Causal Alternatives to Meta-Analysis

Ahmed Boughdiri & Rémi Khellaf

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

Randomized Controlled Trial (RCT)

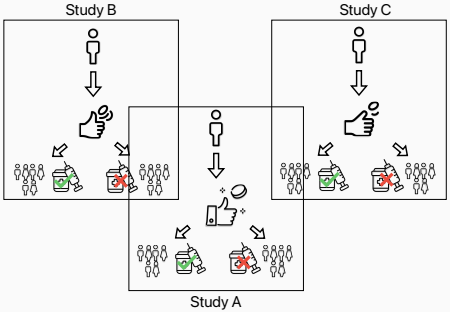
Estimating causal effect of treatment/intervention on outcome

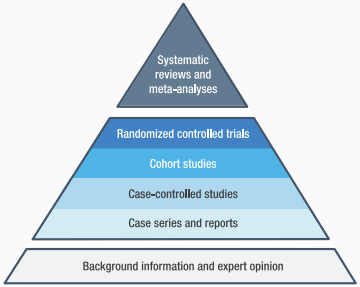
- **gold standard** (allocation )
- same covariate distributions in treated and control groups
  ⇒ High internal validity

- expensive, long, ethical limitations
- restrictive inclusion/exclusion criteria
  ⇒ **small sample size**. Ex: systematic review on 329 RCTs in surgery[1]: median sample size 144 patients
  ⇒ No personalized medicine
  ⇒ trial sample different from the population eligible for treatment
  ⇒ Low external validity

[1] Robinson et al. (2021) Characteristics of RCTs in Surgery From 2008 to 2020: A Systematic Review. *JAMA*.

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. But Meta-analysis still faces significant challenges:

- **Heterogeneity** across studies: sample size, population, center effects
- Lack of causal interpretation: What is the **target population?** How to aggregate causal measures?
- **Difficulty to share individual-level data**: **data silos** & **regulations**

3 alernatives to meta-analysis

Causal meta-analysis by defining a clear target population

- ▷ using aggregated data
- ▷ using individual decentralized data and **federated learning**
- ▷ using an external population and **generalization/transportability** methods

Comparing two average situations

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

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 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
- A 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
- A 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

A desirable property: collapsibility

Collapsibility: 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

$$\tau_R^{RD} = p_R(X=1) \times \tau_R^{RD}(X=1) + p_R(X=0) \times \tau_R^{RD}(X=0)$$
$$-0.0452 = -0.47 \times 0.006 - 0.53 \times 0.08.$$

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)[2] we propose: Weighting & outcome modeling estimators (asymptotic & finite-sample analyses) + Two doubly robust estimators via semi-parametric theory.

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

Causal Meta Analysis using aggregated data

The standard approach: fixed & random-effects meta-analysis

Context: $K \in \mathbb{N}$ studies outputing results of the form

(Study k)

	$Y = 1$	$Y = 0$
$A = 1$	$n_{11}(k)$	$n_{10}(k)$
$A = 0$	$n_{01}(k)$	$n_{00}(k)$

where A is the treatment assignation and Y the outcome.

The standard approach: fixed & random-effects meta-analysis

Context: $K \in \mathbb{N}$ studies outputing results of the form

(Study k)

	$Y=1$	$Y=0$
$A=1$	$n_{11}(k)$	$n_{10}(k)$
$A=0$	$n_{01}(k)$	$n_{00}(k)$

where A is the treatment assignation and Y the outcome.

Based on this table, we can compute:

1. A causal contrast θ_k (Risk Difference, (log-)Risk Ratio, Survival Ratio, Odds Ratio);
2. A estimated variance σ_k^2.

The standard approach: fixed & random-effects meta-analysis

Context: $K \in \mathbb{N}$ studies outputing results of the form

(Study k)

	$Y = 1$	$Y = 0$
$A = 1$	$n_{11}(k)$	$n_{10}(k)$
$A = 0$	$n_{01}(k)$	$n_{00}(k)$

where A is the treatment assignation and Y the outcome.

Based on this table, we can compute:

1. A causal contrast θ_k (Risk Difference, (log-)Risk Ratio, Survival Ratio, Odds Ratio);
2. A estimated variance σ_k^2.

Goal: Aggregate all θ_k.

The standard approach: fixed & random-effects meta-analysis

Fixed-effects models

$$\hat{\theta}_k \sim \mathcal{N}(\theta^*, \sigma_k^2),$$

for some unknown true effect θ^*, with σ_k^2 well approximated by $\hat{\sigma}_k^2$.

Random-effects models

$$\hat{\theta}_k | \theta_k \sim \mathcal{N}(\theta_k, \sigma_k^2) \quad \text{and} \quad \theta_k \sim \mathcal{N}(\theta^*, \tau^2).$$

and estimate τ^2 in some way.

The standard approach: fixed & random-effects meta-analysis

Fixed-effects models

$$\hat{\theta}_k \sim \mathcal{N}(\theta^*, \sigma_k^2),$$

for some unknown true effect θ^*, with σ_k^2 well approximated by $\hat{\sigma}_k^2$.

Random-effects models

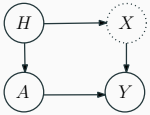
$$\hat{\theta}_k | \theta_k \sim \mathcal{N}(\theta_k, \sigma_k^2) \quad \text{and} \quad \theta_k \sim \mathcal{N}(\theta^*, \tau^2).$$

and estimate τ^2 in some way.

Classical meta-analysis aggregation: inverse variance weighting (IVW)

$$\hat{\theta}^{\text{meta}} = \sum_{k=1}^{K} \omega_k \hat{\theta}_k \quad \text{where} \quad \omega_k \propto \frac{1}{\hat{\sigma}_k^2 + \hat{\tau}^2} \quad \text{and} \quad \sum_{k=1}^{K} \omega_k = 1.$$

Causally interpretable estimand



- ▷ the study membership $H \in [K]$, X not observed
- ▷ $\mathbb{E}_k$: expectation in study k
- ▷ P_k: covariate distribution of study k

Assumption: Transportability of conditional expectation

for all $a \in \{0,1\}$ and $k, k' \in [K]$, $\quad \mathbb{E}_k[Y(a)|X] = \mathbb{E}_{k'}[Y(a)|X] \quad$ a.s.

$\Rightarrow$ No trials effect - distributional shift of the covariate

$$\mathbb{E}_k[Y(a)] = \mathbb{E}_k[\mathbb{E}_k[Y(a)|X]] = \mathbb{E}_k[\mathbb{E}[Y(a)|X]] = \mathbb{E}_{P_k}[\mu_a(X)],$$

θ_k can be written in the form $\theta_k = \theta(P_k)$

Definition: θ^* is causally interpretable

if it can be written as $\theta(P^*)$ for some distribution $P^* \Rightarrow \exists$ a population P^* for which θ^* is the associated population-level measure.

A lack of causal interpretation for standard meta-analysis?

Random model

- a random distribution P_k ($\sim$ from some prior distribution Π)
- $\theta_k = \theta(P_k)$ is random; the random-effects model $\theta_k \sim \mathcal{N}(\theta^{\text{meta}}, \tau^2)$
- θ^{meta}: mean of the θ_k's with respect to the randomness of the model

$$\theta^{\text{meta}} = \mathbb{E}[\theta(P_k)], \quad \text{where } P_k \sim \Pi \text{ for some distribution } \Pi \text{ on } \mathcal{P}(\mathcal{X})$$

A lack of causal interpretation for standard meta-analysis?

