Federated Causal Inference: Multi-Study ATE Estimation beyond Meta-Analysis

Rémi Khellaf **Aurélien Bellet** **Julie Josse**
Inria, PreMeDICaL Team, Université de Montpellier

Abstract

We study Federated Causal Inference, an approach to estimate treatment effects from decentralized data. We compare three classes of Average Treatment Effect (ATE) estimators derived from the Plug-in G-Formula, ranging from simple meta-analysis to one-shot and multi-shot federated learning, the latter leveraging the full data to learn the outcome model (albeit requiring more communication). Focusing on Randomized Controlled Trial studies (RCTs), we derive the asymptotic variance of these estimators for linear models. Our results provide practical guidance on selecting the appropriate estimator for various scenarios, including heterogeneity in sample sizes, covariate distributions, treatment assignment schemes, and study-effects. We validate these findings through experiments on simulated and semi-synthetic data.

1 INTRODUCTION

In modern evidence-based medicine, Randomized Controlled Trials (RCT) are considered the gold standard for estimating the Average Treatment Effect (ATE) because they effectively isolate the treatment effect from confounding factors (Guyatt et al., 2015). The most widely used estimator of the ATE, when expressed as a risk difference, is the difference-in-means (DM) estimator. Recently, however, the U.S. Food and Drug Administration (2023) has recommended to adjust for covariates using linear models for the outcome, as this approach consistently yields more precise ATE estimates than the DM estimator (European Medicines Agency, 2024; Tsiatis et al., 2008; Benkeser et al., 2021) even when the assumption of linearity does not hold (Lin, 2013; Wager, 2020; Lei and Ding, 2021; Van Lancker et al., 2024).

Nevertheless, concerns have been raised about the limited scope of RCTs, including their stringent eligibility criteria, short timeframes, limited sample size, etc. Consequently, regulatory agencies tasked with making high-stakes decisions on drug approvals—decisions that directly impact public health and for which the reimbursement of the drug is often tied to its efficacy (French Health Authority, 2024)—frequently turn to meta-analysis to guide their choices. Meta-analysis, which aggregates estimated effects from multiple studies conducted across various studies (Hunter and Schmidt, 2004; Borenstein et al., 2021), represents the pinnacle of evidence in clinical research (Blunt, 2015). They can lead to increased statistical power and more precise estimates, while also offering valuable insights into rare adverse events.

Despite extensive guidelines on conducting meta-analyses (Moher et al., 1999; Liberati et al., 2009; Higgins et al., 2019), multi-study approaches still face significant challenges. These primarily arise from heterogeneity caused by imbalances in datasets, variations in populations across studies, and study-effects on the outcome due to differing practices across studies (Berlin and Golub, 2014). Moreover, simply aggregating local estimates is not the only approach to conducting meta-analyses. However, implementing "one-stage" meta-analyses (Morris et al., 2018) that pool individual patient data from all studies is practically challenging due to data silos and personal data regulations.

Federated causal inference, an emerging field combining federated learning (Kairouz et al., 2021) and causal inference (Imbens and Rubin, 2015; Hernan, 2020) to estimate causal effects from decentralized data sources, offers a compelling alternative to traditional meta-analysis. Federated Learning (FL) enables multiple studies to collaboratively train a model without sharing raw individual data, instead exchanging only model updates that are iteratively aggregated by an

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

orchestrating server. This decentralized approach is especially valuable in fields like medicine (Sheller et al., 2020; Che et al., 2022; Prosperi et al., 2020), where strong incentives exist to keep data on-site—whether to comply with data protection regulations (Koga et al., 2024), maintain ownership and control over the data, or avoid unwanted knowledge transfer. However, traditional FL algorithms are designed to learn predictive models (Kairouz et al., 2021), rather than estimating causal effects.

Contributions. In this paper, we aim to estimate the ATE for a population represented by multiple RCT studies conducted over potentially heterogeneous populations, using a federated approach. We study and compare three classes of federated estimators of the ATE: (i) *meta-analysis estimators*, which aggregate ATE estimates computed independently at each study; (ii) *one-shot federated estimators*, where outcome model parameters are estimated at each study, aggregated, and shared back for studies to compute and aggregate ATE estimates; and (iii) *gradient-based federated estimators*, where outcome model parameters are learned on joint data using federated gradient descent before each study computes and aggregates its ATE estimate. These estimators entail different communication costs, which often act as the bottleneck in real-world systems (Kairouz et al., 2021).

Our primary contribution is the derivation of asymptotic variances for these estimators under a linear outcome model. This modeling choice is known to provide a variance reduction compared to the classic DM estimator, even when the underlying model is not linear (Lin, 2013; Wager, 2020; Lei and Ding, 2021; Van Lancker et al., 2024), and aligns with recent recommendations from regulatory agencies (U.S. Food and Drug Administration, 2023). We specifically address scenarios involving heterogeneity, including distributional shifts (varying covariate distributions across studies) and study-effects on the outcome. Our results shed light on the trade-offs between the statistical efficiency, communication costs, and underlying modeling assumptions of the considered estimators. We find that, despite their simplicity, low communication overhead and minimal assumptions, meta-analysis estimators can achieve statistical efficiency comparable to pooled data analysis when sufficient data is available at each study, while naturally accommodating study-effects. In contrast, when local datasets are small, gradient-based federated estimators stand out as the only viable option. One-shot estimators offer an interesting middle ground in some cases: they can recover the same ATE estimate as pooled data analysis while being robust to distributional shifts in covariates and differences in treatment assignments, but suffer from increased variance when study-effects are present. These conclusions are supported by experiments on (i) simulated data, illustrating the behavior of the estimators under the different scenarios; and (ii) semi-synthetic data, where we use the real-world Traumatrix database (Mayer et al., 2020) with synthetic outcomes. Ultimately, our work provides clear guidelines on selecting the most suitable estimator for different scenarios, as summarized by a decision diagram designed for practitioners (Figure 6 in Appendix A).

Related work. The work closest to ours is Xiong et al. (2023), which adapts estimators of the ATE through a one-shot federated estimation of the outcome/propensity score model parameters. However, their work does not compare the efficiency of these one-shot federated estimators with traditional meta-analysis and pooled dataset estimators, nor does it consider gradient-based federated alternatives. Our results offer clear guidelines on when the One-Shot estimators proposed by Xiong et al. (2023) should be preferred over other methods.

Other work on federated causal inference (see Brantner et al. (2023) and Edmondson et al. (2023) for an overview) consider different settings and objectives than the ones considered in our work. Vo et al. (2022b) employ a Bayesian framework using Gaussian processes to estimate the ATE under uniform data distributions across studies. Terrail et al. (2023) focus on federating an external control arm for time-to-event outcomes, adapting a gradient-based algorithm for Cox hazard model parameters. While our work aims to estimate the causal effect of treatment across the joint population of studies, other studies (Vo et al., 2022a; Han et al., 2021, 2023; Makhija et al., 2024; Guo et al., 2024) aim at transferring causal estimates from a source study to a target population.

The meta-analysis literature on combining estimates from multiple studies is extensive. One can mention the work of Morris et al. (2018) who discuss the differences and advantages of conducting meta-analysis with individual patient data on stratified ("two-stage") versus pooled data ("one-stage"). However, they require sharing raw data and do not explore any federated strategy. Meta-analysis provides considerable flexibility in the choice of local models (Seo et al., 2021; Tan et al., 2022), but also comes with many subtle statistical challenges. These include the ecological fallacy bias (Piantadosi et al., 1988), which occurs when incorrect conclusions about individuals are drawn from subgroup characteristics, as well as situations when ignoring study sizes can lead to biased ATE estimates (Kahan et al., 2023).

2 GENERAL FRAMEWORK

Notations. We consider a set of K studies, with H denoting the random variable with values in $\{1, \dots, K\}$ indicating membership to a study. Let $\mathcal{Z} = \{Z_i\}_{i=1}^n$ be a sample of n independent and identically distributed (i.i.d.) realizations of the quadruplet $Z = (X, W, Y, H)$, where X denotes a d-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^d$, $W \in \{0, 1\}$ denote the binary treatment, and $Y \in \mathbb{R}$ is the observed outcome of interest.

We denote by $\mathcal{Z}_k$ the local dataset of study k with $n_k = \sum_{i=1}^n \mathbb{1}_{\{H_i=k\}}$ observations, and by $\mathcal{Z}_k^{(w)}$ the $n_k^{(w)} = \sum_{i=1}^n \mathbb{1}_{\{H_i=k, W_i=w\}}$ observations in study k under treatment arm w. We further denote by $X_k^{(w)} = \{X_i \mid W_i = w, H_i = k\}_{i=1}^n \in \mathbb{R}^{n_k^{(w)} \times d}$ (resp. $Y_k(w) = \{Y_i \mid W_i = w, H_i = k\}_{i=1}^n \in \mathbb{R}^{n_k^{(w)}}$) the design matrix of the covariates (resp. the outcome vector) for treatment arm w in study k. Similarly, we denote by $\mathcal{Z}^{(w)}$ the $n^{(w)} = \sum_{k=1}^n n_k^{(w)}$ observations under treatment arm w in the pooled dataset $\mathcal{Z}$, and by $X^{(w)} \in \mathbb{R}^{n^{(w)} \times d}$ and $Y^{(w)} \in \mathbb{R}^{n^{(w)}}$ the corresponding design matrix and outcome vector.

