Federated Causal Inferences: Estimating the Average Treatment Effect in a Decentralized Setting

Rémi Khellaf [1] & Aurélien Bellet [2]& Julie Josse [3]

[1] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France, remi.khellaf@inria.fr*
[2] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France aurelien.bellet@inria.fr*
[3] *Inria Premedical Inria-Idesp (Inserm - University of Montpellier) team, France julie.josse@inria.fr*

Summary This article aims at discussing the practical challenges of conducting causal inference analyses with decentralized data. Federated learning brings tools to deal with distributed datasets where privacy are important stakes. Federated causal inference then appears to show particular promise in domains where the data are siloed, while allowing researches to harness the power of large-scale data to answer causal questions. We focus on estimating the Average Treatment Effect (ATE) within a federated framework.

Résumé Cet article vise à discuter des défis pratiques liés à la réalisation d'analyses d'inférence causale avec des données décentralisées. L'apprentissage fédéré apporte des outils pour traiter les ensembles de données distribués où la confidentialité est une préoccupation importante. L'inférence causale fédérée semble alors prometteuse dans les domaines où les données sont cloisonnées, tout en permettant aux chercheurs d'exploiter la puissance des données à grande échelle pour répondre à des questions causales. Nous nous concentrons sur l'estimation de l'Effet Traitement Moyen (ATE) dans un cadre fédéré.

Keywords Federated learning; Causal inference; regression; privacy; linear model; meta-analysis; average treatment effect; G-Formula; linear models; hospital data

Mots-clés Apprentissage fédéré ; Inférence causale ; régression ; modèle linéaire ; méta ; analyse ; effet de traitement moyen ; G ; Formula ; modèles linéaires ; effets de traitement hétérogènes ; données hospitalières ; données réelles ; confidentialité

Abstract. Federated causal inference is a burgeoning field at the intersection of machine learning and causal inference, aimed at harnessing data from multiple distributed sources to infer causal relationships while preserving privacy and security. Its applications are the most promising in fields where there is a substantial gain at keeping the data decentralized for logistic or privacy reasons, which is particularly the case of hospitals holding patients data.

In this paper, we explore the challenges of federated estimation of the Average Treatment Effect (ATE) by defining and studying several suitable estimators derived from the plugin G-Formula estimator. We provide a comprehensive comparison of the variances of these estimators. We also compare the sample size and inverse variance aggregation methods for the meta estimators in linear models.

1 Introduction

1.1 Causal inference in observational studies

Let X denote the p-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^p$, $W \in \mathcal{W} = \{0, 1\}$ denote the binary treatment, and $Y \in \mathbb{R}$ denote the outcome of interest. We consider the potential outcomes framework, where for $w \in \mathcal{W}$, $Y(w)$ is the outcome had the subject received treatment w. We have access to a sample of n independent and identically distributed (i.i.d.) observations $(Y_i, W_i, X_i)_{i=1,\dots,n}$ and our aim is to identify and estimate the average treatment effect (ATE) $\tau\mathbb{E}[Y(1) - Y(0)]$.

Under the standard assumptions of (a) consistency: $Y = WY(1) + (1 - W)Y(0)$, (b) positivity: there exists $\eta > 0$ such that $1 - \eta \geq \mathbb{P}(W = 1 \mid X = x) \geq \eta$ for every $x \in \mathcal{X}$, (c) unconfoundedness: $W \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X$, we can nonparametrically identify $\mathbb{E}[Y(w))]$ for instance by the outcome regression/G-formula or the inverse probability weighting (IPW) formula. More precisely, the regression identifiability formula can be obtained as follows:

$$\begin{aligned}
\tau &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, |X_i = x]] && \text{(unconfoundedness)} \\
&= \mathbb{E}[\mathbb{E}[Y_i|W_i = 1, X_i] - \mathbb{E}[Y_i|W_i = 0, X_i]] && \text{(consistency)}
\end{aligned}$$

This suggests the popular Plug-in G-formula estimator defined as follows.

Definition 1 (Plug-in G-formula [4]). *The plug-in G-formula estimator, denoted by $\hat{\tau}$, is defined as*

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right),$$

where $\hat{\mu}_w(X)$ is an estimator of $\mu_w(X) = \mathbb{E}\left[Y \mid W = w, X\right]$.

