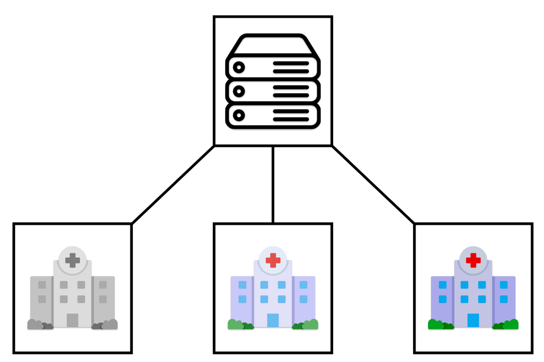
Federated Causal Inference: Estimating the Average Treatment Effect in Decentralized Settings

Rémi Khellaf [1] Aurélien Bellet [1] Julie Josse [1]

[1]INRIA, PreMeDICaL team

Motivation



Estimating the **Average Treatment Effect** (ATE) of a drug/policy/intervention requires **extensive amounts of data** within the counterfactual framework. **Federated Learning** (FL) lets agents **collaborate without sharing their data**. This work leverages FL to study ATE across siloed datasets. We build **federated estimators**, and compare them with pooled-dataset estimators.

Use-case examples: effect of a new drug on a rare disease across several hospitals, increasing minimum wage on employment rate across governments, changing credit scoring rules on clients churn across banks...

Federated Causal Inferences Framework and Modelling

Goal: Estimate the ATE defined as a risk difference $\tau := \mathbb{E}\left[\mathbb{E}\left[Y(1) - Y(0)\right] \mid H\right]$ over K studies, whose individual-level data cannot be exchanged.

- **Data structure**:
  - $X \in \mathbb{R}^{n \times d}$ the n i.i.d. observations of the covariates
  - $W \in \{0,1\}^n$ the treatment. We use the superscript "(w)" to denote the modality $W = w$ of the object.
  - $Y \in \mathbb{R}^n$ the outcome of interest
  - $H \in \{1, \dots, K\}$ the variable of study-membership.
  - $Z_k = \{X_{k,i}, W_{k,i}, Y_{k,i}\}_{i \in [n_k]}$ the data from center k with n_k observations, only accessible by study k.
  - $Z = \bigcup_{k=1}^K Z_k$ the pooled dataset.

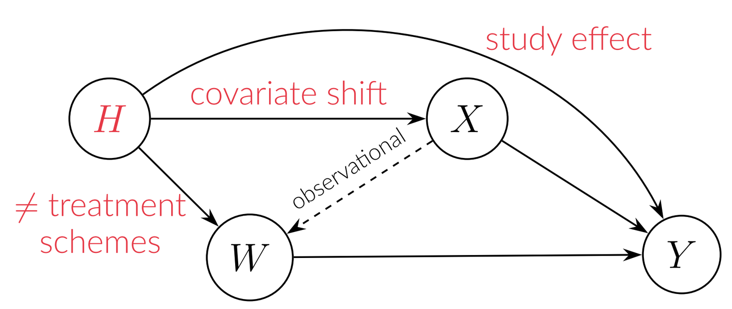


Figure 1. Graphical Model in Decentralized Studies.

Decentralization challenge: handling heterogeneity across studies.

- **ATE identification**:
  - Assuming (a) local unconfoundedness and (b) consistency of treatment and (c) local positivity, the local ATE in study k is identifiable as $\tau_k = \mathbb{E}\left[\mathbb{E}\left[Y_i|W_i = 1, X_i, H_i = k\right] - \mathbb{E}\left[Y_i|W_i = 0, X_i, H_i = k\right]\right]$.
  - The global ATE over all studies is

$$\tau = \mathbb{E}\left[\mathbb{E}\left[Y_i(1) - Y_i(0) \mid H_i\right]\right] = \sum_{k=1}^K \rho_k \tau_k$$

  where $\rho_k = \mathbb{P}(H_i = k)$.
  - Oracle estimators in centralized settings:

G-Formula: $$\tau(\mu_1, \mu_0, Z) = \frac{1}{n}\sum_{i=1}^n (\mu_1(X_i) - \mu_0(X_i))$$

IPW: $$\tau(e, Z) = \frac{1}{n}\sum_{i=1}^n \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)$$

AIPW: $$\tau(\mu_1, \mu_0, e, Z) = \frac{1}{n}\sum_{i=1}^n \Big(\mu_1(X_i) - \mu_0(X_i) + \frac{W_i}{e(X_i)}(Y_i - \mu_1(X_i)) - \frac{1-W_i}{1-e(X_i)}(Y_i - \mu_0(X_i))\Big)$$

  With $\mu_w(X) = \mathbb{E}[Y \mid W = w, X]$ and $e(X) = \mathbb{P}(W = 1 \mid X)$ is the propensity score (PS).