Average treatment effect in K RCTs. We consider the setting of K Bernoulli RCT trials, where each participant i in study k has a fixed probability $\mathbb{P}(W_i = 1 \mid H_i = k) = p_k$ of being assigned to the treatment group, which does not depend on X in this design. We denote $\rho_k = \mathbb{P}(H_i = k)$ the probability that an observation belongs to study k ($0 < \rho_k < 1$). Note that the probability p of being treated within the pooled dataset $\mathcal{Z}$ is then given by $p = \sum_{k=1}^K \rho_k p_k$.

We consider the potential outcomes framework (Rubin, 1974) and we aim to estimate the ATE $\tau \in \mathbb{R}$ defined as the Risk Difference over the K studies as $\tau = \mathbb{E}(\mathbb{E}(Y_i(1) - Y_i(0) | H_i))$, where $Y(w)$ is the outcome had the subject received treatment w. We denote the local ATE in study k by $\tau_k = \mathbb{E}(Y_i(1) - Y_i(0) | H_i = k)$.

We assume that the classical identifiability assumptions for a RCT design hold locally at every study: (a) *SUTVA* (Stable Unit Treatment Value Assumption): $Y = WY(1) + (1 - W)Y(0)$, (b) *Positivity*: $\exists \eta_1 > 0$ such that, almost surely, $\eta_1 \leq \mathbb{P}(W = 1 \mid H) \leq 1 - \eta_1$ and (c) *Ignorability*: $W \perp\!\!\!\perp (Y(0), Y(1))$. Under these conditions, the ATE is identifiable and the simple "Difference-in-Means" (Splawa-Neyman et al., 1990) estimator defined as $\hat{\tau}_{\text{DM}} = \frac{1}{n^{(1)}} \sum_{i=1}^{n^{(1)}} Y_i W_i - \frac{1}{n^{(0)}} \sum_{i=1}^{n^{(0)}} Y_i (1 - W_i)$ is an unbiased estimator of the ATE. However, variance reduction can be obtained by adjusting from covariates and considering the "Plug-in G-formula" or "outcome-based regression" estimator.

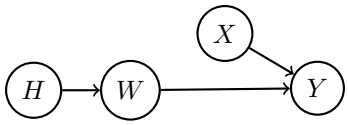


Figure 1: Graphical Model for Homogeneous Settings

Definition 1 (Robins, 1986). *The plug-in G-formula estimator is defined as* $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$, *where* $\hat{\mu}_w(x)$ *is an estimator of the surface response* $\mu_w(x) = \mathbb{E}[Y | W = w, X = x]$.

Throughout the paper, we will consider two different regimes for the sample sizes.

Condition 1 (Local Full Rank). $\forall k \in \{1, \dots, K\}$ *and for* $w \in \{0, 1\}$, *we have* $\text{rank}({X_k^{(w)}}^\top X_k^{(w)}) = d$, *which implies that* $\forall (k, w), n_k^{(w)} \geq d$.

Condition 2 (Federated Full Rank). *For* $w \in \{0, 1\}$, *we have* $\text{rank}({X^{(w)}}^\top X^{(w)}) = d$, *which implies that* $\forall w, \sum_{k=1}^K n_k^{(w)} \geq d$.

3 HOMOGENEOUS POPULATION

In this section, we focus on estimating the ATE over K RCTs studying the same population. We assume that the joint distribution of $Z = (X, W, Y, H)$ decomposes as

$$\mathbb{P}(Z) = \mathbb{P}(H)\mathbb{P}(W|H)\mathbb{P}(X)\mathbb{P}(Y|X, W). \tag{1}$$

This corresponds to the graphical model shown in Figure 1. We refer to this setting as *homogeneous* because X and $Y|X, W$ are independent of H: in other words, there is no distributional shift for the covariates and the conditional outcomes across studies. *Heterogeneous* settings will be addressed in Section 4.

We consider a linear model for the potential outcomes

$$Y_{k,i}(w) = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}(w), \tag{2}$$

with $c^{(w)} \in \mathbb{R}$ the intercept, $\beta^{(w)} \in \mathbb{R}^d$ the coefficients, $\mathbb{E}(\varepsilon_{k,i}(w) \mid X_{k,i}) = 0$ and $\mathbb{V}(\varepsilon_{k,i}(w) \mid X_{k,i}) = \sigma^2$.

Note that $\beta^{(1)}$ and $\beta^{(0)}$ can be different, so that the treatment effect can be heterogeneous (*i.e.* depends on the covariates). We denote $\theta^{(w)} = (c^{(w)}, \beta^{(w)}) \in \mathbb{R}^{d+1}$and $X' = (1, X) \in \mathbb{R}^{n,d+1}$ the covariate matrix augmented with a column of ones. The model parameters $\theta^{(w)}$ are equal in every study, in accordance with the homogeneous setting defined by the decomposition in Eq. 1. We assume that X has finite first two moments $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \Sigma = \mathbb{E}((X - \mu)^\top (X - \mu)$.

Under the above model, the ATE can be written as

$$\tau = \mathbb{E}(X_i')(\theta^{(1)} - \theta^{(0)}) \tag{3}$$

and the ATEs per study $\tau_k = \mathbb{E}(X'_{k,i})(\theta^{(1)} - \theta^{(0)})$ are homogeneous across studies, i.e., $\tau_1 = \cdots = \tau_K = \tau$, as the covariate distribution is the same across studies.

The G-formula estimator on the pooled data $\mathcal{Z}$ can be used to estimate the ATE over the K studies

$$\hat{\tau}_{\text{pool}} = \frac{1}{n}\sum_{i=1}^{n}\left(X'_i\hat{\theta}^{(1)}_{\text{pool}} - X'_i\hat{\theta}^{(0)}_{\text{pool}}\right), \tag{4}$$

where $\hat{\theta}^{(1)}_{\text{pool}}$ and $\hat{\theta}^{(0)}_{\text{pool}}$ are the regression coefficients estimated by fitting two OLS regressions over $\mathcal{Z}^{(1)}$ and $\mathcal{Z}^{(0)}$ respectively. This *pooled* estimator satisfies (see Appendix B.2.6 for a proof extending the standard result in Wager (2020) to non-centered covariates)

$$\sqrt{n}(\hat{\tau}_{\text{pool}} - \tau) \xrightarrow{d} \mathcal{N}(0, V_{\text{pool}}), \tag{5}$$

with $V_{\text{pool}} = \frac{\sigma^2}{p(1-p)} + \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}$. However, computing $\hat{\tau}_{\text{pool}}$ requires access to the *pooled* dataset $\mathcal{Z}$, which is not accessible in the decentralized setting we consider. Under Condition 1, each study in isolation can only estimate the ATE using its local dataset $\mathcal{Z}_k$

$$\hat{\tau}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\left(X'_{k,i}\hat{\theta}^{(1)}_k - X'_{k,i}\hat{\theta}^{(0)}_k\right), \tag{6}$$

where $\hat{\theta}^{(w)}_k = ({X'}^{(w)}_k{}^{\top} {X'}^{(w)}_k)^{-1} {X'}^{(w)}_k{}^{\top} Y^{(w)}_k$ is the OLS estimator computed over $\mathcal{Z}^{(w)}_k$. To improve upon this baseline, we now introduce several estimation strategies of the ATE over the K studies, with the aim of obtaining estimates as if one had access to $\mathcal{Z}$.

3.1 Definition of the Estimators

3.1.1 Meta-Analysis Estimators

A first strategy under Condition 1 is to aggregate the local ATE estimates $\hat{\tau}_k$ in Eq. 6. The studies then send their local ATE estimates to the server, which aggregates them using non-negative weights ω_k that sum to 1 over the K studies to obtain a global ATE estimate. For selecting the weights ω_k, Hunter and Schmidt (2004) describes two common methods for absolute measures like the risk difference: sample size weighting (SW) and inverse variance weighting (IVW). These meta-analysis estimators involve a single round of communication: each study k sends $\hat{\tau}_k$ to the server.

Definition 2 (Meta-Analysis - SW Aggregation)**.**

$$\hat{\tau}_{\text{Meta-SW}} = \sum_{k=1}^{K}\frac{n_k}{n}\hat{\tau}_k. \tag{7}$$

Definition 3 (Meta-Analysis - IVW Aggregation)**.**

$$\hat{\tau}_{\text{Meta-IVW}} = \frac{\sum_{k=1}^{K}\mathbb{V}(\hat{\tau}_k)^{-1}\hat{\tau}_k}{\sum_{k=1}^{K}\mathbb{V}(\hat{\tau}_k)^{-1}}. \tag{8}$$

Proposition 1 (proof in Appendix B.1.2)**.** $\hat{\tau}_{\text{Meta-IVW}}$ *is the minimum-variance estimator of* τ *among the class of aggregation-based estimators.*

Remark 1. *In practice,* $\mathbb{V}(\hat{\tau}_k)$ *is often unknown and must be estimated, leading to an approximation of* $\hat{\tau}_{\text{Meta-IVW}}$*. In contrast,* $\hat{\tau}_{\text{Meta-SW}}$ *only requires knowledge of the local sample sizes.*

3.1.2 One-Shot Federated Estimators

To go beyond the mere aggregation of local ATEs, we can follow Xiong et al. (2023) and aggregate the local outcome model parameters using a single round of communication (hence the term "one-shot" federated) to build better local ATE estimates, before aggregating them.