In this work, we focus on linear outcome models and tackle the challenges of estimating the ATE in a Federated setting using G-formula-based estimators.

1.2 Federated learning

When data is collected and stored in distinct centers, it is sometimes impossible to pool them in one place due to privacy or organizational reasons. It is particularly the case when data is sensitive and cannot be shared. For example, hospitals often cannot share their data easily due to medical confidentiality, and often seek to retain control and ownership of their data. On the other hand, gathering enough data to conduct statistically significant observational studies at a single hospital can be challenging. Federated learning [3] aims at addressing these issues by allowing analysts to train a machine learning model across several centers while keeping data decentralized. Federated learning algorithms typically rely on an orchestrating server and operate across multiple rounds. At each round, (i) the server transmits the current global model to the centers, (ii) each center updates the model using its respective data and sends this model update (not raw data) to the server, and (iii) the server aggregates the updates to form a new global model.

We consider K centers and denote by $Z = \{Z_1, \ldots, Z_K\}$ the pooled dataset, where $Z_k = \{X_k, W_k, Y_k\}$ the data from center $k \in [\![1, K]\!]$ with n_k observations, and $n = \sum_{k=1}^K n_k$. The dataset Z_k is only accessible by center k. Federated Learning aims at recovering the same results had the researcher had direct access to dataset Z.

In the following, we will consider either of the following two assumptions on the sample sizes.

Assumption 1 (Local Large Sample Size)**.** *For each $w \in \{0, 1\}$ and for each center k, we assume $n_k^{(w)} > p$, where $n_k^{(w)}$ is the number of individuals in $\{Z_k \mid W = w\}$ (i.e., the number of treated/control individuals in each center) and p the number of features in X.*

Assumption 2 (Federated Large Sample Size)**.** *For each $w \in \{0, 1\}$, we assume $\sum_{k=1}^K n_k^{(w)} > p$.*

In the sequel and unless stated otherwise, we assume that Assumption 1 holds.

1.3 Related work

Federated learning has mostly focused on training predictive models [3], and only a handful of recent work considers the problem of federated causal inference. The field of federated causal inference is recent and only few works are available. One can cite [5] who studied the estimation of the ATE in the simplest one-shot federated learning framework (a single round of communication between the server and the centers). In this work, we introduce additional federated estimators and systematically compare the bias and variance of the different strategies under linear models.

2 Federated causal inference

2.1 Model definition

We consider a random design model, where individuals follow a similar distribution among the K centers. In other words, each center can be seen as a random sample drawn from the same population. We consider the following potential outcomes model with an homogeneous binary treatment effect $\tau \in \mathbb{R}$ with $w \in \{0, 1\}$:

$$Y_i^k(w) = c^{(w)} + X_i^k \beta^{(w)} + \varepsilon_i(w)$$

We denote :

- $Y^k(w) := \{Y_i^k(w)\}_{i=1}^{n_k}$ and $X^k := \{X_i^k\}_{i=1}^{n_k}$ the outcome and covariate vectors for center k.
- $Y(w) := \{Y_i(w)\}_{i=1}^n = \cup_{k=1}^K Y^k(w)$ and $X := \{X_i\}_{i=1}^n = \cup_{k=1}^K X^k$ the outcome and covariate vectors of the pooled dataset.

and assume:

- (same covariates distribution) $\forall (i,k), X_i^k \sim \mathcal{D}$
- (centered covariates) $\forall k, \mathbb{E}(X) = \mathbb{E}(X^k) = 0$ and $\mathbb{V}(X) = \mathbb{V}(X^k) = \Sigma$
- (regression assumptions) $\forall k, \mathbb{E}(X^{k\top}\varepsilon) = 0, \quad \mathbb{V}(\varepsilon \mid X^k) = \sigma^2, \quad \text{rank}(X^{k\top}X^k) = d$
- (RCT) The treatment variable W is random and does not depend on the covariates: $W \sim \mathcal{B}(p)$.

Finally we define $\tau := \mathbb{E}(Y(1) - Y(0)) = c^{(1)} - c^{(0)} + \mathbb{E}(X)(\beta^{(1)} - \beta^{(0)})$ and $\tau_k := \mathbb{E}(Y^k(1) - Y^k(0)) = c^{(1)} - c^{(0)} + \mathbb{E}(X^k)(\beta^{(1)} - \beta^{(0)})$.