Random model

- a random distribution P_k ($\sim$ from some prior distribution Π)
- $\theta_k = \theta(P_k)$ is random; the random-effects model $\theta_k \sim \mathcal{N}(\theta^{\text{meta}}, \tau^2)$
- θ^{meta}: mean of the θ_k's with respect to the randomness of the model

$$\theta^{\text{meta}} = \mathbb{E}[\theta(P_k)], \quad \text{where } P_k \sim \Pi \text{ for some distribution } \Pi \text{ on } \mathcal{P}(\mathcal{X})$$

θ is a linear contrast (i.e. Risk Difference)

θ and the expectation commute: $\theta^{\text{meta}} = \mathbb{E}[\theta(P_k)] = \theta(\mathbb{E}[P_k])$

$\Rightarrow$ the average of the ATEs on each population is equal to the **treatment effect on the average population**

A lack of causal interpretation for standard meta-analysis?

Random model

- a random distribution P_k ($\sim$ from some prior distribution Π)
- $\theta_k = \theta(P_k)$ is random; the random-effects model $\theta_k \sim \mathcal{N}(\theta^{\text{meta}}, \tau^2)$
- θ^{meta}: mean of the θ_k's with respect to the randomness of the model

$$\theta^{\text{meta}} = \mathbb{E}[\theta(P_k)], \quad \text{where } P_k \sim \Pi \text{ for some distribution } \Pi \text{ on } \mathcal{P}(\mathcal{X})$$

θ is a linear contrast (i.e. Risk Difference)

θ and the expectation commute: $\theta^{\text{meta}} = \mathbb{E}[\theta(P_k)] = \theta(\mathbb{E}[P_k])$

$\Rightarrow$ the average of the ATEs on each population is equal to the **treatment effect on the average population**

θ is not a linear (i.e. (log-)Risk/Odds Ratios)

Not possible to express θ^{meta} in the form $\theta(P^*)$ for an univocally defined target population P^*

$\Rightarrow$ difficult to interpret θ^{meta} in a causal way

New approach: causally interpretable meta-analysis (CIMA)

The methodology

1. Define a target population P^* the one of all studies together:

$$P^* = \sum_{k \in [K]} \mathbb{E}\left[\frac{n_k}{n}\right] P_k \quad \text{with } n_k \text{ the number of people in the } k\text{-th study}$$

2. Use the full data $n_{ay}(k)$ to estimate $\theta(P^*)$.

New approach: causally interpretable meta-analysis (CIMA)

The methodology

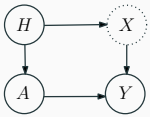
1. Define a target population P^* the one of all studies together:

$$P^* = \sum_{k \in [K]} \mathbb{E}\left[\frac{n_k}{n}\right] P_k \quad \text{with } n_k \text{ the number of people in the } k\text{-th study}$$

2. Use the full data $n_{ay}(k)$ to estimate $\theta(P^*)$.

(Study k)

	$Y=1$	$Y=0$
$A=1$	$n_{11}(k)$	$n_{10}(k)$
$A=0$	$n_{01}(k)$	$n_{00}(k)$



Collapsibility yields

$$\mathrm{RD}(P^*) = \mathbb{E}\left[\mathrm{RD}(P_H)\right] \qquad \mathrm{RR}(P^*) = \mathbb{E}\left[\frac{\mathbb{E}[Y_0|H]}{\mathbb{E}[Y_0]}\mathrm{RR}(P_H)\right]$$

Such formula exists for other more general first-moment causal contrast.

Causal meta-analysis

Let $\hat{\psi}_a(k) = n_{a1}(k)/n_a(k)$ the natural estimator for $\mathbb{E}[Y(a)|H = k]$

Causal Meta Risk Difference

$$\hat{\theta}_{\mathrm{RD}}^{\mathrm{causal}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\theta}_{\mathrm{RD}}(k) \quad \text{where} \quad \hat{\theta}_{\mathrm{RD}}(k) = \hat{\psi}_1(k) - \hat{\psi}_0(k)$$

$\Rightarrow$ weighted average of $\hat{\theta}_{\mathrm{RD}}(k)$ with $\neq$ weights than classical meta

Causal Meta Risk Ratio

$$\hat{\theta}_{\mathrm{RR}}^{\mathrm{causal}} = \frac{\hat{\psi}_1}{\hat{\psi}_0} \quad \text{where} \quad \hat{\psi}_a := \sum_{k=1}^{K} \frac{n_k}{n} \hat{\psi}_a(k)$$

$\Rightarrow$ weighted average of $\hat{\theta}_{\mathrm{RR}}(k)$, with collapsibility weights

$\Rightarrow$ **averages both outcomes separately**

Causal Meta Odds Ratio

$\hat{\theta}_{\mathrm{OR}}^{\mathrm{causal}} = \frac{\hat{\psi}_1}{1-\psi_1} \times \frac{1-\hat{\psi}_0}{\hat{\psi}_0} \Rightarrow$ Not a weighted average of the per-study estimates but of separate outcomes

Difference between classical and causal meta-analysis

Ex log-RR: $\mathrm{RR}_{\mathrm{Meta}} = \exp \mathbb{E}[\log \mathrm{RR}(H)]$.

That is geometric mean of the stratified risk ratios.

Ex: a 2-study $\mathbb{P}(H = 1) = \mathbb{P}(H = 2) = 1/2$, $\psi_a(k) = \mathbb{E}[Y(a)|H = k]$

$$\theta_{\mathrm{RR}}^{\mathrm{meta}} = \sqrt{\frac{\psi_1(1)\psi_1(2)}{\psi_0(1)\psi_0(2)}} \quad \text{and} \quad \theta_{\mathrm{RR}}^{\mathrm{causal}} = \frac{\psi_1(1) + \psi_1(2)}{\psi_0(1) + \psi_0(2)}.$$

No general bounds that compare $\theta_{\mathrm{RR}}^{\mathrm{meta}}$ and $\theta_{\mathrm{RR}}^{\mathrm{causal}}$

Difference between classical and causal meta-analysis

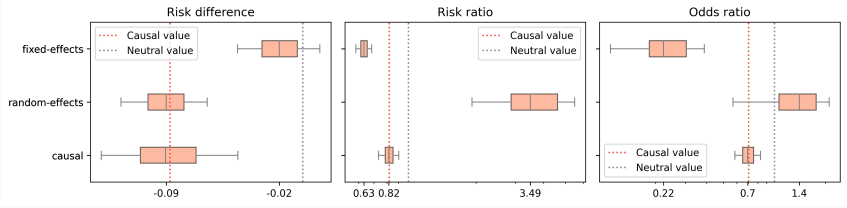
Ex log-RR: $\mathrm{RR}_{\mathrm{Meta}} = \exp \mathbb{E}[\log \mathrm{RR}(H)]$.

That is geometric mean of the stratified risk ratios.