Step 1. Local estimation of outcome parameters: Under Condition 1, each study k estimates $\hat{\theta}^{(w)}_k$ locally with an OLS regression.

Step 2. One-shot federation of parameters: We perform a meta-analysis (local estimation then weighted aggregation) of the local outcome model parameters $\hat{\theta}^{(w)}_k$. Specifically, the studies send their local estimates to the server, which computes $\hat{\theta}^{(w)}_{\text{1S}} = \sum_{k=1}^{K}\omega^{(\theta)}_k\hat{\theta}^{(w)}_k$ (where "1S" stands for "one-shot"), with $\omega^{(\theta)}_k$ some federation weights (summing to 1 over the K studies) like SW or IVW. The server sends back the obtained $\hat{\theta}^{(w)}_{\text{1S}}$ to all the studies.

Definition 4 (SW Federation of $\hat{\theta}^{(w)}_k$)**.**

$$\hat{\theta}^{(w)}_{\text{1S-SW}} = \sum_{k=1}^{K}\frac{n^{(w)}_k}{n^{(w)}}\hat{\theta}^{(w)}_k. \tag{9}$$

Definition 5 (IVW Federation of $\hat{\theta}^{(w)}_k$)**.**

$$\hat{\theta}^{(w)}_{\text{1S-IVW}} = V^{-1}\sum_{k=1}^{K}\left(\mathbb{V}(\hat{\theta}^{(w)}_k)^{-1}\hat{\theta}^{(w)}_k\right). \tag{10}$$

where $V = \sum_{k=1}^{K}\mathbb{V}(\hat{\theta}^{(w)}_k)^{-1} = \sum_{k=1}^{K}\frac{1}{\sigma^2}{X'}^{(w)}_k{}^{\top}{X'}^{(w)}_k$.

Theorem 1 (proof in Appendix B.2.1)**.** *Under Condition 1,* $\hat{\theta}^{(w)}_{\text{1S-IVW}} = \hat{\theta}^{(w)}_{\text{pool}}$.

Remarkably, Theorem 1 shows that one can obtain similar estimates as $\hat{\theta}_{\text{pool}}$ (which has the lowest variance among the class of linear unbiased estimators (Giraud et al., 2012)) by federating the local estimates with a one-shot IVW procedure, even in finite sample sizes, whenever the local datasets are of full rank. This gives a very strong argument in favor of this approach in comparison to the One-Shot SW which thus necessarily has higher variance than $\hat{\theta}_{\text{1S-IVW}}$ and $\hat{\theta}_{\text{pool}}$. However, note that the communication cost of computing $\hat{\theta}^{(w)}_{\text{1S-IVW}}$ is $O(d)$ times larger than for $\hat{\theta}^{(w)}_{\text{1S-SW}}$, as each study must send to the server its $(d+1)\times(d+1)$ local variance matrix ${X'}^{(w)}_k{}^{\top}{X'}^{(w)}_k$.

Step 3. Aggregation of the ATEs: Each study estimates its local ATE using the federated outcome

model parameters:

$$\hat{\tau}_k^{\text{1S-agg}} = \frac{1}{n_k}\sum_{i=1}^{n_k}(X'_{k,i}\hat{\theta}^{(1)}_{\text{1S-agg}} - X'_{k,i}\hat{\theta}^{(0)}_{\text{1S-agg}}) \quad (11)$$

with agg $\in$ {SW, IVW}. Finally, a second communication round is used where studies each send their $\hat{\tau}_k^{\text{1S-agg}}$ to the server for aggregation with weights $\omega^{(\tau)}$:

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

It turns out that using SW or IVW for $\omega^{(\tau)}$ is asymptotically equivalent (see Appendix B.2.6), so in the following we focus on sample size aggregation weights.

Definition 6 (1S SW Federation - SW Aggregation)**.**

$$\hat{\tau}_{\text{1S-SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{1S-SW}} \quad (12)$$

Definition 7 (1S IVW Federation - SW Aggregation)**.**

$$\hat{\tau}_{\text{1S-IVW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{1S-IVW}} \quad (13)$$

3.1.3 Gradient-based Federated Estimators

Neither the meta nor the one-shot estimators can be used when Condition 1 does not hold (e.g., as soon as one study has its sample size $n_k^{(w)} < d$), since $\hat{\theta}_k^{(w)}$ is not defined. In such cases, we propose to leverage gradient-based federated estimators.

Step 1. Multi-shot federation of parameters: Studies jointly estimate $\hat{\theta}^{(w)}_{\text{pool}}$ by solving the underlying OLS problem in a federated fashion. To the best of our knowledge, we are the first to propose this approach to estimate the ATE in Federated Causal Inference. Finding $\hat{\theta}^{(w)}_{\text{pool}}$ amounts to minimizing the Mean Squared Error loss function, defined as $\ell(\theta^{(w)}, \bigcup_{k=1}^{K} \mathcal{Z}_k^{(w)}) = \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} (Y_i^{(w)} - X'^{(w)}_i \theta^{(w)})^2$. This optimization problem can be solved by a federated gradient descent-based algorithm. We propose to use the FedAvg algorithm (McMahan et al., 2017) for its intuitive simplicity, strong convergence guarantees, and good empirical performance on both homogeneous (Stich, 2019; Khaled et al., 2020) and heterogeneous data (Wang et al., 2024). FedAvg alternates for T rounds between performing E local gradient steps in each study and aggregating the parameters at the server, see Appendix D for the detailed algorithm. The output of this procedure is an estimate $\hat{\theta}^{(w)}_{\text{GD}}$ ("GD" stands for Gradient Descent).

Let $\lambda_{\max,k}$ be the largest eigenvalue of the covariance matrix in study k. If we set $T = 1$ (a single communication round) and the local learning rate for study k to $2/\lambda_{\max,k}$, then $\hat{\theta}^{(w)}_{\text{GD}}$ is guaranteed to converge to the one-shot estimate as $\hat{\theta}^{(w)}_{\text{SW}}$ as the number of local steps $E \to \infty$. Conversely, if we set the number of local steps $E = 1$ and the global learning rate to $\frac{2}{\sum_{k=1}^{K} \lambda_{\max,k}}$, then $\hat{\theta}^{(w)}_{\text{GD}}$ is guaranteed to converge to the pooled estimate $\theta^{(w)}_{\text{pool}}$ as $T \to \infty$. More details are provided in Appendix D. FedAvg thus allows to learn a better estimate of the outcome model parameters than one-shot approaches (in particular when Condition 1 does not hold) at the cost of more communication. In our homogeneous case, an extremely accurate estimate of $\hat{\theta}^{(w)}_{\text{pool}}$ can be obtained with a small number of communication rounds T (see Appendix D). Then, the estimation error on the ATE with $\hat{\theta}^{(w)}_{\text{GD}}$ compared to $\hat{\theta}^{(w)}_{\text{pool}}$ satisfies $(\hat{\tau}_{\text{pool}} - \hat{\tau}_{\text{GD}})^2 \leq \|\frac{2}{n}\sum_{i=1}^{n} X'_i\|^2 (\epsilon^{(1)} + \epsilon^{(0)})$, with $\epsilon^{(w)} = \|\hat{\theta}^{(w)}_{\text{pool}} - \hat{\theta}^{(w)}_{\text{GD}}\|_2^2$ the error on the parameters.

In the following, we consider that the parameters of FedAvg are chosen such that $\hat{\theta}^{(w)}_{\text{GD}} = \hat{\theta}^{(w)}_{\text{pool}}$.

Step 2. Aggregation of the ATEs: Each study k then computes a local estimate of the ATE using $\hat{\theta}^{(w)}_{\text{GD}}$

$$\hat{\tau}_k^{\text{GD}} = \frac{1}{n_k}\sum_{i=1}^{n_k} \left(X'_{k,i}\hat{\theta}^{(1)}_{\text{GD}} - X'_{k,i}\hat{\theta}^{(0)}_{\text{GD}}\right) \quad (14)$$

Finally, these estimates are aggregated in a last round of communication with sample weighting (IVW is asymptotically equivalent, see Section 3.1.2), which yields the following GD estimator of the global ATE.

Definition 8 (GD Federation - SW Aggregation)**.**

$$\hat{\tau}_{\text{GD}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}_k^{\text{GD}} \quad (15)$$

3.2 Comparison of the Federated Estimators

General case. Under the graphical model in Figure 1 and Condition 1, all ATE estimators presented so far are unbiased (as proved in Appendix B.2.5).

For each estimator, we report in Table 1 its asymptotic variance $\mathbb{V}^{\infty}$ (with proofs in Appendix B.2.6), the sample size required (as per Condition 1 or the weaker Condition 2), the number of communication rounds needed between the studies and the server, and the total communication cost per study (in number of floats). We observe that the one-shot and gradient-based federated estimators achieve the same variance of the pooled-data estimator. The differences lie in the sample size conditions and communication costs. While one-shot estimators require two communication rounds (and generally lower total communication costs), they require that the *local* sample size $n_k^{(w)}$ at each study k and arm w be larger than the dimension d. In contrast, the GD estimator requires this only for the *pooled* sample size $n^{(w)} = \sum_{k=1}^{K} n_k^{(w)}$.

Theorem 2 (Proof in Appendix B.2.7)**.** *Under graphical model in Fig. 1, the estimators compare as:*

$$\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{fed}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}})$$

Table 1: Properties of the (unbiased) estimators of the ATE in the homogeneous setting: asymptotic variance, number of communication rounds and total communication cost (in number of floats per study).