Remark that in this model, we have $\tau_1 = \tau_2 = \cdots = \tau_K = \tau$, which means that the ATE is homogeneous across centers.

Estimating the treatment effect on the pooled data can be done by fitting two ordinary least squares (OLS) regressions, one for the treated and one for the control group. For simplicity, we now - and unless stated otherwise - consider that the first column of X contains the constant vector to serve as intercept, so that The Plug-in G-formula estimator is:

$$\hat{\tau}_{\text{OLS}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{c}^{(1)} + X_i\hat{\beta}_{(1)} - \hat{c}^{(0)} + X_i\hat{\beta}_{(0)}\right)$$

We get the following asymptotic result using the central limit theorem:

$$\sqrt{n}(\hat{\tau}_{\text{OLS}} - \tau) \xrightarrow{d} \mathcal{N}(0, n\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + n\|\beta^{(1)} - \beta^{(0)}\|_{\Sigma}^2) \tag{1}$$

However, the $\hat{\tau}_{\text{OLS}}$ estimator of the ATE cannot be computed in the federated setting since data is decentralized across K centers. Therefore, alternative strategies need to be considered.

2.2 Estimation with meta-analysis

A first strategy is to estimate the ATE locally in each center, yielding for each center $k \in [\![1, \ldots, K]\!]$ its associated $\hat{\tau}_k$, and then to aggregate them with weights ω_k (summing to 1 over the K centers) to get an estimation of the ATE, $\hat{\tau}_{\text{meta}}$, for the whole data, *i.e.*

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

We call this approach Meta Analysis in reference to the scientific literature that aims at combining the conclusions of multiple studies in order to get a more precise estimate of the quantities of interest. Regarding the aggregation weights, [2] presented two natural choices as we focus on the risk difference: sample size weighting (SW) or weighting the local estimates by the inverse of their variance (IVW).

Definition 2 (Meta-Analysis with Sample Size Weighting)**.**

$$\hat{\tau}_{\text{meta-SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k$$

Property. *If* $\forall k \in [\![1; K]\!], \mathbb{E}(\hat{\tau}_k) = \tau$*, then* $\hat{\tau}_{\text{meta-SW}}$ *is an unbiased estimator of* τ*.*

Definition 3 (Meta-Analysis with Inverse Variance Weighting)**.**

$$\hat{\tau}_{\text{meta-IVW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k)^{-1} . \hat{\tau}_k}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k)^{-1}}$$

Property. *If* $\forall k \in [\![1; K]\!], \mathbb{E}(\hat{\tau}_k) = \tau$ *then* $\hat{\tau}_{\text{meta-IVW}}$ *is the minimum-variance unbiased estimator of* τ *based on aggregation.*

2.3 Estimation with one-shot federated learning

A second approach consists in estimating the outcome models parameters in a federated way, using these parameters to estimate the ATE on each center, and ultimately to aggregate these ATE estimates over all centers. In Section 2.3.1, we present an approach to federate the $\hat{\beta}_k$ using a single round of communication ("one-shot" federated learning), following [5]. Then, in Section 2.3.2, we explain how to recover the ATE over all the centers, defining and discussing four estimators of the ATE that rely on the federated parameters.

We define $\hat{\beta}_k^{(w)} := ({X_k^{(w)}}^\top X_k^{(w)})^{-1} {X_k^{(w)}}^\top Y_k^{(w)}$ the local OLS estimator, computed over Z_k and where $X_k^{(w)} = \{X_i^k | W_i^k = w\}_{i=1,\ldots,n}$, $Y_k^{(w)} = \{Y_i^k | W_i^k = w\}_{i=1,\ldots,n}$ and $w = \{0, 1\}$.

2.3.1 One-shot federation of parameters

The first step, which we call the one-shot federation step, is the equivalent of a meta-analysis (local estimation then weighted aggregation) over the $\hat{\beta}_k^{(w)}$. One computes $\hat{\beta}_{\text{fed}}^{(w)} = \sum_{k=1}^{K} \omega_k^{(\beta)} \hat{\beta}_k^{(w)}$ in order to approach the $\hat{\beta}_{\text{pool}}^{(w)}$, with $\omega_k^{(\beta)}$ some chosen federation weights (summing to 1 over the K centers) like sample size weighting and inverse variance weighting.