  - **(Fixed-effect) Meta-Analysis Estimators**:

$$\hat{\tau}_{\text{meta}} = \sum_{k=1}^K \omega_k \hat{\tau}_k$$

  with weights $\omega_k = \frac{n_k}{n}$ sample size weighting (SW) or $\frac{\hat{\mathbb{V}}(\hat{\tau}_k)^{-1}}{\sum_{l=1}^K \hat{\mathbb{V}}(\hat{\tau}_l)^{-1}}$: inverse variance weighting (IVW).

$\rightarrow$ We leverage Federated Learning to learn more accurately the ATE.

RCT vs. Observational settings

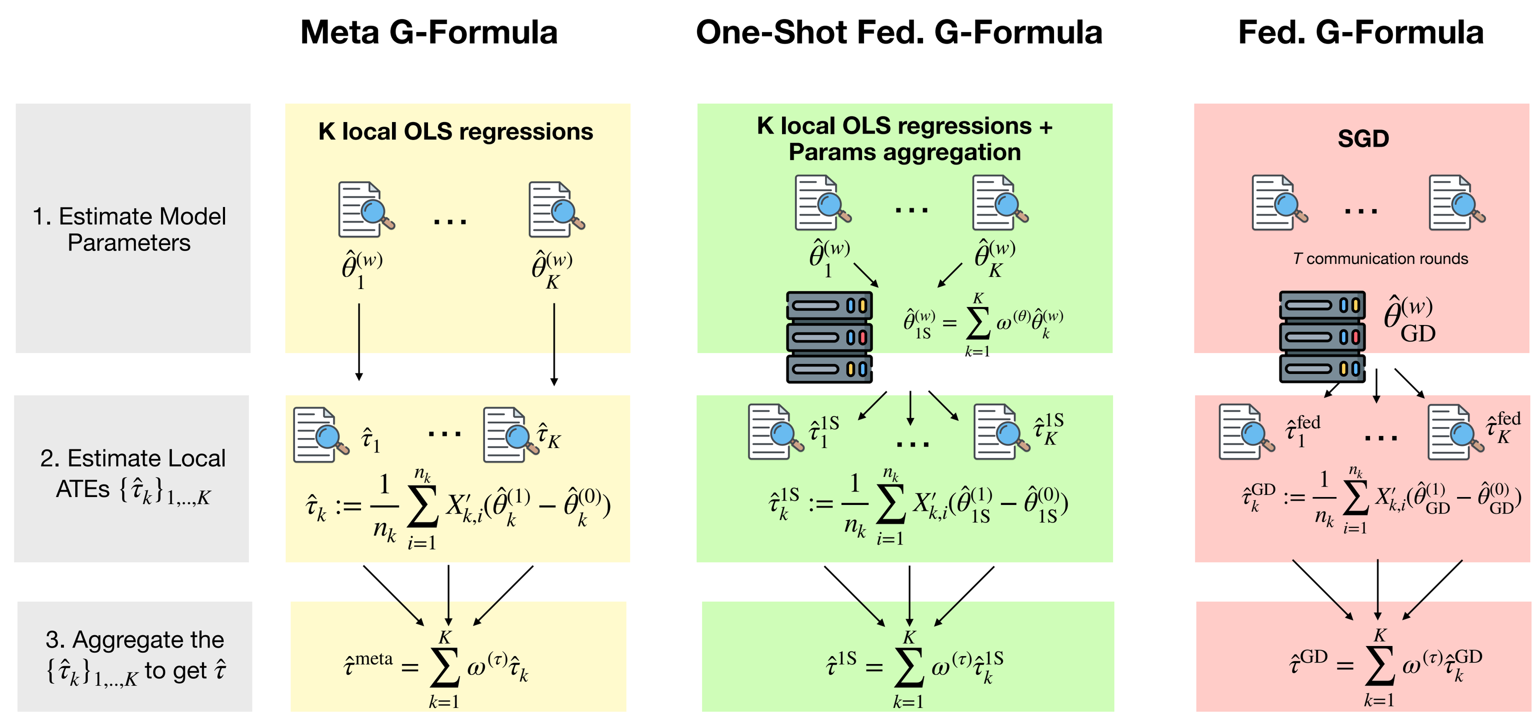
RCT: $W_i \mid H_i = k \sim \mathcal{B}(p_k) \rightarrow$ Linearly-adjusted G-Formulas are unbiased + variance reduction compared to Difference-in-Means, even if linearity is misspecified.