Estimator	Notation	Condition	$\mathbb{V}^\infty$	Com. rounds	Com. cost
Local	$\hat{\tau}_k$ (Eq. 6)	Cond. 1	$\frac{\sigma^2}{n_k}\frac{1}{p_k(1-p_k)} + \frac{1}{n_k}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2$	0	0
Meta-SW	$\hat{\tau}_{\text{Meta-SW}}$ (Eq. 7)	Cond. 1	$\frac{\sigma^2}{n}\sum_{k=1}^K \frac{\rho_k}{p_k(1-p_k)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2$	1	$O(1)$
Meta-IVW	$\hat{\tau}_{\text{Meta-IVW}}$ (Eq. 8)	Cond. 1	$\Big(\sum_{k=1}^K \Big(\sigma^2\frac{n\rho_k}{p_k(1-p_k)} + \frac{1}{n_k}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2\Big)^{-1}\Big)^{-1}$	1	$O(1)$
1S-SW	$\hat{\tau}_{\text{1S-SW}}$ (Eq. 12)	Cond. 1	$\frac{\sigma^2}{n}\frac{1}{p(1-p)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2$	2	$O(d)$
1S-IVW	$\hat{\tau}_{\text{1S-IVW}}$ (Eq. 13)	Cond. 1	$\frac{\sigma^2}{n}\frac{1}{p(1-p)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2$	2	$O(d^2)$
GD	$\hat{\tau}_{\text{GD}}$ (Eq. 15)	Cond. 2	$\frac{\sigma^2}{n}\frac{1}{p(1-p)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2$	$T+1$	$O(Td)$
Pool	$\hat{\tau}_{\text{pool}}$ (Eq. 4)	Cond. 2	$\frac{\sigma^2}{n}\frac{1}{p(1-p)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_\Sigma^2$	—	—

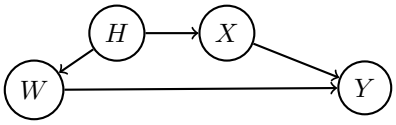


Figure 2: Graphical Model in the Heterogeneous Distributions Setting.

where fed $\in$ {GD, 1S-IVW, 1S-SW}. Theorem 2 shows that meta-analysis estimators typically exhibit larger variance. This happens as soon as treatment probabilities are not equal across studies, and the variance difference increases as the treatment probabilities $\{p_k\}_k$ become more distinct. Moreover, $\mathbb{V}^\infty(\hat{\tau}_{\text{Meta-IVW}}) < \mathbb{V}^\infty(\hat{\tau}_{\text{Meta-SW}})$ whenever $\{p_k(1-p_k)\}_k$ differ across studies, and the difference increase as the difference between these quantities increase. We provide examples in Appendix B.2.7.

Special case: one RCT conducted across K studies. Consider the special case where $W \perp\!\!\!\perp H$, i.e., there is no edge from H to W in the graphical model of Figure 1. This corresponds to a single Bernoulli RCT with treatment probability p implemented across multiple studies. Then, under Condition 1, all estimators are asymptotically equivalent (see Appendix B.2.8) and thus meta-analysis estimators should be used as they require a single round of communication.

4 HETEROGENEOUS SCENARIOS

4.1 Distributional Shifts in Covariates

In model (2), we do not assume $H \perp\!\!\!\perp X$ and consider the graphical model in Figure 2, depicting *Heterogeneous* Bernoulli Trials in the distribution of $X|H$.

Here, $\mathbb{P}(Z) = \mathbb{P}(Y|X,W)\mathbb{P}(X|H)\mathbb{P}(W|H)\mathbb{P}(H)$, hence the observations $(X_{k,i}, W_{k,i}, Y_{k,i})_i$ are i.i.d. within study k but not necessarily across studies. We denote $\mathbb{E}(X_{k,i}) = \mu_k$, $\mathbb{V}(X_{k,i}) = \Sigma_k$ and we have $\Sigma = \mathbb{V}(X) = \sum_{k=1}^K \rho_k \Sigma_k$ and $\mu = \mathbb{E}(X) = \sum_{k=1}^K \rho_k \mu_k$. Furthermore, the local ATEs $\tau_k = \mathbb{E}(Y_k(1) - Y_k(0)) = \mathbb{E}(X_k')(\theta^{(1)} - \theta^{(0)})$ generally differ from each other and from the global ATE $\tau = \mathbb{E}(Y(1) - Y(0)) = \mathbb{E}(X')(\theta^{(1)} - \theta^{(0)})$.

Meta estimators. The ATE can be written as $\tau = \sum_{k=1}^K \rho_k \tau_k$, so the Meta-SW estimator remains unbiased (as n_k/n is an unbiased estimate of ρ_k) and has the same variance as in Table 1 with $\Sigma = \sum_{k=1}^K \rho_k \Sigma_k$ (proof in Appendix B.3.1). In contrast, the Meta-IVW estimator becomes unsuitable under distributional shifts: IVW weights give biased estimates of the ρ_k's, leading to a biased estimate of τ.

Pool and GD estimators. The pooled and GD estimators are robust to covariate shifts, leading them to be also unbiased with same variance as in Table 1 (as proved in Appendix B.3.1).

One-shot estimators. Although the One-Shot IVW outcome parameters estimator still enjoys Theorem 1, the variance of the One-Shot SW one is impacted by the difference in population means at each study.

Proposition 2 (Larger Variance of 1S-SW, proof in B.3.1)**.** *Under Condition 1, the one-shot federated estimators are unbiased and*

$$\mathbb{V}(\hat{\theta}_{\text{pool}}) = \mathbb{V}(\hat{\theta}_{\text{GD}}) = \mathbb{V}(\hat{\theta}_{\text{1S-IVW}}) \leq \mathbb{V}(\hat{\theta}_{\text{1S-SW}})$$

which yields

$$\mathbb{V}^\infty(\hat{\tau}_{\text{pool}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{GD}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{1S-IVW}}) \leq \mathbb{V}^\infty(\hat{\tau}_{\text{1S-SW}})$$

Theorem 3 (Comparison of asymptotic variances under distributional shift)**.**

$$\mathbb{V}^\infty(\hat{\tau}_{\text{pool}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{GD}}) = \mathbb{V}^\infty(\hat{\tau}_{\text{1S-IVW}}) \leq \mathbb{V}^\infty(\hat{\tau}_{\text{Meta-SW}})$$

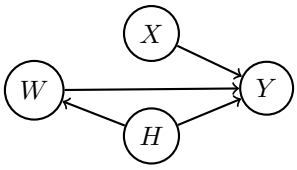


Figure 3: Graphical model for the heterogeneous study-effects setting.

Note that in this setting the DM estimators are no longer unbiased as there is a path between W and Y in Figure 2. Here, X is a sufficient adjustment set and adding H could be harmful, particularly when the association between X and Y is weak and the association between H and W is strong (Rotnitzky and Smucler, 2020, Lemma 2 therein).

4.2 Study-Effects

We now consider the graphical model shown in Figure 3, exhibiting an effect of study H onto the outcome Y. The distribution of Z decomposes as $\mathbb{P}(Z) = \mathbb{P}(Y|X,W,H)\mathbb{P}(X|H)\mathbb{P}(W|H)\mathbb{P}(H)$. We modify model (2) to account for a constant study effect on individual outcomes by adding a term $h_k \in \mathbb{R}$:

$$Y_{k,i}(w) = c^{(w)} + h_k + X_{k,i}\beta^{(w)} + \varepsilon_i(w) \tag{16}$$

Model (16) accounts for the possibility that studies may have different baselines in individual outcomes resulting from varying practices or organizational contexts. Here, τ and τ_k are still defined as $\mathbb{E}(Y(1)-Y(0)) = c^{(1)} - c^{(0)} + \mathbb{E}(X)(\beta^{(1)} - \beta^{(0)})$ and $\mathbb{E}(Y_k(1)-Y_k(0)) = c^{(1)} - c^{(0)} + \mathbb{E}(X_k)(\beta^{(1)} - \beta^{(0)})$ respectively since the h_k terms cancel out in the differences. In other words, this modeling assumes that the Conditional Average Treatment Effect (CATE) remains consistent across studies, while allowing for an additive shift in the outcomes at each study. In this setting, aggregating multiple RCTs into the pooled data is not itself an RCT because H is now a confounder, affecting both the outcome variable $Y_{k,i}(w)$ through h_k and the treatment variable $W_{k,i}$ through the treatment probability p_k, thereby violating the ignorability assumption (c). Therefore, the (unadjusted) pooled OLS estimator (4) is biased (see proof in Appendix B.4.1). An unbiased estimator can be obtained by adjusting the model to incorporate the study effect.

Meta estimators. While most estimators need to be adjusted and thus require prior knowledge on the underlying model, Meta-analysis estimators can be applied directly without such modifications. We prove in Appendix B.4 that under Graphical Model 3 and model (16), the Meta-IVW (which is relevant here as $H \perp\!\!\!\perp X$) and Meta-SW estimators remain unbiased, with asymptotic variances as in Table 1.

Adjusted gradient-based estimator. We augment the design matrix X with $K-1$ dummy variables $H = \{H_2, \dots, H_K\}$ and note $\tilde{X}'_{k,i} = (1, X_{k,i}, H_{2,i}, \dots, H_{K,i})$. Then, the method is the same as in Section 3.1.3: $\hat{\theta}^{(w)}_{\text{GD}\circ} = \left(\hat{c}^{(w)}, \hat{\beta}^{(w)}, \hat{h}_2, \dots, \hat{h}_K\right) \in \mathbb{R}^{d+K}$ is obtained with FedAvg and used to compute the local ATEs before final aggregation.