Definition 4 (Sample Size Weighting Federation of outcome model parameters)**.**

$$\hat{\beta}_{\text{SW}}^{(w)} = \sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \hat{\beta}_k^{(w)}$$

Property. *If* $\forall k \in [\![1;K]\!], \mathbb{E}(\hat{\beta}_k^{(w)}) = \beta^{(w)}$, *then* $\hat{\beta}_{\text{SW}}^{(w)}$ *is an unbiased estimate of* $\beta^{(w)}$

Definition 5 (Inverse Variance Weighting Federation of outcome model parameters)**.**

$$\hat{\beta}_{\text{IVW}}^{(w)} = \frac{\sum_{k=1}^{K} \left(\mathbb{V}(\hat{\beta}_k^{(w)})^{-1} \hat{\beta}_k^{(w)} \right)}{\sum_{k=1}^{K} \mathbb{V}(\hat{\beta}_k^{(w)})^{-1}}$$

with $\mathbb{V}(\hat{\beta}_k^{(w)})^{-1} = \frac{1}{\sigma^2} {X_k^{(w)}}^{\top} X_k^{(w)}$

Property (IVW: Minimum-Variance Estimator)**.** *If* $\forall k \in [\![1;K]\!], \mathbb{E}(\hat{\beta}_k^{(w)}) = \beta^{(w)}$, *then* $\hat{\beta}_{\text{IVW}}^{(w)}$ *is an unbiased estimator of* $\beta^{(w)}$ *with minimum-variance.*

2.3.2 Aggregation of the ATE

Once the outcome model parameters have been computed via federation, one can apply the resulting $\hat{\beta}_{\text{fed}}^{(w)} = \{\hat{\beta}_{\text{SW}}^{(w)}$ or $\hat{\beta}_{\text{IVW}}^{(w)}\}$ to each local dataset Z_k to recover $\hat{\tau}_k^{\text{fed}} = \{\hat{\tau}_k^{\text{SW}}$ or $\hat{\tau}_k^{\text{IVW}}\}$, the estimated ATE over center k: $\hat{\tau}_k^{\text{fed}} = \frac{1}{n_k} \sum_{i=1}^{n_k} (X_i^k \hat{\beta}_{\text{fed}}^{(1)} - X_i^k \hat{\beta}_{\text{fed}}^{(0)})$.

Finally, we perform an aggregation step of the $\hat{\tau}_k^{\text{fed}}$ to recover $\hat{\tau}_{\text{agg}}^{\text{fed}}$, with a chosen $\omega^{(\tau)}$ aggregation. This yields:

$$\hat{\tau}_{\text{agg}}^{\text{fed}} = \sum_{k=1}^{K} \omega_k^{(\tau)} \hat{\tau}_k^{\text{fed}}$$

Note that we use the superscript "fed" for a federation technique of the outcome model parameters, and the subscript "agg" for the aggregation of the local ATE $\hat{\tau}_k^{\text{fed}}$. We define four estimators of the ATE over the K centers.

Definition 6 (One-Shot Sample Size Federation - Sample Size Aggregation)**.**

$$\hat{\tau}_{\text{SW}}^{\text{SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{SW}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_i^k \hat{\beta}_{\text{SW}}^{(1)} - X_i^k \hat{\beta}_{\text{SW}}^{(0)})$$

Definition 7 (One-Shot Inverse Variance Weighting Federation - Sample Size Aggregation)**.**

$$\hat{\tau}_{\text{SW}}^{\text{IVW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{IVW}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_i^k \hat{\beta}_{\text{IVW}}^{(1)} - X_i^k \hat{\beta}_{\text{IVW}}^{(0)})$$

Definition 8 (One-Shot Sample Size Federation - Inverse Variance Aggregation)**.**

$$\hat{\tau}_{\text{IVW}}^{\text{SW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{SW}})^{-1} \cdot \hat{\tau}_k^{\text{SW}}}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{SW}})^{-1}}$$