Observational: $W_i \mid H_i = k \sim \mathcal{B}\left(e_k(X_i)\right)$

Federated linearly-adjusted G-formula with multiple RCTs

Modelling: with $X' = (1, X)$, $\theta^{(w)} = (c^{(w)}, \beta^{(w)}) \in \mathbb{R}^{d+1}$, individual i's outcome in study k under treatment arm w is:

$$Y_{k,i}(w) = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_i(w) = X'_{k,i}\theta^{(w)} + \varepsilon_i(w)$$
$$\rightarrow \tau = c^{(1)} - c^{(0)} + \mathbb{E}(X)(\beta^{(1)} - \beta^{(0)}) = \mathbb{E}(X')(\theta^{(1)} - \theta^{(0)})$$



Properties of the estimators

Condition 1. Locally Sufficient Sample Size: $\forall (k, w), n_k^{(w)} \geq d)$
Condition 2. Globally Sufficient Sample Size: $\forall w, \sum_{k=1}^K n_k^{(w)} \geq d$

$\rightarrow$ Under Condition 2, only $\hat{\tau}_{\text{GD}}$ can be computed.

Homogeneous settings ($X|H \sim X \sim \mathcal{D}$):

- Meta and Federated estimators are **unbiased**.

Federated estimators offer a gain in asymptotic variances:

$$\mathbb{V}^\infty(\hat{\tau}_{\text{pool}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{1S}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{GD}}) = \frac{\sigma^2}{n}\frac{1}{p(1-p)} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$$
$$\leq \mathbb{V}^\infty(\hat{\tau}_{\text{meta-IVW}}) = \left(\sum_{k=1}^K \left(\frac{\sigma^2 n \rho_k}{p_k(1-p_k)} + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2\right)^{-1}\right)^{-1}$$
$$\leq \mathbb{V}^\infty(\hat{\tau}_{\text{meta-SW}}) = \frac{\sigma^2}{n}\sum_{k=1}^K \frac{\rho_k}{p_k(1-p_k)} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2$$

Strict if p_k differs across studies.

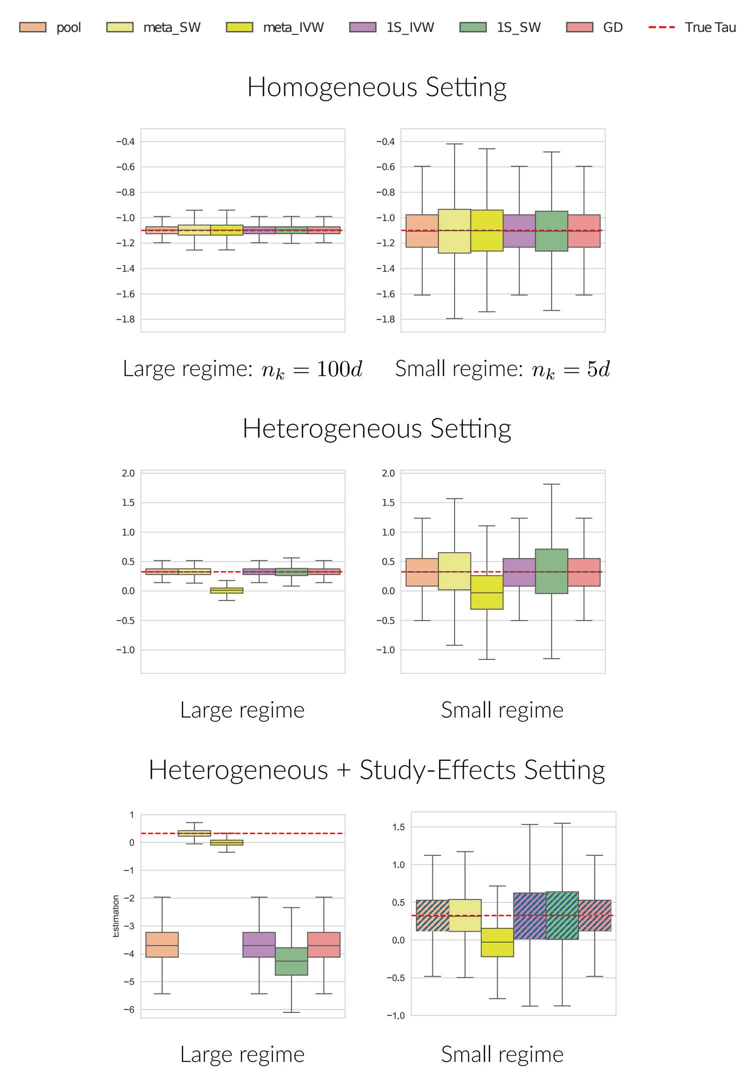
Heterogeneous population settings ($X_i|H_i = k \sim \mathcal{D}_k$):