Definition 9 (Adjusted GD estimator)**.**

$$\hat{\tau}_{\text{GD}\circ} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\tilde{X}'_{k,i}(\hat{\theta}^{(1)}_{\text{GD}\circ} - \hat{\theta}^{(0)}_{\text{GD}\circ})\right)$$

This estimator is unbiased, and we do not pay a price in terms of asymptotic variance in adjusting the variables H. Indeed, its asymptotic variance is equal to the unadjusted $\hat{\tau}_{\text{GD}}$'s one in Table 1 (proof in Appendix B.4.2) since h_k is equal in both treatment arms and cancel out in the difference of the true parameters.

Adjusted one-shot federated estimators. Like the vanilla pooled and GD estimators, the one-shot estimators are biased in the presence of study-effects. However, their adjustment procedure is different because the one-shot procedure does not allow the inclusion of membership variables, as the variance matrices of the local dummy-augmented datasets are not full rank, violating Condition 1. Instead, we compute the OLS $\hat{\theta}_k$ at each study, then share and aggregate only the coefficients $\hat{\beta}^{(w)}_k$, without federating the locally estimated intercepts $\hat{a}^{(w)}_k = \hat{c}^{(w)} + \hat{h}_k$.

Definition 10 (Adjusted one-shot ATE estimators)**.**

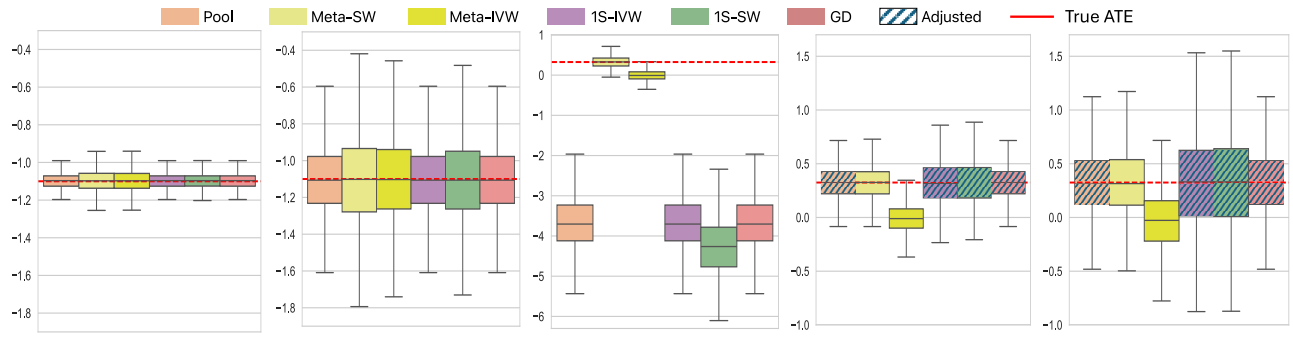
$$\hat{\tau}_{\text{1S-SW}\circ} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\hat{a}^{(1)}_k - \hat{a}^{(0)}_k + X_{k,i}(\hat{\beta}^{(1)}_{\text{SW}\circ} - \hat{\beta}^{(0)}_{\text{SW}\circ})\right)$$

$$\hat{\tau}_{\text{1S-IVW}\circ} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\hat{a}^{(1)}_k - \hat{a}^{(0)}_k + X_{k,i}(\hat{\beta}^{(1)}_{\text{IVW}\circ} - \hat{\beta}^{(0)}_{\text{IVW}\circ})\right)$$

$$\textit{with } \hat{\beta}^{(w)}_{\text{IVW}\circ} = \frac{\sum_{k=1}^{K}\mathbb{V}(\hat{\beta}^{(w)}_k)^{-1}\hat{\beta}^{(w)}_k}{\sum_{l=1}^{K}\mathbb{V}(\hat{\beta}^{(w)}_l)^{-1}},\ \hat{\beta}^{(w)}_{\text{SW}\circ} = \sum_{k=1}^{K}\frac{n^{(w)}_k}{n^{(w)}}\hat{\beta}^{(w)}_k.$$

Note that, in general, $\hat{\beta}^{(w)}_{\text{IVW}\circ} \neq \hat{\beta}^{(w)}_{\text{IVW}}$ as the aggregation weights $\mathbb{V}(\hat{\beta}^{(w)}_k)^{-1}$ and $\mathbb{V}(\hat{\theta}^{(w)}_k)^{-1}$ are different.

The variances of the adjusted one-shot estimators are affected by the lack of federation of the local intercepts $\{\hat{a}^{(w)}_k\}_k$, which converge at a rate of $1/n^{(w)}_k$ rather than $1/n^{(w)}$. Additionally, they estimate the study-effects twice in each study on independent data, whereas GD$\circ$ and pool$\circ$ estimate them on $n^{(w)}$ observations. As a result, adjusted one-shot estimators generally underperform compared to other methods (see simulations in Appendix C.2).



(a) Homogeneous Case (Figure 1): Large (left) and Small (right) regimes

(b) Fully Heterogeneous Case (Figure 5): Large regime, unadjusted (left); Large regime, adjusted (middle); Small regime, adjusted (right)

Figure 4: Multi-study ATE Estimation under Homogeneous and Heterogeneity Scenarios

5 EXPERIMENTS

We now present some numerical experiments on simulated and semi-synthetic data.

We compare the following estimators: Meta-SW (Definition 2), Meta-IVW (Definition 3), 1S-SW (Definition 6), 1S-IVW (Definition 7) and GD (Definition 8), and the adjusted estimators GD$\circ$ (Definition 9), 1S-SW$\circ$ and 1S-IVW$\circ$ (Definition 10). We also include the pooled estimator (Pool, Eq. 4) as a baseline. For the Meta-IVW estimator, we use empirical estimates of $\mathbb{V}^{\infty}(\hat{\tau}_k)^{-1}$ for the aggregation weights.

5.1 Synthetic Data

We generate data according to the graphical models in Figures (1) and (2), with $K = 5$ studies and $d = 10$ covariates. We consider two magnitudes of sample sizes, referred to as *Large* ($\forall k, n_k = 20^*d$) and *Small* ($\forall k, n_k = 6^*d$) local sample sizes. We consider the following treatment assignment probabilities for the *Large* setting: $p_1 = p_2 = p_3 = 0.9, p_4 = p_5 = 0.1$. In the *Small* regime we choose less extreme probabilities in order to guarantee Condition 1: $p_1 = p_2 = p_3 = 0.65$, $p_4 = p_5 = 0.35$. For each scenario considered, we perform 2000 simulations and display the distribution of the global estimates of the ATE given by each estimator. More details about the simulations can be found in Appendix C.1.

Homogeneous case. The results displayed in Figure 4a (left) are in agreement with Theorem 2: the variances of the meta estimators are larger than the pooled, one-shot and gradient-based estimators when the assignment probabilities p_k differ from one study to another, with the variance of the Meta-IVW being smaller than that of Meta-SW (Proposition 1). In the *Small* regime (Figure 4a, right), the variance of One-Shot SW is larger than compared to One-Shot IVW and GD, but we still have $\mathbb{V}(\hat{\tau}_{\text{pool}}) = \mathbb{V}(\hat{\tau}_{\text{GD}}) = \mathbb{V}(\hat{\tau}_{\text{1S-IVW}})$, in line with Theorem 1.

Heterogeneous case. We now consider a "fully heterogeneous" setting combining three sources of heterogeneity (see graphical model in Figure 5): different covariate distributions across studies, presence of study-effects, and different p_k (see Appendix C.5 for the chosen values of $\{(\mu_k, \Sigma_k, h_k, p_k)\}_k$). Figure 4b (left) shows that the pooled, one-shot and gradient-based estimators are biased when we do not adjust for the study-effects. Here, only the Meta SW estimator is unbiased.

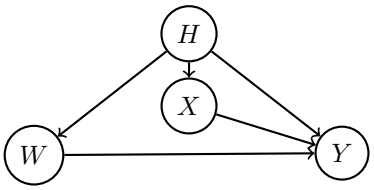


Figure 5: Graphical model in fully heterogeneous case.

Figure 4b (middle, right) shows the results of adjusted estimators, as described in Section 4.2. Adjusting removes their bias but comes at a cost for the one-shot estimators, whose variances are impacted by study-effects, unlike the meta and gradient-based estimators. In the *Small* regime, the one-shot estimators further suffer from the combination of differences in covariate means and unshared local intercepts, which inflate $\mathbb{V}^{\infty}(\hat{a}_k^{(w)})$. On the other hand, the variance of the Meta-SW estimator is nearly equal to that of the adjusted pooled and adjusted GD estimators.

5.2 Semi-Synthetic Data

We now provide results on the Traumatrix database (Mayer et al., 2020) to test our theory on real-world data with both linear and non-linear synthetic outcomes. We focus on brain trauma patients and consider 8,097 patients scattered across 13 sites, suffer-

ing from traumatic brain injury (TBI), with 15 covariates such as systolic and diastolic blood pressure, heart rate, oxygen saturation, and information on interventions like catecholamine administration. We consider each site as an independent study, with different treatment probabilities and potential covariate shifts, corresponding to the graphical model in Figure 2.