Ex: a 2-study $\mathbb{P}(H = 1) = \mathbb{P}(H = 2) = 1/2$, $\psi_a(k) = \mathbb{E}[Y(a)|H = k]$

$$\theta_{\mathrm{RR}}^{\mathrm{meta}} = \sqrt{\frac{\psi_1(1)\psi_1(2)}{\psi_0(1)\psi_0(2)}} \quad \text{and} \quad \theta_{\mathrm{RR}}^{\mathrm{causal}} = \frac{\psi_1(1) + \psi_1(2)}{\psi_0(1) + \psi_0(2)}.$$

No general bounds that compare $\theta_{\mathrm{RR}}^{\mathrm{meta}}$ and $\theta_{\mathrm{RR}}^{\mathrm{causal}}$



Comparison of random effects meta-analysis with CIMA

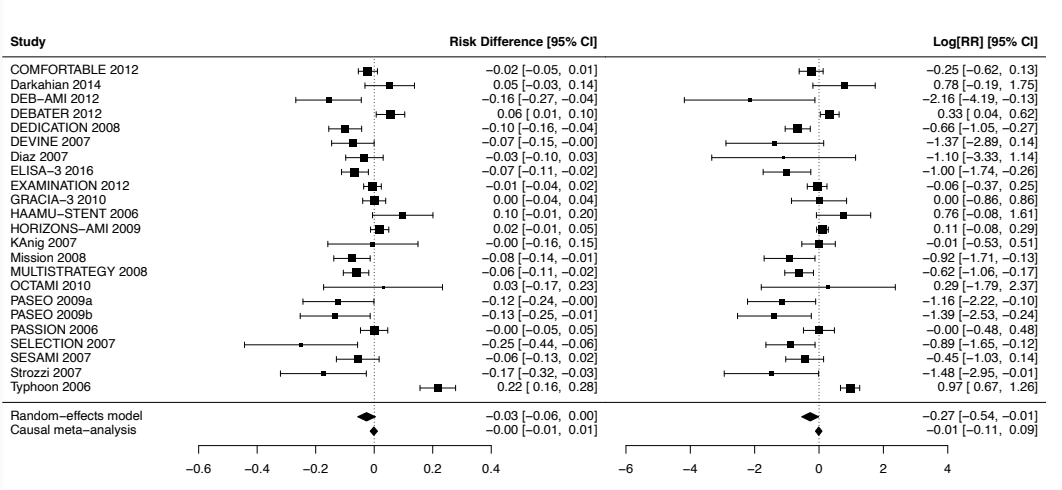


Figure 1: Effect of drug-eluting versus bare-metal stents in acute coronary syndrome

Comparison of random effects meta-analysis with CIMA

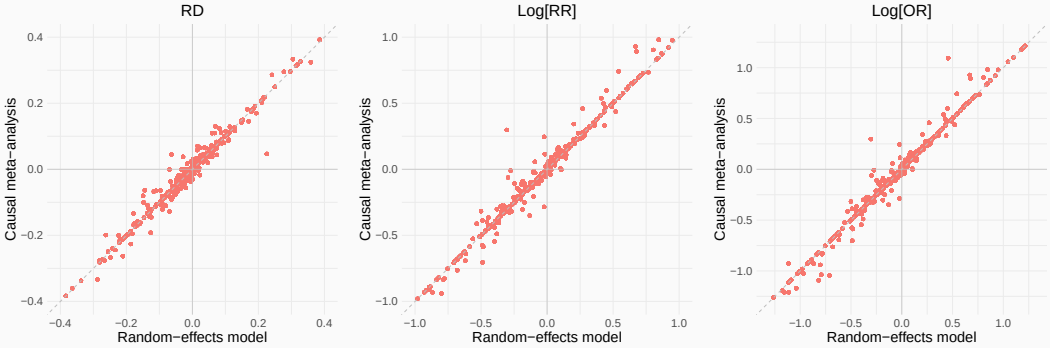


Figure 1: Comparison across 597 real-world meta-analyses from Cochrane Library.

Causal meta-analysis from individual decentralized data

Going beyond meta-analysis with federated causal inference

Algorithm 1: FedAvg (server-side)

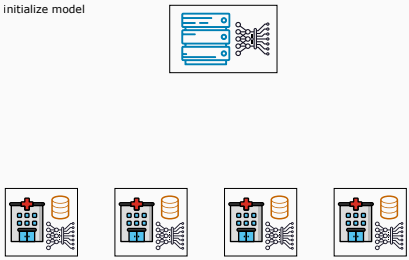
initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

Algorithm 2: $\textsc{ClientUpdate}(k, \theta)$

$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

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

Going beyond meta-analysis with federated causal inference



Algorithm 3: FedAvg (server-side)

initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

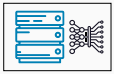
Algorithm 4: $\textsc{ClientUpdate}(k, \theta)$

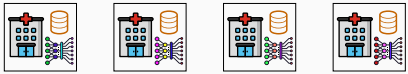
$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

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

Going beyond meta-analysis with federated causal inference

each party makes an update using its local dataset





Algorithm 5: FedAvg (server-side)

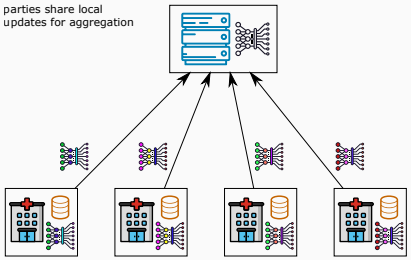
initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

Algorithm 6: $\textsc{ClientUpdate}(k, \theta)$

$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

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

Going beyond meta-analysis with federated causal inference



Algorithm 7: FedAvg (server-side)

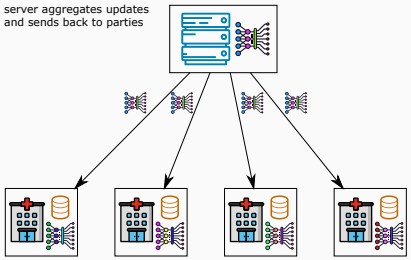
initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

Algorithm 8: $\textsc{ClientUpdate}(k, \theta)$

$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

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

Going beyond meta-analysis with federated causal inference



Algorithm 9: FedAvg (server-side)

initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

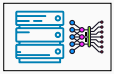
Algorithm 10: $\textsc{ClientUpdate}(k, \theta)$

$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

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

Going beyond meta-analysis with federated causal inference

parties update their copy of the model and iterate



Algorithm 11: FedAvg (server-side)

initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

Algorithm 12: $\textsc{ClientUpdate}(k, \theta)$

$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

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

Going beyond meta-analysis with federated causal inference

- T comm. rounds: larger improves accuracy but increases comm. cost. Typically $100 - 1000$ for deep learning models.
- E local updates: larger improves local convergence but can cause drift in heterogeneous settings. $1 - 5$ works well.
- η learning rate: typically $0.01 - 0.1$ for logistic regression, $0.001 - 0.01$ for deep learning models.