Algorithm 1 Federated Averaging

Input: K centers, E local epochs, B batch size, η the learning rate
Server executes:
  Initialize $\beta_0^{(w)}$
for each round $t = 0, 1, \dots$ **do**
  for each center $k \in [\![1, K]\!]$ **in parallel do**
    $\beta_{t+1}^{k^{(w)}} \leftarrow \text{LocalUpdate}(k, \beta_t^{(w)})$
  end for
  $\beta_{t+1}^{(w)} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \beta_{t+1}^{k^{(w)}}$
end for
LocalUpdate$(k,\ \beta_k^{(w)})$:
for each local epoch $e = 0, 1, \dots, E-1$ **do**
  $\nabla \ell(\beta_k^{(w)}; \mathcal{D}_k) \leftarrow -\frac{2}{n_k} X^{k^\top} (Y^k - X^k \beta_k^{(w)})$
  $\beta_k^{(w)} \leftarrow \beta_k^{(w)} - \eta \nabla \ell(\beta_k^{(w)}; Z_k)$
end for
return $\beta_k^{(w)}$

Definition 9 (One-Shot Inverse Variance Weighting Federation - Inverse Variance Aggregation)**.**

$$\hat{\tau}_{\text{IVW}}^{\text{IVW}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{IVW}})^{-1} \cdot \hat{\tau}_k^{\text{IVW}}}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{IVW}})^{-1}}$$

Property. $\hat{\tau}_{\text{IVW}}^{\text{IVW}}$ *is unbiased*

Property (Minimum Variance of IVW)**.**

2.4 Estimation with Federated Learning

A last strategy consists in estimating the outcome model parameters over the pooled data using the FedAvg algorithm [1], which requires multiple communication rounds. The idea is to estimate $\hat{\beta}_{\text{pool}}^{(w)}$ by solving the underlying OLS problem by gradient descent, which corresponds to minimizing the Mean Squared Error loss function $\ell(\beta^{(w)}; Z_1^{(w)}, \dots, Z_K^{(w)}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta)^2$. The algorithm, shown in Algorithm 1, alternates between local gradient steps in each center and aggregation at the server, to eventually produce an estimate $\hat{\beta}_{\text{GD}}^{(w)}$ ("GD" stands for Gradient Descent).

Under an appropriate choice of hyper-parameters, the FedAvg algorithm converges [3] and we have $\hat{\beta}_{\text{GD}}^{(w)} = \hat{\beta}_{\text{pool}}^{(w)}$. Once $\hat{\beta}_{\text{GD}}^{(w)}$ is obtained, we compute $\hat{\tau}_k^{\text{GD}} := \frac{1}{n_k} \sum_{i=1}^{n_k} \left(X_i^k \hat{\beta}_{\text{GD}}^{(1)} - X_i^k \hat{\beta}_{\text{GD}}^{(0)} \right)$, and then calculate the estimates of the ATE over the K centers $\hat{\tau}_{\text{agg}}^{\text{GD}} = \sum_{k=1}^{K} \omega_k^{(\tau)} \hat{\tau}_k^{\text{GD}}$ with a chosen aggregation method as in Section 2.3.

Definition 10 (Gradient Descent Federation - Sample Size Aggregation)**.**

$$\hat{\tau}_{\text{SW}}^{\text{GD}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{GD}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(X_i^k \hat{\beta}_{\text{GD}}^{(1)} - X_i^k \hat{\beta}_{\text{GD}}^{(0)} \right)$$

Definition 11 (Gradient Descent Federation - Inverse Variance Aggregation)**.**

$$\hat{\tau}_{\text{IVW}}^{\text{GD}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{GD}})^{-1} \hat{\tau}_k^{\text{GD}}}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{GD}})^{-1}} = \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{GD}})^{-1} \frac{1}{n_k} \sum_{i=1}^{n_k} \left(X_i^k \hat{\beta}_{\text{GD}}^{(1)} - X_i^k \hat{\beta}_{\text{GD}}^{(0)} \right)}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k^{\text{GD}})^{-1}}$$

Property. *If* $\mathbb{E}(\hat{\beta}_{\text{GD}}^{(w)}) = \beta^{(w)}$, $\hat{\tau}_{\text{SW}}^{\text{GD}}$ *and* $\hat{\tau}_{\text{SW}}^{\text{GD}}$ *are unbiased.*

2.5 Properties, Bias and Variance of the Federated Estimators

Under the random design, one can derive asymptotic properties of the estimators, summarized in this table.