We regenerate the treatment and outcome variables while keeping the covariates as is. The synthetic treatment, independent of the covariates as in an RCT setting, is generated as a Bernoulli variable that has probability p_k at site k, with p_k ranging from 0.2 to 0.8. Then, in the linear setting, we generate two continuous outcomes $Y^{(1)}$ and $Y^{(0)}$, respectively as $X'\theta^{(1)}+\varepsilon$ and $X'\theta^{(0)}+\varepsilon$, with $X' \in \mathbb{R}^{8,097\times 16}$, $\theta^{(1)}, \theta^{(0)} \in \mathbb{R}^{16}$ the parameters generated (drawn from $U([-1,1]^d)$) once with each coordinate $\theta_l^{(1)} = \theta_l^{(0)} + 0.05$ so that the true ATE $\tau = \mathbb{E}(X)(\theta^{(1)} - \theta^{(0)}) \approx \overline{X}(\theta^{(1)} - \theta^{(0)}) \approx 0.37$, and $\varepsilon \sim N(0, 2)$. Finally, we bootstrapped 3,000 times the dataset and computed the estimators on these resampled data to estimate their means and variances.

We also consider non-linear outcomes, repeating the same steps as above but with polynomial non-linear outcome model equations with interaction terms $\mu_w(X)=\theta_0^{(w)}+\sum_{j=1}^{3} X_j^j\theta_j^{(w)}+\sum_{j=4}^{16} X_j\theta_j^{(w)}+\theta_{\text{int}}^{(w)}X_{\text{int}}+\varepsilon$, where the interaction terms are given by $X_{\text{int}} = (-X_2\times X_3, X_1\times X_4)$. This leads to a true ATE $\tau \approx 1.07$. For these non-linear outcomes, the estimators still rely on linear regressions to model the outcomes. As stated above, even if the model is misspecified in this case, adjusting for covariates is still recommended in RCTs to reduce variance.

Note that we chose p_k and $n_{\text{bootstrap}} = 4,000$ to ensure that Condition 1 holds, as some sites have very small datasets. In this regard, the semi-synthetic simulation does not match the asymptotic regime on which our results rely, but it echoes the empirical results on the finite-sample regime.

Estimator	Squared Bias	RMSE
Meta SW	0.000	1.238
Meta IVW	0.002	0.116
One-Shot SW, SW Agg	0.004	1.161
One-Shot IVW, SW Agg	0.005	0.117
GD	0.001	0.073
Pool	0.001	0.073

Table 2: ATE estimation on semi-synthetic data with linear outcome

As expected, most estimators appear to achieve a low (empirical) bias, but differ significantly in terms of Root Mean Square Error (RMSE). Additionally, as the p_k are different, the meta estimators struggle more. In the linear setting (Table 2), the variance ranking of the estimators aligns with our theory: the Pool and GD-based estimators yield the same results, with the lowest variance: we have $\mathbb{V}(\hat{\tau}_{\text{pool}}) = \mathbb{V}(\hat{\tau}_{\text{GD}}) \leq \mathbb{V}(\hat{\tau}_{\text{meta-IVW}}) \leq \mathbb{V}(\hat{\tau}_{\text{meta-SW}})$.

Estimator	Squared Bias	RMSE
Meta SW	0.004	1.923
Meta IVW	0.007	0.136
One-Shot SW, SW Agg	0.010	1.064
One-Shot IVW, SW Agg	0.002	0.120
GD	0.001	0.078
Pool	0.001	0.078

Table 3: ATE estimation on semi-synthetic data with nonlinear outcome

In the non-linear setting (Table 3), the same conclusions seem to hold, although a more thorough analysis should be conducted to theoretically corroborate these results.

6 CONCLUSION

After clearly defining the population and estimand of interest, our findings can be turned into clear guidelines for practitioners that we summarize as a decision diagram (Figure 6 in Appendix A). We recommend the one-shot IVW estimator when each study can perform a local OLS regression (Condition 1) and there are no study-effects. However, this estimator requires studies to share sample covariance matrices, which can be impractical in high-dimensional settings or when privacy is a concern. On the other hand, meta estimators are preferable when study-effects are present or when there is limited prior knowledge about the underlying model, provided Condition 1 holds. Caution is required with Meta-IVW, which has lower variance than Meta-SW but is biased when studies analyze different populations. Finally, our (adjusted) GD estimator allows us to estimate the ATE under the weaker Condition 2 with the same precision as if the data were pooled, regardless of the setting.

Several challenges remain for the deployment of these methods, particularly in medical contexts. Future work includes addressing covariate mismatch, where some features are missing in specific studies, estimating non-collapsible causal measures (e.g., odds ratio), and extensions to observational studies. Finally, a key practical challenge in multi-study settings is to ensure consistent data encoding, especially for outcomes.

References

David Benkeser, Iván Díaz, Alex Luedtke, Jodi Segal, Daniel Scharfstein, and Michael Rosenblum. Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*, 77(4):1467–1481, 2021.

Jesse A. Berlin and Robert M. Golub. Meta-analysis as evidence: Building a better pyramid. *JAMA*, 312 (6):603–606, 2014. doi: 10.1001/jama.2014.8167.

Christopher Blunt. *Hierarchies of evidence in evidence-based medicine*. PhD thesis, London School of Economics and Political Science, 2015.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2021.

Carly Lupton Brantner, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A Stuart. Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical Science*, 38(4):640–654, 2023.

Sicong Che, Zhaoming Kong, Hao Peng, Lichao Sun, Alex Leow, Yong Chen, and Lifang He. Federated multi-view learning for private medical data integration and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23, 2022.

Mackenzie J Edmondson, Chongliang Luo, and Yong Chen. Statistical analysis—meta-analysis/reproducibility. In *Clinical Applications of Artificial Intelligence in Real-World Data*, pages 125–139. Springer, 2023.

European Medicines Agency. ICH E9 Statistical Principles for Clinical Trials: Scientific Guideline, 2024.

French Health Authority. Pricing & reimbursement of drugs and hta policies in france, 2024. URL https://www.has-sante.fr/upload/docs/application/pdf/2014-03/pricing_reimbursement_of_drugs_and_hta_policies_in_france.pdf.

Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. 2012.

Tianyu Guo, Sai Praneeth Karimireddy, and Michael I. Jordan. Collaborative heterogeneous causal inference beyond meta-analysis. *arXiv preprint arXiv:2404.15746*, 2024.

Gordon Guyatt, Drummond Rennie, Maureen O. Meade, and Deborah J. Cook. *Users' Guides to the Medical Literature : A Manual for Evidence-Based Clinical Practice*. McGraw-Hill Education, New York, 2015.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *arXiv preprint arXiv:2112.09313*, 2021.

Larry Han, Zhu Shen, and José R. Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

JM Hernan, MA Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC., 2020.

Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Chichester, UK, 2nd edition, 2019.

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

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge UK, 2015.

Brennan C Kahan, Fan Li, Andrew J Copas, and Michael O Harhay. Estimands in cluster-randomized trials: choosing analyses that answer the right question. *International Journal of Epidemiology*, 52(1): 107–118, 2023.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *AISTATS*, 2020.

Tatsuki Koga, Kamalika Chaudhuri, and David Page. Differentially private multi-site treatment effect estimation. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 472–489. IEEE, 2024.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5381–5393. PMLR, 13–18 Jul

2020. URL https://proceedings.mlr.press/v119/koloskova20a.html.

Lihua Lei and Peng Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828, 2021.

A Liberati, DG Altman, J Tetzlaff, and C Mulrow. G0tzsche pc, ioannidis jp, et al. the prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol*, 62(10):e1–34, 2009.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. 2013.

Disha Makhija, Joydeep Ghosh, and Yejin Kim. Federated learning for estimating heterogeneous treatment effects. *CoRR*, abs/2402.17705, 2024. doi: 10.48550/ARXIV.2402.17705. URL https://doi.org/10.48550/arXiv.2402.17705.

Imke Mayer, Erik Sverdrup, Tobias Gauss, Jean-Denis Moyer, Stefan Wager, and Julie Josse. Doubly robust treatment effect estimation with missing attributes. *The Annals of Applied Statistics*, 14(3): 1409 – 1431, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

David Moher, Deborah J Cook, Susan Eastwood, Ingram Olkin, Drummond Rennie, and Donna F Stroup. Improving the quality of reports of meta-analyses of randomised controlled trials: the quorom statement. *The Lancet*, 354(9193):1896–1900, 1999.

Tim P Morris, David J Fisher, Michael G Kenward, and James R Carpenter. Meta-analysis of gaussian individual patient data: Two-stage or not two-stage? *Statistics in medicine*, 37(9):1419–1438, 2018.

Steven Piantadosi, David P Byar, and Sylvan B Green. The ecological fallacy. *American journal of epidemiology*, 127(5):893–904, 1988.

Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2 (7):369–375, 2020.

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.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188): 1–86, 2020. URL http://jmlr.org/papers/v21/19-1026.html.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66(5):688–701, 1974. ISSN 0022-0663.

Michael Seo, Ian R White, Toshi A Furukawa, Hissei Imai, Marco Valgimigli, Matthias Egger, Marcel Zwahlen, and Orestis Efthimiou. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. *Statistics in medicine*, 40(6):1553–1573, 2021.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472, 1990. doi: 10.1214/ss/1177012031. URL https://doi.org/10.1214/ss/1177012031.