- Meta-IVW is **biased** (IVW is biased for ρ_k).
- $\mathbb{V}^\infty(\hat{\tau}_{\text{pool}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{1S}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{GD}}) \leq \mathbb{V}^\infty(\hat{\tau}_{\text{meta-SW}})$, strict if different p_k.
- 1S-SW is robust to Σ_k variation, but not $\mathbb{E}[X_i|H_i]$.

Study-effects: $Y_{k,i}(w) = c^{(w)} + \underline{h_k} + X_{k,i}\beta^{(w)} + \varepsilon_i(w)$

$\rightarrow$ H is now a confounder!

- Meta estimators are unbiased (via stratification).
- Federated estimators are biased and need to be adjusted:
  - GD: adjust for H in the descent.
  - 1S: avoid sharing intercepts.



Federated (A)IPW with multiple Observational Studies

Observational studies offer large amounts of data, but likely come with differing treatment assignment practices which mitigates the benefits of learning a global model of the propensity score (PS).

The global PS $e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)$ decomposes into $\sum_{k=1}^K \mathbb{P}(H_i = k \cap W_i = 1 \mid X_i)$ which yields two expressions of the form $e(X_i) = \sum_{k=1}^K \omega_k(X_i) e_k(X_i)$ that make use of the local PS.

Membership Weights (MW)

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

Decentralized estimation of the weights using Gradient Descent-based procedures:

1. Server and studies iterate and find estimates $\hat{\omega}_k(X_i)$ of $\mathbb{P}(H_i = k \mid X_i)$ via FedAvg GD. — Federated Multinomial Logistic regression, Neural networks, Boosted Random Forests
2. Studies estimate locally and share $\hat{e}_k$ to build — $\hat{e}_{\text{MW}}^{\text{fed}}(X_i) = \sum_{k=1}^K \hat{\omega}_k(X_i)\hat{e}_k(X_i)$
3. Studies estimate their local ATE with federated PS. — $\hat{\tau}_k^{\text{MW}} = \tau(\hat{e}_{\text{MW}}^{\text{fed}}, Z_k)$
4. Server aggregates the local ATEs. — $\hat{\tau}_{\text{MW}} = \sum_{k=1}^K \boldsymbol{\omega}_k^{\tau}\hat{\tau}_k^{\text{MW}}$

Use when: X explains H (e.g., geographical repartition of individuals) and small K.

Densities ratio Weights (DW)

$$e(X_i) = \sum_{k=1}^K \underbrace{\rho_k \frac{f_k(X_i)}{f(X_i)}}_{\text{density weights}} e_k(X_i)$$

where the global distribution of X is a mixture of the K local distributions with mixing weights ρ_k, i.e. $f(x) = \sum_{l=1}^K \rho_l f_l(x)$.