Estimator	Notation	Definition	$\mathbb{E}$	$\mathbb{V}^{\infty}$
Pool	$\hat{\beta}_{\text{pool}}^{(w)}$	$\left({X^{(w)}}^{\top} X^{(w)} \right)^{-1} {X^{(w)}}^{\top} y^{(w)}$	$\beta^{(w)}$	$\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$
Local	$\hat{\beta}_k^{(w)}$	$\left({X_k^{(w)}}^{\top} X_k^{(w)} \right)^{-1} {X_k^{(w)}}^{\top} y_k^{(w)}$	$\beta^{(w)}$	$\frac{\sigma^2}{n_k^{(w)}} \Sigma^{-1}$
Federated	$\hat{\beta}_{\text{GD}}^{(w)}$	obtained by gradient descent	$\beta^{(w)}$	$\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$
Sample Size Weighted	$\hat{\beta}_{\text{SS}}^{(w)}$	$\sum_{k=1}^{K} \frac{n_k}{n} \hat{\beta}_k^{(w)}$	$\beta^{(w)}$	$\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$
Inverse Variance Weighted	$\hat{\beta}_{\text{IVW}}^{(w)}$	$\frac{\sum_{k=1}^{K} \left(\mathbb{V}^{\infty}(\hat{\beta}_k^{(w)})^{-1} \hat{\beta}_k^{(w)} \right)}{\sum_{k=1}^{K} \mathbb{V}^{\infty}(\hat{\beta}_k^{(w)})^{-1}}$	$\beta^{(w)}$	$\frac{\sigma^2}{n^{(w)}} \Sigma^{-1}$

Table 1: Estimators of the outcome model parameters

Property 1. $\forall k \in [\![1; K]\!]$, $\frac{n_k}{n} = \frac{\mathbb{V}^{\infty}(\hat{\tau}_k \text{fed})^{-1}}{\sum_{k=1}^{K} \mathbb{V}^{\infty}(\hat{\tau}_k \text{fed})^{-1}}$, *which yields* $\hat{\tau}_{\text{SW}}^{\text{fed}} = \hat{\tau}_{\text{IVW}}^{\text{fed}}$

Then, using that $\sqrt{n}(\hat{\tau}_{\text{OLS}} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \left(\frac{1}{n^1} + \frac{1}{n^0}\right) + \|\beta^{(1)} - \beta^{(0)}\|_{\Sigma}^2\right)$, we get the following asymptotic results for the ATE estimators:

Estimator	Notation	Definition	$\mathbb{E}$	$\mathbb{V}^\infty$
Pool	$\hat{\tau}_{\text{pool}}$	$\frac{1}{n}\sum_{i=1}^n X_i\left(\hat{\beta}^{(1)^\top}_{\text{pool}} - \hat{\beta}^{(0)^\top}_{\text{pool}}\right)$	τ	$\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
Local	$\hat{\tau}_k$	$\frac{1}{n}\sum_{i=1}^n X_i\left(\hat{\beta}^{(1)^\top}_{k} - \hat{\beta}^{(0)^\top}_{k}\right)$	τ	$\sigma^2\left(\frac{1}{n_k^{(1)}} + \frac{1}{n_k^{(0)}}\right) + \frac{1}{n_k}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
fed-Local	$\hat{\tau}_k^{\text{fed}}$	$\frac{1}{n}\sum_{i=1}^n X_i\left(\hat{\beta}^{(1)^\top}_{\text{fed}} - \hat{\beta}^{(0)^\top}_{\text{fed}}\right)$	τ	$\frac{\sigma^2}{n_k}\left(\frac{n}{n^{(1)}} + \frac{n}{n^{(0)}}\right) + \frac{1}{n_k}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
meta-SW	$\hat{\tau}_{\text{SW}}$	$\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k$	τ	$\sum_{k=1}^K\left(\frac{n_k}{n^2}\left(\left(\frac{\sigma^2 n_k}{n_k^{(1)}} + \frac{\sigma^2 n_k}{n_k^{(0)}}\right) + \lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2\right)\right)$
meta-IVW	$\hat{\tau}_{\text{IVW}}$	$\frac{\sum_{k=1}^K\left(\mathbb{V}^\infty(\hat{\tau}_k)^{-1}\hat{\tau}_k\right)}{\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k)^{-1}}$	τ	$\left(\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k)^{-1}\right)^{-1}$
fed-SW-agg	$\hat{\tau}_{\text{SW}}^{\text{fed}}$	$\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k^{\text{fed}}$	τ	$\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
fed-IVW-agg	$\hat{\tau}_{\text{IVW}}^{\text{fed}}$	$\frac{\sum_{k=1}^K\left(\mathbb{V}^\infty(\hat{\tau}_k^{\text{fed}})^{-1}\hat{\tau}_k^{\text{fed}}\right)}{\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k^{\text{fed}})^{-1}}$	τ	$\sigma^2\left(\frac{1}{n^{(1)}} + \frac{1}{n^{(0)}}\right) + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$