Algorithm 13: FedAvg (server-side)

initialize global model parameters θ_0 each round $t = 1$ to T each client $k \in K$ in parallel $\theta_k \leftarrow$ $\textsc{ClientUpdate}(k, \theta)$ $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

Algorithm 14: $\textsc{ClientUpdate}(k, \theta)$

$\theta^{(k)} \leftarrow \theta$ local step $e = 1$ to E $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$ compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$ return $\theta^{(k)}$ to server

Same principle to estimate global outcome models μ_1, μ_0

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

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 E steps until convergence

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

$\eta < \frac{2}{L}$, 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 $E = 1$ 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 & E. $T = 1$ & $E \to \infty$: One-shot federated learning, meta analysis on β

Our setting: decentralized heterogeneous RCTs

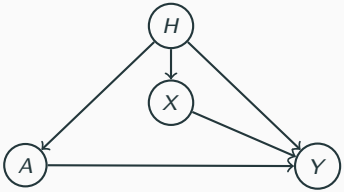
We consider K **decentralized and potentially heterogeneous** RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

Source	Obs.	Covariates			Treat.	Outcome
H	i	X_1	X_2	X_3	A	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	A	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

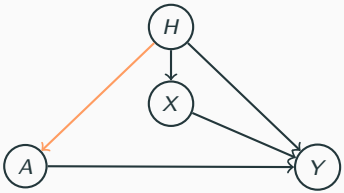


Our setting: decentralized heterogeneous RCTs

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	A	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

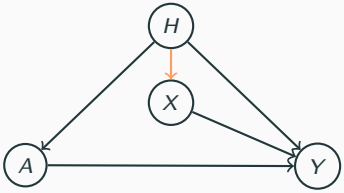


Our setting: decentralized heterogeneous RCTs

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	A	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

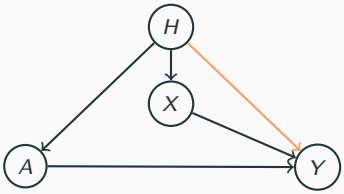


Our setting: decentralized heterogeneous RCTs

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	A	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

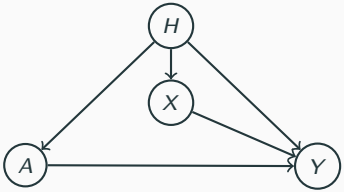
Presence of
center effects



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	A	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



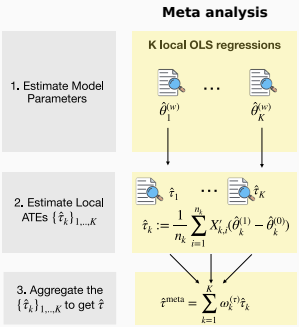
How to estimate τ without pooling together individual-level data?

Three types of federated estimators - collapsible measures (RD)

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

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

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



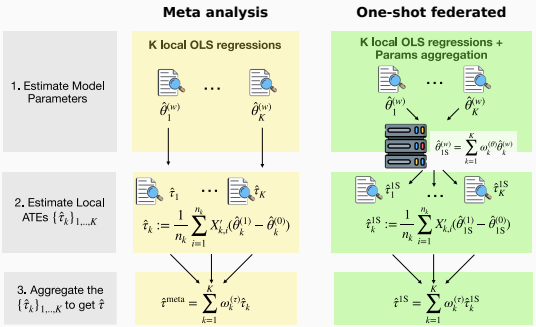
Aggregation w_k: sample size weights (SW, causal) or inverse variance (IVW)

Three types of federated estimators - collapsible measures (RD)

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

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

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



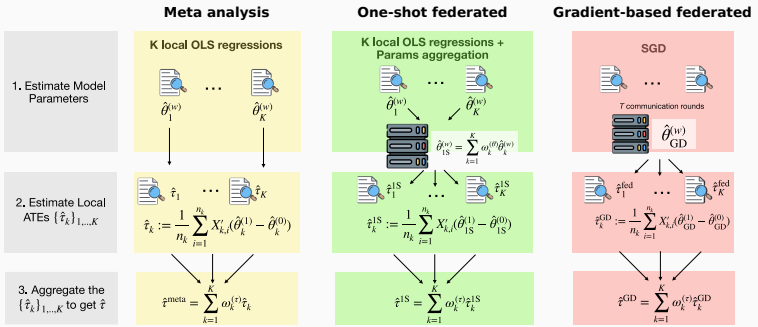
Aggregation w_k: sample size weights (SW, causal) or inverse variance (IVW)

Three types of federated estimators - collapsible measures (RD)

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

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

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



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

Statistical perf. & communication costs

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

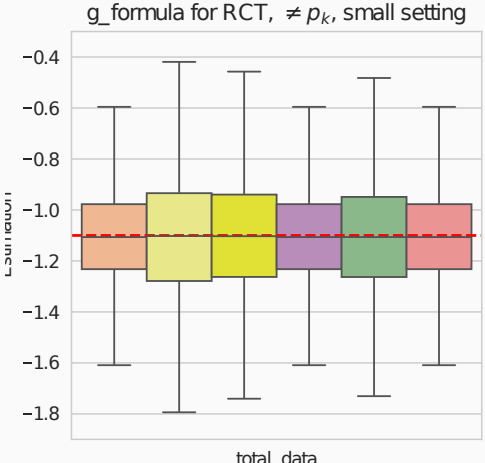
Unbiased estimators but different asymptotic variance & communication costs:

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

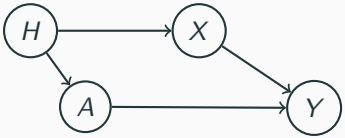
with $\rho_k = \mathbb{P}(H=k)$ and $p=\sum_{k=1}^{K}\frac{n_k}{n}p_k$

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



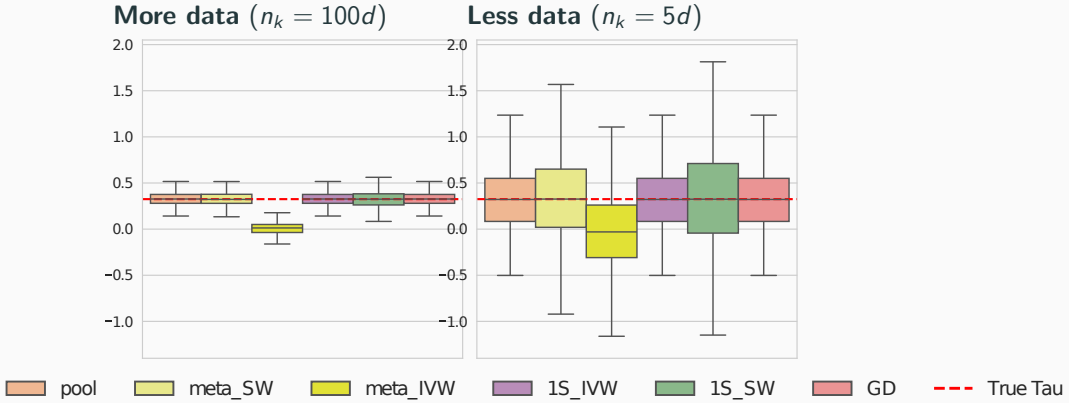
- **Distributional shift** across sources $\implies \tau_k \neq \tau_{k'}$
- Global ATE is given by $\tau = \sum_{k=1}^{K} \rho_k \tau_k$ with $\rho_k = \mathbb{P}(H = k)$