Sebastian U. Stich. Local sgd converges fast and communicates little. In *ICLR*, 2019.

Xiaoqing Tan, Chung-Chou H Chang, Ling Zhou, and Lu Tang. A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources. In *International Conference on Machine Learning*, pages 21013–21036. PMLR, 2022.

Jean Ogier du Terrail, Quentin Klopfenstein, Honghao Li, Imke Mayer, Nicolas Loiseau, Mohammad Hallal, Félix Balazard, and Mathieu Andreux. Fedeca: A federated external control arm method for causal inference with time-to-event data in distributed settings. *arXiv preprint arXiv:2311.16984*, 2023.

Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677, 2008.

U.S. Food and Drug Administration. Adjusting for covariates in randomized clinical trials for drugs and biological products, 2023.

Kelly Van Lancker, Frank Bretz, and Oliver Dukes. Covariate adjustment in randomized controlled tri-

als: General concepts and practical considerations. *Clinical Trials*, page 17407745241251568, 2024.

Thanh Vinh Vo, Arnab Bhattacharyya, Young Lee, and Tze-Yun Leong. An adaptive kernel approach to federated learning of heterogeneous causal effects. *Advances in Neural Information Processing Systems*, 35:24459–24473, 2022a.

Thanh Vinh Vo, Young Lee, Trong Nghia Hoang, and Tze-Yun Leong. Bayesian federated estimation of causal effects from observational data. In *UAI*, 2022b.

Stefan Wager. Stats 361: Causal inference. Technical report, 2020.

Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. A new theoretical perspective on data heterogeneity in federated optimization. In *International Conference on Machine Learning (ICML)*, 2024.

Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *Statistics in Medicine*, 42(24):4418–4439, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A DECISION DIAGRAM

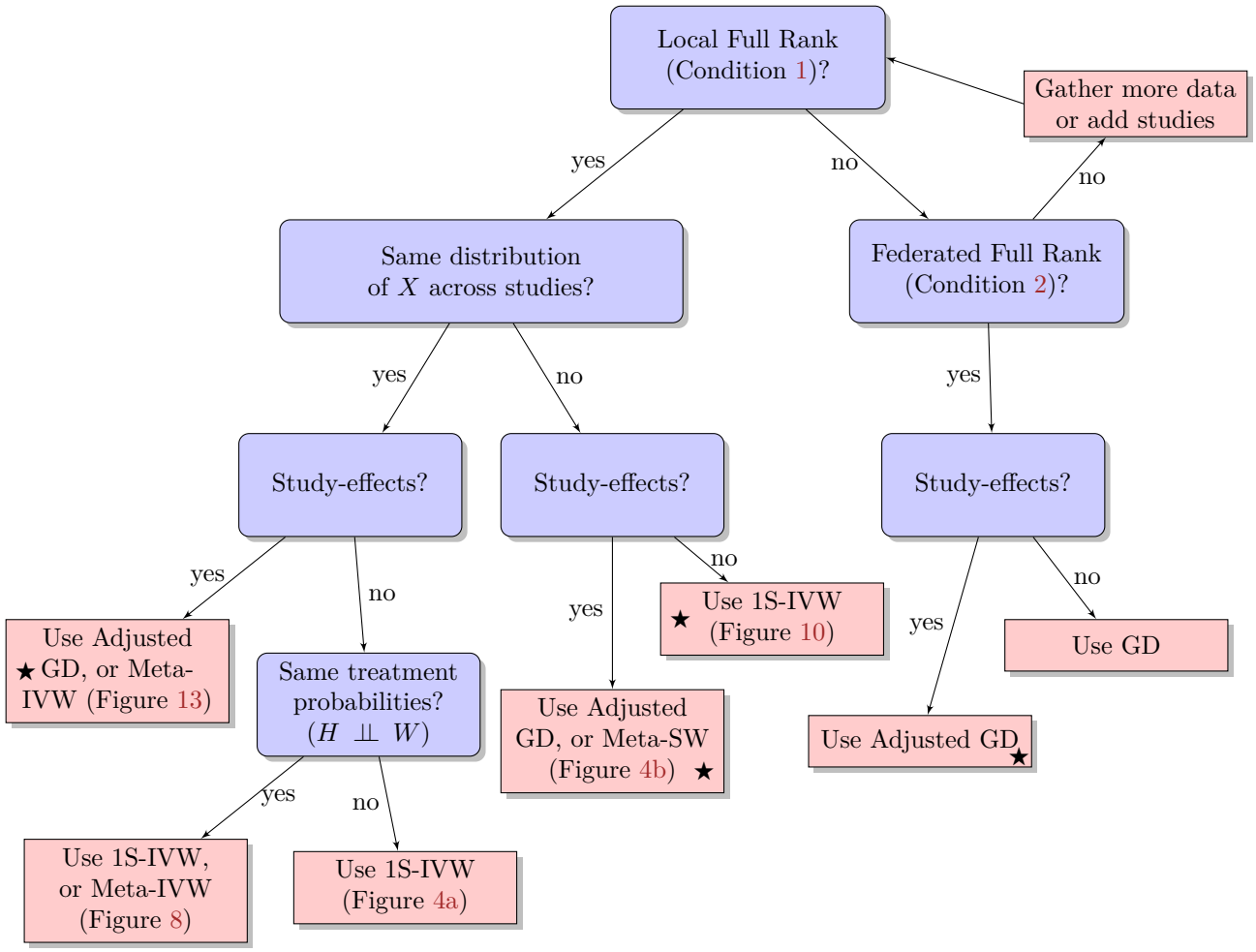


Figure 6: Decision Diagram for Practitionners. The sign ★ denotes scenarios where the DM estimator is biased.

Our results yield clear guidelines for practitioners to select the most suitable estimator for different scenarios, which we present as a decision diagram in Figure 6. Note that each time we mention the use of a meta estimator in second position, it means that this meta estimator yields a valid unbiased estimator, but with a higher variance than the estimator in first position.

It is also worth mentioning that the Difference-in-Means estimator on the pooled individual data is biased whenever H acts as a confounder between W and Y, which happens when studies have distinct treatment probabilities and study effects. The DM estimator is also biased in the graphical model in Figure 2, although H is not technically a confounder in this setting.

B PROOFS

B.1 Weighting methods

B.1.1 Probability of treatment in pooled dataset

We consider that a study is included in the federated study if its sample size is strictly larger than zero, i.e. $\forall k, n_k > 0$. Since n_k is a binomial random variable of parameters n and ρ_k, we have $\mathbb{E}(n_k) = n\rho_k$ which yields $\rho_k = \mathbb{E}(\frac{n_k}{n})$.

B.1.2 Meta-IVW has minimum variance among unbiased aggregation-based estimators

Proof of Proposition 1: Let $\hat{\tau}_k \sim \mathcal{N}(\tau, \mathbb{V}(\hat{\tau}_k))$ and K independent studies. We denote as $\hat{\tau} = \frac{\sum_{k=1}^K w_k \hat{\tau}_k}{\sum_{k=1}^K w_k}$ a w-weighted average of the local estimators of τ. We have:

$$\begin{aligned}
\mathbb{V}(\hat{\tau}) &= \mathbb{V}\left(\frac{\sum_{k=1}^K w_k \hat{\tau}_k}{\sum_{k=1}^K w_k}\right) \\
&= \frac{1}{\left(\sum_{k=1}^K w_k\right)^2} \sum_{k=1}^K w_k^2 \mathbb{V}\left(\hat{\tau}_k\right) \\
&= \sum_{k=1}^K \left(\frac{w_k}{\sum_{k=1}^K w_k}\right)^2 \mathbb{V}\left(\hat{\tau}_k\right) \\
&= \sum_{k=1}^K u_k^2 \mathbb{V}\left(\hat{\tau}_k\right)
\end{aligned}$$

with $u_k = \frac{w_k}{\sum_{k=1}^K w_k}$ and $\sum_{k=1}^K u_k = 1$

We want to minimize $\mathbb{V}(\hat{\tau})$ under the constraint $\sum_{k=1}^K u_k = 1$. We can use the Lagrange multiplier method to find the minimum of $\mathbb{V}(\hat{\tau})$ under this constraint. We define the Lagrangian function:

$$\mathcal{L}(u_1, \ldots, u_K, \lambda) = \sum_{k=1}^K u_k^2 \mathbb{V}\left(\hat{\tau}_k\right) + \lambda\left(1 - \sum_{k=1}^K u_k\right)$$

Then we have:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial u_k} &= 2u_k \mathbb{V}\left(\hat{\tau}_k\right) - \lambda \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= 1 - \sum_{k=1}^K u_k
\end{aligned}$$

To cancel out the two derivatives above, we get that $\forall k \in [1; K], u_k = \frac{\lambda}{2\mathbb{V}(\hat{\tau}_k)}$. Then starting back from the constraint we have:

$$\begin{cases} \sum_{k=1}^K u_k = 1 \\ u_k = \frac{\lambda}{2\mathbb{V}(\hat{\tau}_k)} \end{cases} \Rightarrow \begin{cases} u_k = \frac{\mathbb{V}(\hat{\tau}_k)^{-1}}{\sum_{k=1}^K \mathbb{V}(\hat{\tau}_k)^{-1}} \end{cases}$$

Injecting this result in $\mathbb{V}(\hat{\tau})$, we get:

$$\begin{aligned}\mathbb{V}(\hat{\tau}) &= \sum_{k=1}^{K} u_k^2 \mathbb{V}\left(\hat{\tau}_k\right) \\ &= \sum_{k=1}^{K} \left(\frac{\mathbb{V}\left(\hat{\tau}_k\right)^{-1}}{\sum_{k=1}^{K} \mathbb{V}\left(\hat{\tau}_k\right)^{-1}} \right)^2 \mathbb{V}\left(\hat{\tau}_k\right) \\ &= \frac{1}{\sum_{k=1}^{K} \mathbb{V}\left(\hat{\tau}_k\right)^{-1}}\end{aligned}$$

Finally, we get that $\forall k \in [1; K], u_k = \frac{\mathbb{V}(\hat{\tau}_k)^{-1}}{\sum_{k=1}^{K} \mathbb{V}(\hat{\tau}_k)^{-1}}$.