Decentralized estimation of the weights with parametric assumptions on $X \mid H$ ($f_k(x) = f_k(x; \theta_k)$):

1. Studies estimate locally and share $\hat{\theta}_k$ (e.g., empirical means, covariance) and $\hat{e}_k$. — $\hat{e}_{\text{DW}}^{\text{fed}}(X_i) = \frac{\sum_{k=1}^K \frac{n_k}{n} f(X_i; \hat{\theta}_k)\hat{e}_k(X_i)}{\sum_{l=1}^K \frac{n_l}{n} f(X_i; \hat{\theta}_l)}$
2. Studies estimate their local ATE with federated PS. — $\hat{\tau}_k^{\text{DW}} = \tau(\hat{e}_{\text{DW}}^{\text{fed}}, Z_k)$
3. Server aggregates the local ATEs. — $\hat{\tau}_{\text{DW}} = \sum_{k=1}^K \boldsymbol{\omega}_k^{\tau}\hat{\tau}_k^{\text{DW}}$

Use when: knowledge on $X \mid H$ and low dimensionality (Kd^2 floats exchanged for covariance matrices)

Properties of the estimators

Assumption 1. Local Overlap: $\forall k, 1/(e_k(x)(1 - e_k(x)) < \infty$
Assumption 2. Global Overlap: $1/(e(x)(1 - e(x)) < \infty$

$\rightarrow$ **Our Federated estimators only require Assumption 2**

Heterogeneous Treatment schemes settings ($X_i|H_i = k \sim \mathcal{D}_k$):

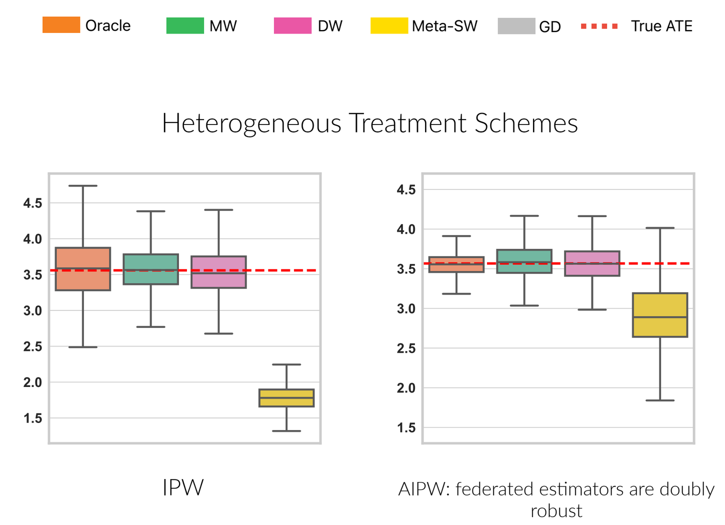
- Meta-IVW are biased if $X \mid H \not\sim X$.
- **IPW and AIPW Meta estimators are biased if Assumption 1 does not hold** unlike the DW and MW ones.
- The global overlap is always "better" than the worst local one:

$$\max_{x \in \text{supp}(X)} (1/(e(x)(1 - e(x))) < \max_{\substack{x \in \text{supp}(X)\\ k \in [K]}} (1/(e_k(x)(1 - e_k(x)))$$

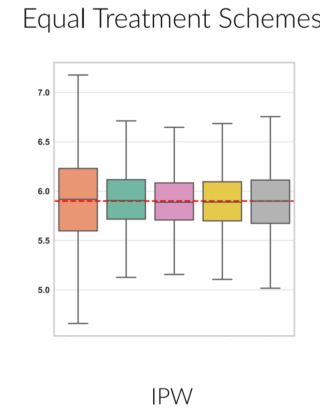
  $\rightarrow$ improved stability.
- Meta estimators have larger variance than the Federated ones.
- Federated IPW estimators with estimated $\hat{e}$ have lower variance than the oracle PS one.

Homogeneous Treatment Schemes settings ($\forall k, e_k(x) = e(x)$):

- Assumption 1 $\iff$ Assumption 2: Meta and Federated estimators with oracle PS are equal. Learning e with GD yields more accurate estimates than locally.



Equal Treatment Schemes



Simulations description:

- Assumption 1 holds left only, up (top and left) only 2 stands.
- Local PS e_k estimated with Generalized Random Forests.
- Non-linear μ_w with interaction terms estimated with misspecified linear-GD (top right).
- $X \mid H = k \sim \mathcal{N}(\mu_k, \Sigma)$.

Conclusion

We provide practical guidance on the advantages of using Federated estimators vs. Meta estimators in both RCTs and Observational studies, accounting for heterogeneity across studies (in populations, treatment schemes and study-effects) which data is decentralized.

Soon: generalization of the ATE and Conditional ATE estimation in decentralized settings.

remi.khellaf@inria.fr