Summary of results

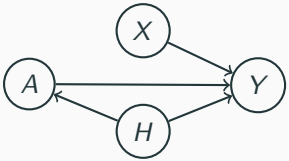
- $\hat{\tau}_{\text{meta-IVW}}$ is biased because inverse variance weights give biased estimates of the ρ_k
- $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{GD}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-IVW}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{meta-SW}})$
- $\hat{\tau}_{\text{1S-SW}}$ 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**:

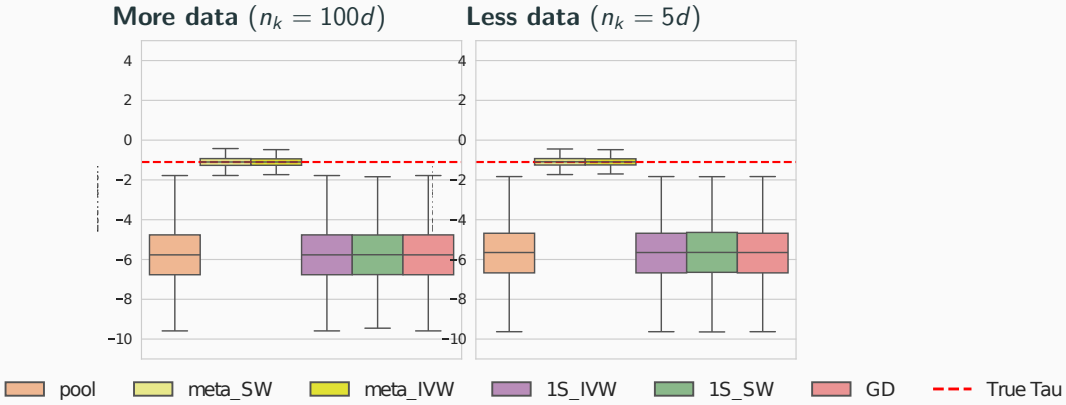
$$Y_{k,i}^{(a)} = c^{(a)} + h_k + X_{k,i}\beta^{(a)} + \varepsilon_i(a)$$

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

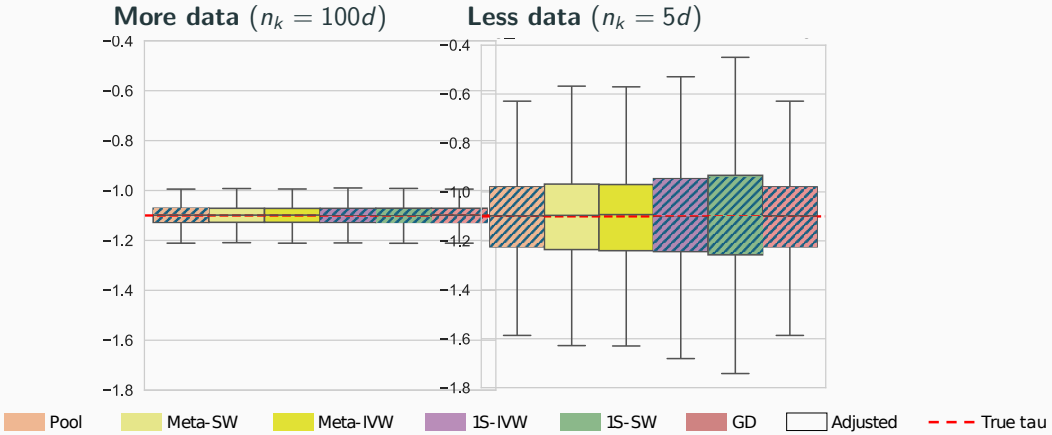
Summary of results

- $\hat{\tau}_{\text{meta}-\text{SW}}$ and $\hat{\tau}_{\text{meta}-\text{IVW}}$ 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

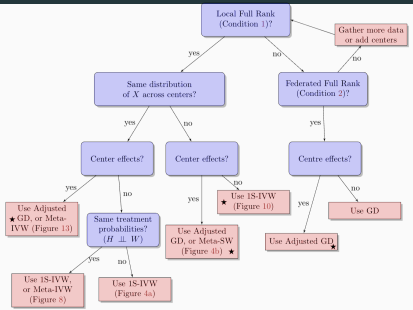


Figure 6: Decision Diagram for Practitionners. The sign ★ denotes scenarios where the DM estimator is biased.

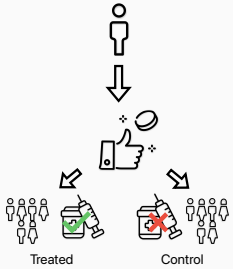
Federated RCTs: Guidelines Meta & GD predilection regimes

- Small **sample size**: Gradient Descent: other need $n_k^{(a)} \geq d$ for k, a
- **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**: step 2: estimate local $\mathbb{E}[Y(a)]$

The future of meta-analysis: using external data Generalization from RCTs to target population

Introduction to Generalization

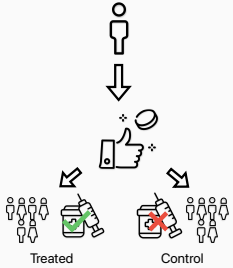
Randomized Controlled Trials:

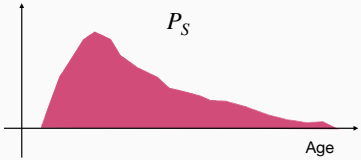


+ direct causal association
− selective population, small sample, not always feasible

Introduction to Generalization

Randomized Controlled Trials:

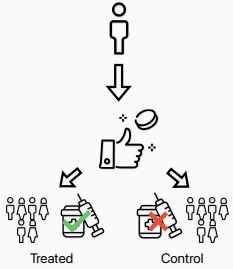


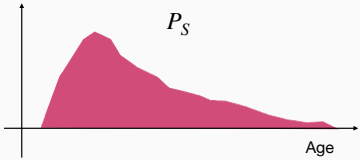


+ direct causal association
− selective population, small sample, not always feasible

Introduction to Generalization

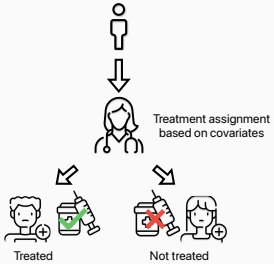
Randomized Controlled Trials:





+ direct causal association
− selective population, small sample, not always feasible

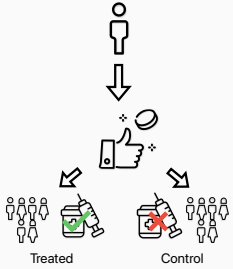
Observational Data:

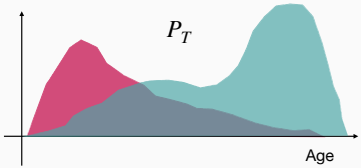


+ abundant, representative population
− confounding factors

Introduction to Generalization

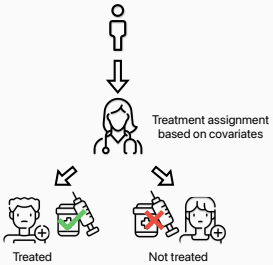
Randomized Controlled Trials:





+ direct causal association
− selective population, small sample, not always feasible

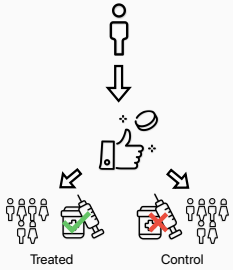
Observational Data:



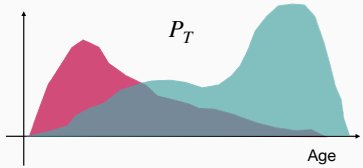
+ abundant, representative population
− confounding factors

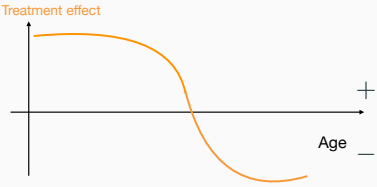
Introduction to Generalization

Randomized Controlled Trials:

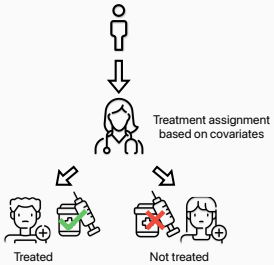


+ direct causal association
− selective population, small sample, not always feasible





Observational Data:



+ abundant, representative population
− confounding factors

$$p_S(x) \neq p_T(x) \Rightarrow \underbrace{\text{ATE in the RCT}}_{>0} \neq \underbrace{\text{Target ATE}}_{<0}$$

Problem setting: Generalization from one RCT to a Target pop.

Goal: estimate the treatment effect on the target population.

Source	Obs.	Covariates			Treatment	Outcomes	Potential Outcomes	
S	i	X^1	X^2	X^3	A	Y	$Y^{(1)}$	$Y^{(0)}$
0	1	37	2.0	F	0	1.7	??	1.7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	m	52	1.7	M	1	2.4	2.4	??

IPW, G-formula, AIPW under $Y(1), Y(0) \perp A \mid X \implies$ Strong assumption !

Problem setting: Generalization from one RCT to a Target pop.

Generalization: estimate the treatment effect on the target population using the RCT

Source	Obs.	Covariates			Treatment	Outcomes	Potential Outcomes	
S	i	X^1	X^2	X^3	A	Y	$Y^{(1)}$	$Y^{(0)}$
1	1	23	1.5	M	1	3.2	3.2	??
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	n	17	2.9	M	0	1.5	??	1.5
0	1	37	2.0	F	??	??	??	??
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	m	52	1.7	M	??	??	??	??

Note that here we do not need the treatment and the outcome in the target population.

The age-old question of how to report treatment effects

Count the dead

$$\tau_{RR} = \frac{\mathbb{E}\left[Y^{(1)}\right]}{\mathbb{E}\left[Y^{(0)}\right]}$$

Count the living

$$\tau_{SR} = \frac{1 - \mathbb{E}\left[Y^{(1)}\right]}{1 - \mathbb{E}\left[Y^{(0)}\right]}$$

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

Risk Difference

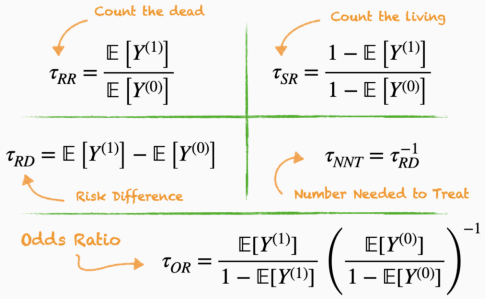
$$\tau_{NNT} = \tau_{RD}^{-1}$$

Number Needed to Treat

Odds Ratio

$$\tau_{OR} = \frac{\mathbb{E}[Y^{(1)}]}{1 - \mathbb{E}[Y^{(1)}]} \left(\frac{\mathbb{E}[Y^{(0)}]}{1 - \mathbb{E}[Y^{(0)}]} \right)^{-1}$$

The age-old question of how to report treatment effects



Risk Ratio, odds ratio, risk difference...
Which causal measure is easier to generalize?

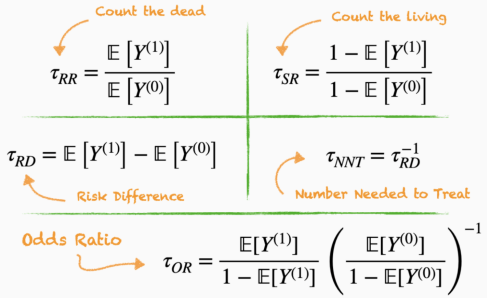
Bénédicte Colnet *Julie Josse* *Gael Varoquaux* *Erwan Scornet*

Measure	Dir. collapsible	Collapsible	Logic-respecting
RD	Yes	Yes	Yes
NNT	No	No	Yes
RR	No	Yes	Yes
SR	No	Yes	Yes
OR	No	No	No

e.g. Huitfeldt et al., 2021; Lapointe-Shaw et al., 2022; Liu et al., 2022; Colnet, et al. J.J. 2022; Colnet, J.J et al. 2023; Boughdiri, J.J et al 2025; Dumas, E., Stensrud (2025) ...

• **CONSORT guidelines recommend to report all of them** •

The age-old question of how to report treatment effects



Risk Ratio, odds ratio, risk difference...
Which causal measure is easier to generalize?

Bénédicte Colnet *Julie Josse* *Gael Varoquaux* *Erwan Scornet*

Measure	Dir. collapsible	Collapsible	Logic-respecting
RD	Yes	Yes	Yes
NNT	No	No	Yes
RR	No	Yes	Yes
SR	No	Yes	Yes
OR	No	No	No

e.g. Huitfeldt et al., 2021; Lapointe-Shaw et al., 2022; Liu et al., 2022; Colnet, et al. J.J. 2022; Colnet, J.J et al. 2023; Boughdiri, J.J et al 2025; Dumas, E., Stensrud (2025) ...

• **CONSORT guidelines recommend to report all of them** •

Existing studies have generalized the RD but not other causal measures. Here, we propose

A Unified Framework for the Transportability of Population-Level Causal Measures

First moment population-level measure

- τ^P a 1st moment population-level[3] measure if $\exists\ \Phi : D_\Phi \to \mathbb{R}$, $D_\Phi \subset \mathbb{R}^2$

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

Note that a 1st moment population-level measure depends on a population P: $\tau^S_\Phi \neq \tau^T_\Phi$

[3] Fay & Li. (2024). Causal interpretation of the hazard ratio in RCTs. *Clinical Trials*.

First moment population-level measure

- τ^P a 1st moment population-level[3] measure if $\exists\ \Phi : D_\Phi \to \mathbb{R}$, $D_\Phi \subset \mathbb{R}^2$

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

Note that a 1st moment population-level measure depends on a population P: $\tau^S_\Phi \neq \tau^T_\Phi$

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	$\Phi(x, y) = \frac{1}{x-y}$	$\{(x, y) \in \mathbb{R}^2 \mid x - y \neq 0\}$

[3] Fay & Li. (2024). Causal interpretation of the hazard ratio in RCTs. *Clinical Trials*.