Table 2: Estimators of the ATE in random design, RCT

One can rewrite the table above for the special RCT case of a completely balanced design, where $n_k^{(1)} = n_k^{(0)} = n_k/2$ for all $k \in [\![1, K]\!]$:

Estimator	Notation	Definition	$\mathbb{E}$	$\mathbb{V}^\infty$
Pool	$\hat{\tau}_{\text{pool}}$	$\frac{1}{n}\sum_{i=1}^n\left(X_i\hat{\beta}_1^{\text{pool}\top} - X_i\hat{\beta}_{(0)}^{\text{pool}\top}\right)$	τ	$\frac{4\sigma^2}{n} + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
Local	$\hat{\tau}_k$	$\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^k\hat{\beta}_k^{(1)\top} - X_i^k\hat{\beta}_k^{(0)\top}\right)$	τ	$\frac{4\sigma^2}{n_k} + \frac{1}{n_k}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
fed-Local	$\hat{\tau}_k^{\text{fed}}$	$\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X_i^k\hat{\beta}_1^{\text{fed}\top} - X_i^k\hat{\beta}_{(0)}^{\text{fed}\top}\right)$	τ	$\frac{4\sigma^2}{n_k} + \frac{1}{n_k}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
meta-SW	$\hat{\tau}_{\text{SW}}$	$\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k$	τ	$\frac{4\sigma^2}{n} + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
meta-IVW	$\hat{\tau}_{\text{IVW}}$	$\frac{\sum_{k=1}^K\left(\mathbb{V}(\hat{\tau}_k)^{-1}\hat{\tau}_k\right)}{\sum_{k=1}^K \mathbb{V}(\hat{\tau}_k)^{-1}}$	τ	$\frac{4\sigma^2}{n} + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
fed-SW-agg	$\hat{\tau}_{\text{SW}}^{\text{fed}}$	$\sum_{k=1}^K \frac{n_k}{n}\hat{\tau}_k^{\text{fed}}$	τ	$\frac{4\sigma^2}{n} + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$
fed-IVW-agg	$\hat{\tau}_{\text{IVW}}^{\text{fed}}$	$\frac{\sum_{k=1}^K\left(\mathbb{V}(\hat{\tau}_k^{\text{fed}})^{-1}\hat{\tau}_k^{\text{fed}}\right)}{\sum_{k=1}^K \mathbb{V}(\hat{\tau}_k^{\text{fed}})^{-1}}$	τ	$\frac{4\sigma^2}{n} + \frac{1}{n}\lVert\beta^{(1)} - \beta^{(0)}\rVert_\Sigma^2$

Table 3: Estimators of the ATE in fixed design

3 Conclusion

In this paper, we have defined and studied several estimators for the ATE in the context of federated causal inference. We have provided a comprehensive comparison of the variance of these estimators, and we have shown that in the simple case of homogeneous settings with no centre-effect, the meta-analysis estimators are asympotically as efficient as the federated estimators. Simulation results that are not presented in this paper also back up this finding.

The natural next steps of this work would be to explore the context of heterogeneous settings, where populations or treatment assignment rules are different from centers to centers. Heterogeneous treatment effects as well as nonlinear models are interesting horizons, and other estimators of the ATE like the Augmented Inverse Propensity Score (AIPW) can also be studied. Covariate mismatch between centers and high dimensionality are important topics to consider, especially in the context of hospital data. Finally, the development of federated causal inference methods that can be applied to real-world data is an important direction for future research.

References

[1] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[2] John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

[3] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, and Mehdi Bennis et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[4] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[5] Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*, 2021.