learners (parametric/non param., etc) /software/ missing

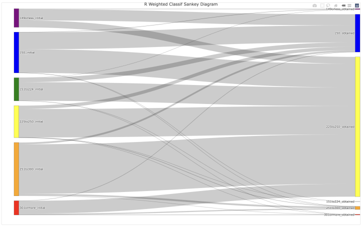


Policy learning for personalized treatment

- **Policy estimation**
  - Counterfactual outcome estimation/CATE: T-learners, R-learner, X-learner, DR-learner, Causal Forest, etc.
  - Direct treatment rule estimation approach: Single stage outcome weighted learning, weighted classification
- **Policy evaluation**: Substitution estimator, AIPW, TMLE value

$\Rightarrow +$ Choice of learners (parametric/non param., etc) /software/ missing





Recommended optimal dose never matched the one prescribed by the MD!

Identifying Gaps in the Literature

Mihaela's Quote:
"A big part in a researcher's workflow is to identify gaps in the literature."

- **Question: Should we also focus on consolidation?**
  - Too many methods and papers, leading to an overwhelming number of choices.
  - Users are lost in the multitude of options available.
- **Gap Between Theory and Practice:**
  - Many theoretical advancements do not translate effectively into practice.
- **Incentive Structures:**
  - Need for incentives for sustained/maintained software beyond just hosting code on GitHub.
  - Incentive for consolidation?

Importance of Careful Design Over New Methods

Key Considerations:

- Careful design is more critical than merely creating new methods.
- **Which data should be collected?**

Example 1: In Vitro Fertilization (IVF)

- Goal: Find the optimal dose of gonadotropin to maximize oocyte size.
- **Question:** Can we expect reliable results without understanding the patient's psychological state?

Example 2: Personalized Medicine

- Goal: Determine the best treatment for each individual.
- **Observation:** In oncology, similar profiles can have vastly different treatment outcomes.
- **Hypothesis:** External factors (e.g., exercise, acupuncture, dietary supplements, hypnosis) could influence outcomes.
- **Conclusion:** We must encourage the collection of such additional information.

The Limits of AutoML

Question: Can we rely on AutoML?

- **Lack of Contextual Information:**
  - Important information is missing from datasets, which is often uncovered through collaborative discussions.
  - This context affects how data is coded and interpreted.
- **Examples:**
  - Distribution changes in gravity scores due to funding tied to patient severity.
  - Missing values due to team disagreements; Orientation depends of trust/reputation

Context is crucial to **access algorithms**. Go beyond the model: what is its impact on all stakeholders?

Mihaela's Quote:
"Having clear communication between both parties may avoid researchers wasting valuable resources and time on problems that need not be solved."

Generalization of first moment population-level estimands

Transportability (Ignorability on trial participation)

$$\forall w \in \{0,1\} \quad \mathbb{E}_{\mathrm{R}}[Y(w) \mid X] = \mathbb{E}_{\mathrm{T}}[Y(w) \mid X]$$

Identifiability formulae

$$\mathbb{E}_{\mathrm{T}}\left[Y(w)\right] = \mathbb{E}_{\mathrm{T}}\left[\mathbb{E}_{\mathrm{R}}\left[Y(w)|X\right]\right]$$

Estimator: G-formula transported

$$\hat{\tau}_{\Phi,G} = \Phi\left(\frac{1}{m}\sum_{i=1}^{m}\mu_{(1)}^{R}(X_i), \frac{1}{m}\sum_{i=1}^{m}\mu_{(0)}^{R}(X_i)\right)$$

Estimator: Doubly robust

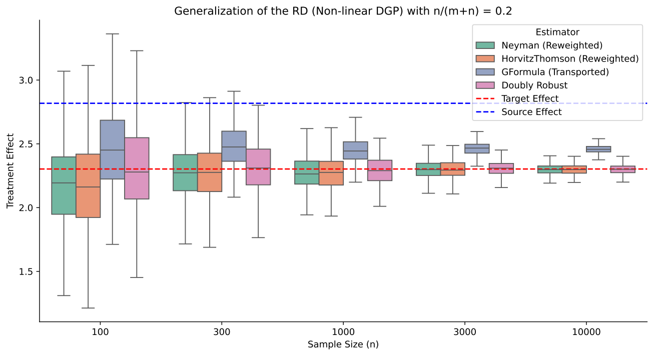
$$\hat{\tau}_{\Phi,DR}^{\pi,\boldsymbol{\beta}} = \Phi\left(\tilde{Y}(1), \tilde{Y}(0)\right)$$

$$\tilde{Y}(w) = \frac{1}{m}\sum_{i=1}^{m}\mu_{(w)}^{R}(X_i) + \frac{1}{n}\sum_{i=1}^{n} r(X_i,\boldsymbol{\beta})\mathbf{1}_{W_i=w}\frac{Y_i - \mu_{(w)}^{R}(X_i)}{\mathbb{P}_R(W=w)}$$

Generalization of RD, RR and OR

Under assumption $\mathbb{E}_{\text{R}}[Y(w) \mid X] = \mathbb{E}_{\text{T}}[Y(w) \mid X]$

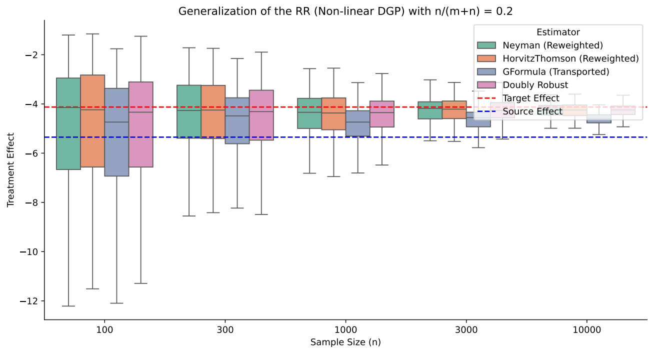
- $Y_{\text{R}}(w) = c(w) + X_{\text{R}}\beta(w) + \epsilon_{\text{R}(w)}$
- $X_{\text{R}} \sim \mathcal{N}(\mu_{\text{R}}, \Sigma)$
- $X_{\text{T}} \sim \mathcal{N}(\mu_{\text{T}}, \Sigma)$



Generalization of RD, RR and OR

Under assumption $\mathbb{E}_{\text{R}}[Y(w) \mid X] = \mathbb{E}_{\text{T}}[Y(w) \mid X]$

- $Y_{\text{R}}(w) = c(w) + X_{\text{R}}\beta(w) + \epsilon_{\text{R}(w)}$
- $X_{\text{R}} \sim \mathcal{N}(\mu_{\text{R}}, \Sigma)$
- $X_{\text{T}} \sim \mathcal{N}(\mu_{\text{T}}, \Sigma)$



Generalization of RD, RR and OR

Under assumption $\mathbb{E}_{\text{R}}[Y(w) \mid X] = \mathbb{E}_{\text{T}}[Y(w) \mid X]$

- $Y_{\text{R}}(w) = c(w) + X_{\text{R}}\beta(w) + \epsilon_{\text{R}(w)}$
- $X_{\text{R}} \sim \mathcal{N}(\mu_{\text{R}}, \Sigma)$
- $X_{\text{T}} \sim \mathcal{N}(\mu_{\text{T}}, \Sigma)$