First moment population-level measure

- τ^P a 1st moment population-level[3] measure if $\exists\ \Phi : D_\Phi \to \mathbb{R}$, $D_\Phi \subset \mathbb{R}^2$

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

Note that a 1st moment population-level measure depends on a population P: $\tau_\Phi^S \neq \tau_\Phi^T$

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	$\Phi(x, y) = \frac{1}{x-y}$	$\{(x, y) \in \mathbb{R}^2 \mid x - y \neq 0\}$

- In contrast an individual-level measure depends on the joint distribution. Most are non identifiable but workarounds exist. Ex: $\mathbb{E}\left[\frac{Y_i(1)}{Y_i(0)}\right] \neq \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Y_i(0)]}$ or $\mathbb{P}[Y(1) > Y(0)]$

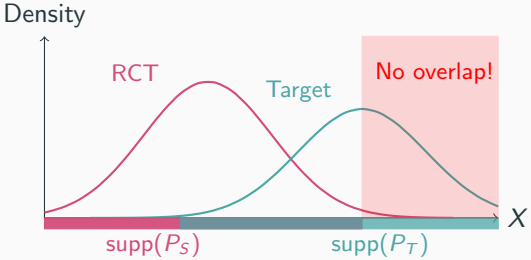
[3] Fay & Li. (2024). Causal interpretation of the hazard ratio in RCTs. *Clinical Trials*.

Assumptions for ATE identifiability in generalization

Overlap

$\forall x \in \mathbb{X}, p_S(x) > 0$ and

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



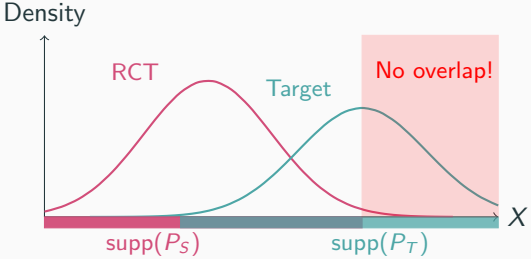
Intuition: Every covariate profile in the target population must be represented in the RCT. We cannot generalize on people not represented in S

Assumptions for ATE identifiability in generalization

Overlap

$\forall x \in \mathbb{X}, p_S(x) > 0$ and

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



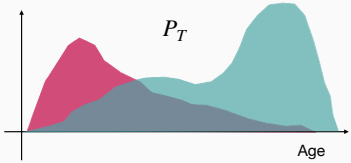
Intuition: Every covariate profile in the target population must be represented in the RCT. We cannot generalize on people not represented in S

Exchangeability in mean

$\forall a \in \{0, 1\}$,

$$\mathbb{E}_S[Y(a) \mid X] = \mathbb{E}_T[Y(a) \mid X]$$

In general $\mathbb{E}_S[Y(a)] \neq \mathbb{E}_T[Y(a)]$ since:



- what about: $\mathbb{E}_S[Y(a)|age] = \mathbb{E}_T[Y(a)|age]$?
- what if we have $p_s(weight) \neq p_T(weight)$?

Intuition: X must contain all shifted and prognostic covariates.

Reweighting the RCT: reweight Horvitz-Thomson

Reweighted Horvitz-Thomson

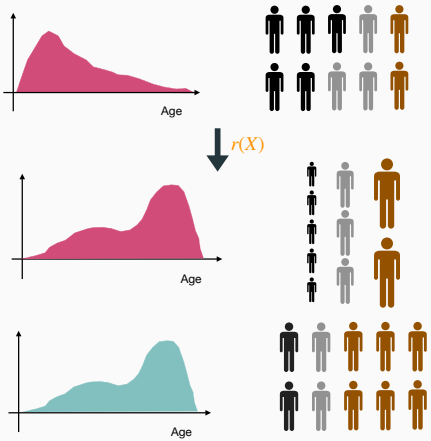
$$\hat{\tau}_{\Phi,\text{wHT}} = \Phi\left(\frac{1}{n}\sum_{S_i=1}\hat{r}(X_i)\frac{A_i Y_i}{\pi}, \frac{1}{n}\sum_{S_i=1}\hat{r}(X_i)\frac{(1-A_i)Y_i}{1-\pi}\right)$$

Estimate the **ratio of densities** with a logistic regression

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

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

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



Transport the RCT: G-formula

Fit models on RCT data
$\hat{\mu}_a^S(X) = \mathbb{E}_S[Y|A = a, X]$

↓

Predict on the target data
$Y(a) = \hat{\mu}_a^S(X)$ where $X \sim P_T(X)$

↓

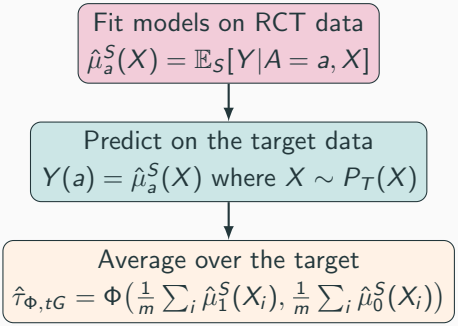
Average over the target
$\hat{\tau}_{\Phi,tG} = \Phi\left(\frac{1}{m}\sum_i \hat{\mu}_1^S(X_i), \frac{1}{m}\sum_i \hat{\mu}_0^S(X_i)\right)$

X^1	X^2	X^3	$Y^{(1)}$	$Y^{(0)}$
37	2.0	F	$\hat{\mu}_0^S(X_1)$	$\hat{\mu}_1^S(X_1)$
⋮	⋮	⋮	⋮	⋮
52	1.7	M	$\hat{\mu}_0^S(X_m)$	$\hat{\mu}_1^S(X_m)$

Transported G-formula

$$\hat{\tau}_{\Phi,\mathrm{tG}} = \Phi\left(\frac{1}{m}\sum_{S_i=0} \hat{\mu}^{\mathrm{S}}_{(1)}(X_i), \frac{1}{m}\sum_{S_i=0} \hat{\mu}^{\mathrm{S}}_{(0)}(X_i)\right)$$

Transport the RCT: G-formula



X^1	X^2	X^3	$Y^{(1)}$	$Y^{(0)}$
37	2.0	F	$\hat{\mu}_0^S(X_1)$	$\hat{\mu}_1^S(X_1)$
⋮	⋮	⋮	⋮	⋮
52	1.7	M	$\hat{\mu}_0^S(X_m)$	$\hat{\mu}_1^S(X_m)$

Transported G-formula

$$\hat{\tau}_{\Phi,\mathrm{tG}} = \Phi\left(\frac{1}{m}\sum_{S_i=0}\hat{\mu}^{\mathrm{S}}_{(1)}(X_i), \frac{1}{m}\sum_{S_i=0}\hat{\mu}^{\mathrm{S}}_{(0)}(X_i)\right)$$

We use data from the RCT to train $\hat{\mu}_{(1)}$ and $\hat{\mu}_{(0)}$ using

- ▷ Linear Regressions
- ▷ Random Forests

Proposition

Under a logistic $S|X$ and linear $Y(a)|X$ model for respectively the source and the outcome we have,

$$V^{\mathrm{OLS}}_{\Phi,\mathrm{tG}} \leq V_{\Phi,\mathrm{HT}},$$

Doubly robust estimator

Estimated equation estimator

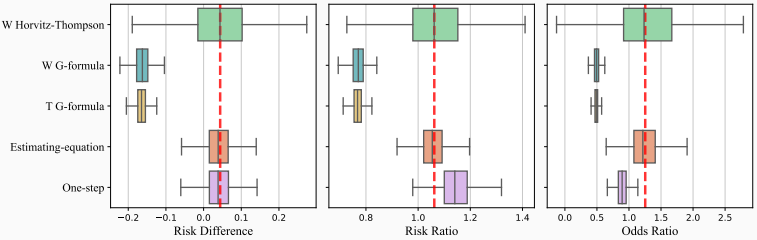
Given estimators $\hat{\mu}_{(a)}$ (resp. $\hat{r}$) of $\mu_{(a)}$ (resp. r), an estimating equation estimator $\hat{\tau}_{\Phi}^{\mathrm{EE}}$ of τ_{Φ} is given by $\hat{\tau}_{\Phi}^{\mathrm{EE}} = \Phi(\hat{\psi}_1^{\mathrm{EE}}, \hat{\psi}_0^{\mathrm{EE}})$ where for all $a \in \{0, 1\}$

$$\hat{\psi}_a^{\mathrm{EE}} := \frac{1}{m} \sum_{S_i=0} \hat{\mu}_{(a)}(X_i) + \frac{1}{n} \sum_{S_i=1} \frac{1\{A = a\}}{\mathbb{P}(A = a)} \hat{r}(X_i)(Y - \hat{\mu}_{(a)}(X_i))$$

Doubly Robust: The estimator $\hat{\tau}_{\Phi}^{\mathrm{EE}}$ is consistent as soon as either $\hat{\mu}_{(a)} = \mu$ or $\hat{r} = r$.

Robust to miss-specification :

- ▷ Logistic model on $S|X$
- ▷ Non-linear model on $Y(a)|X$



Doubly robust estimator

Estimated equation estimator

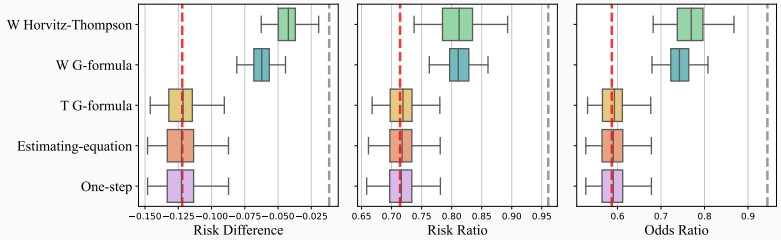
Given estimators $\hat{\mu}_{(a)}$ (resp. $\hat{r}$) of $\mu_{(a)}$ (resp. r), an estimating equation estimator $\hat{\tau}_\Phi^{\mathrm{EE}}$ of τ_Φ is given by $\hat{\tau}_\Phi^{\mathrm{EE}} = \Phi(\hat{\psi}_1^{\mathrm{EE}}, \hat{\psi}_0^{\mathrm{EE}})$ where for all $a \in \{0,1\}$

$$\hat{\psi}_a^{\mathrm{EE}} := \frac{1}{m}\sum_{S_i=0} \hat{\mu}_{(a)}(X_i) + \frac{1}{n}\sum_{S_i=1} \frac{1\{A=a\}}{\mathbb{P}(A=a)}\hat{r}(X_i)(Y - \hat{\mu}_{(a)}(X_i))$$

Doubly Robust: The estimator $\hat{\tau}_\Phi^{\mathrm{EE}}$ is consistent as soon as either $\hat{\mu}_{(a)} = \mu$ or $\hat{r} = r$.

Robust to miss-specification :

- ▷ Non-Logistic model on $S|X$
- ▷ Linear model on $Y(a)|X$



Relaxing exchangeability in mean

Exchangeability in mean

$\forall a \in \{0,1\}$,

$$\mathbb{E}_S[Y(a) \mid X] = \mathbb{E}_T[Y(a) \mid X]$$

X contains shifted prognostic variables

Exchangeability in effect measure

For a given ϕ, we have

$$\tau^{\mathrm{R}}_{\Phi}(x_c) = \tau^{\mathrm{T}}_{\Phi}(x_c)$$

X_c contains all shifted eï¬¬ectmodiï¬¬ers.$X_c \subseteq X$

If $Y^{(0)}$ is known in the target population, then the target treatment effect (for a given Φ) is **identifiable:**

$$\tau^T_\Phi = \Phi\left(\mathbb{E}_T\left[\Gamma\left(\tau^S_\Phi(X_c), \mu^T_{(0)}(X_c)\right)\right], \mathbb{E}_T\left[Y^{(0)}\right]\right)$$

where Γ is the inverse of $\psi_1 \mapsto \Phi(\psi_1, \psi_0)$ It leads naturally to:

- Weighted estimators
- Regression-based estimators
- **Doubly robust** estimators combining both approaches

Estimate the treament effect on the Target pop.

	Observational studies	Gen with Conditional Outcome	Gen with Local effects
Ass.	$Y(1), Y(0) \perp A \mid X$	$\mathbb{E}_{\mathrm{R}}[Y(a) \mid X] = \mathbb{E}_{\mathrm{T}}[Y(a) \mid X]$	$\tau_{\Phi}^{\mathrm{R}}(x) = \tau_{\Phi}^{\mathrm{T}}(x)$
Var.	confounding variables	shifted prognostic covariates	shifted effect modifiers $Y(0)$ in the target population
Meas.	RD, RR[4] but can be extended	RD, RR, NNT, OR, SR, ...	only Φ

[4] Boughdiri, J.J., Scornet. Estimating Risk Ratios in Causal Inference. ICML2025

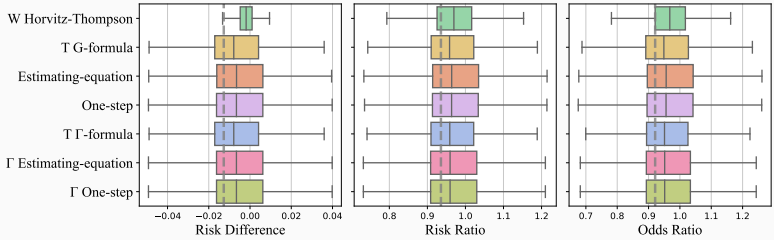
Generalizing CRASH-3 findings to the Traumabase population

CRASH-3 trial (Mostly Pakistan)

- Randomized trial ($n \approx 9{,}000$)
- Patients with TBI, GCS ≤ 12, within 3h
- Treatment: Tranexamic Acid (TXA)
- Outcome: Head injury-related death at 28 days

Traumabase cohort (France)

- Observational registry ($m \approx 8{,}200$)
- Selected CRASH-3-eligible patients
- Treatment: Tranexamic Acid (TXA)
- Deleterious/No evidence



[4] Colnet, J.J et al (2023). *Causal inference methods for combining randomized trials and observational studies: a review.*

Conclusion & Perspectives

- Identification
  - **Exchangeability in mean/effect measure:** $\mathbb{E}_{\text{S}}[Y(a) \mid X] = \mathbb{E}_{\text{T}}[Y(a) \mid X]$ or $\tau_{\Phi}^{\text{S}}(x) = \tau_{\Phi}^{\text{T}}(x)$
  - **Overlap:** $\text{supp}(P_T(X)) \subseteq \text{supp}(P_S(X))$

What we did:

- Generalized RD, RR and OR under Overlap and Exchangeability.
- Build and studied weighted, regression and doubly robust estimators.
- Applied this to **transported** the effect of TXA using CRASH-3 and Traumabase.

Thank you for your attention!
Questions?