Therefore, $\hat{\tau}_{\text{Meta-IVW}}$ is the minimum-variance unbiased estimator of τ among the class of aggregation-based estimators.

B.2 Homogeneous setting

We prove in this section the results in Section 3, assuming Condition 1 and the graphical model in Figure 1.

B.2.1 Properties of federated outcome estimators

Proof of Theorem 1:

$$\begin{aligned}\hat{\theta}^{(w)}_{\text{IVW}} &= \frac{\sum_{k=1}^{K} \mathbb{V}(\hat{\theta}^{(w)}_k)^{-1} \hat{\theta}^{(w)}_k}{\sum_{k=1}^{K} \mathbb{V}(\hat{\theta}^{(w)}_k)^{-1}} \\ &= \frac{\sum_{k=1}^{K} \left(\frac{1}{\sigma^2} {X'}^{(w)\top}_k {X'}^{(w)}_k \right) \left({X'}^{(w)\top}_k {X'}^{(w)}_k \right)^{-1} {X'}^{(w)\top}_k y^{(w)}_k}{\sum_{k=1}^{K} \left(\frac{1}{\sigma^2} {X'}^{(w)\top}_k {X'}^{(w)}_k \right)} \\ &= \frac{\sum_{k=1}^{K} {X'}^{(w)\top}_k y^{(w)}_k}{\sum_{k=1}^{K} {X'}^{(w)\top}_k {X'}^{(w)}_k} \\ &= \frac{{X'}^{(w)\top}_{\text{pool}} y^{(w)}_{\text{pool}}}{{X'}^{(w)\top}_{\text{pool}} {X'}^{(w)}_{\text{pool}}} && \text{by sum of matrix product} \\ &= \hat{\theta}^{(w)}_{\text{pool}}\end{aligned}$$

B.2.2 Bias of the outcome model estimators

Unbiasedness of $\hat{\theta}_{\text{pool}}$:

$$\begin{aligned}\mathbb{E}\left(\hat{\theta}^{(w)}_{\text{pool}}\right) &= \mathbb{E}\left(\left({X'}^{(w)\top} {X'}^{(w)} \right)^{-1} {X'}^{(w)\top} y \right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\left({X'}^{(w)\top} {X'}^{(w)} \right)^{-1} {X'}^{(w)\top} y \mid X'^{(w)} \right) \right) \\ &= \mathbb{E}\left(\left({X'}^{(w)\top} {X'}^{(w)} \right)^{-1} {X'}^{(w)\top} \mathbb{E}\left({X'}^{(w)} \theta + \varepsilon \mid X'^{(w)} \right) \right) \\ &= \mathbb{E}\left(\left({X'}^{(w)\top} {X'}^{(w)} \right)^{-1} {X'}^{(w)\top} {X'}^{(w)} \theta + 0 \right) \\ &= \theta\end{aligned}$$

Unbiasedness of $\hat{\theta}_{\text{GD}}$: under convergence of Algorithm D, $\hat{\theta}^{(w)}_{\text{GD}} = \hat{\theta}^{(w)}_{\text{pool}}$ which implies $\mathbb{E}\left(\hat{\theta}^{(w)}_{\text{GD}}\right) = \mathbb{E}\left(\hat{\theta}^{(w)}_{\text{pool}}\right) = \theta$.

Unbiasedness of $\hat{\theta}_{\text{1S-IVW}}$:

$$\begin{aligned}\mathbb{E}\left(\hat{\theta}^{(w)}_{\text{IVW}}\right) &= \mathbb{E}\left(\hat{\theta}^{(w)}_{\text{pool}}\right) && \text{using Theorem 1}\\ &= \theta\end{aligned}$$

Unbiasedness of $\hat{\theta}_{\text{1S-SW}}$:
We condition on the realization of H to account for the variability in the $\{n_k^{(w)}\}_k$ terms.

$$\begin{aligned}\mathbb{E}\left(\mathbb{E}\left(\hat{\theta}^{(w)}_{\text{1S-SW}} \mid H\right)\right) &= \mathbb{E}\left(\mathbb{E}\Big(\sum_{k=1}^{K}\frac{n_k^{(w)}}{n^{(w)}}\hat{\theta}_k^{(w)} \mid H\Big)\right)\\ &= \mathbb{E}\left(\sum_{k=1}^{K}\frac{n_k^{(w)}}{n^{(w)}}\mathbb{E}\left(\hat{\theta}_k^{(w)} \mid H\right)\right)\\ &= \sum_{k=1}^{K}\frac{\mathbb{E}(n_k^{(w)})}{n^{(w)}}\theta\\ &= \theta\end{aligned}$$

B.2.3 (Non-asymptotic) variance comparison of the outcome model parameters

Variance of the local outcome model OLS estimators:

$$\begin{aligned}\mathbb{V}\left(\hat{\theta}_k^{(w)}\right) &= \mathbb{E}\left(\mathbb{V}\left(\left(\hat{\theta}_k^{(w)} \mid {X'}_k^{(w)}\right)\right)\right) + \mathbb{V}\left(\mathbb{E}\left(\hat{\theta}_k^{(w)} \mid {X'}_k^{(w)}\right)\right)\\ &= \mathbb{E}\left(\mathbb{V}\left(\left({X'}_k^{(w)\top}{X'}_k^{(w)}\right)^{-1}{X'}_k^{(w)\top}y_k^{(w)} \mid {X'}_k^{(w)}\right)\right) + 0\\ &= \mathbb{E}\left(\left({X'}_k^{(w)\top}{X'}_k^{(w)}\right)^{-1}{X'}_k^{(w)\top}\mathbb{V}\left(y_k^{(w)} \mid {X'}_k^{(w)}\right){X'}_k^{(w)}\left({X'}_k^{(w)\top}{X'}_k^{(w)}\right)^{-1}\right)\\ &= \mathbb{E}\left(\left({X'}_k^{(w)\top}{X'}_k^{(w)}\right)^{-1}{X'}_k^{(w)\top}\sigma^2 I_p {X'}_k^{(w)}\left({X'}_k^{(w)\top}{X'}_k^{(w)}\right)^{-1}\right)\\ &= \sigma^2\mathbb{E}\left(\left({X'}_k^{(w)\top}{X'}_k^{(w)}\right)^{-1}\right)\end{aligned}$$

Similarly we get $\mathbb{V}\left(\hat{\theta}^{(w)}_{\text{IVW}}\right) = \mathbb{V}\left(\hat{\theta}^{(w)}_{\text{pool}}\right) = \mathbb{V}\left(\hat{\theta}^{(w)}_{\text{GD}}\right) = \sigma^2\mathbb{E}\left(\left({X'}^{(w)\top}{X'}^{(w)}\right)^{-1}\right)$.

The variance of the One-Shot SW outcome model parameters estimator is obtained using the law of total variance over H:

$$\begin{aligned}\mathbb{V}\left(\hat{\theta}^{(w)}_{\text{SW}}\right) &= \mathbb{E}\left(\mathbb{V}(\hat{\theta}^{(w)}_{\text{SW}} \mid H)\right) + \mathbb{V}\left(\mathbb{E}(\hat{\theta}^{(w)}_{\text{SW}} \mid H)\right)\\ &= \sum_{k=1}^{K}\mathbb{E}\left(\frac{n_k^2}{n^2}\right)\mathbb{V}\left(\hat{\theta}_k^{(w)}\right) + 0\\ &= \sigma^2\sum_{k=1}^{K}\frac{\mathbb{E}(n_k^2)}{n^2}\mathbb{E}\left({X'}_k^{(w)\top}{X'}_k^{(w)^{-1}}\right)\end{aligned}$$

by independence of the studies for the last equality.

We now compare the variances above. First notice that:

$$\mathbb{E}\left(\left({X'}^{(w)\top}_{\text{pool}}X^{(w)}_{\text{pool}}\right)^{-1}\right) = \frac{1}{n}\mathbb{E}\left(\frac{1}{\frac{1}{n}{X'}^{(w)\top}_{\text{pool}}X^{(w)}_{\text{pool}}}\right) = \frac{1}{n}\mathbb{E}\left(\frac{1}{\sum_{k=1}^{K}\frac{\mathbb{E}(n_k)}{n}\left(\frac{1}{\mathbb{E}(n_k)}{X'}_k^{(w)\top}X_k^{(w)}\right)}\right)$$

By applying Jensen's inequality on the inverse function over the space of semi positive definite matrices and with weights summing to 1 $\{\mathbb{E}(n_k)/n\}_{k\in[\![1,K]\!]}$:

$$\begin{aligned}
\mathbb{E}\left(\left(X'^{(w)\top}_{\text{pool}} X^{(w)}_{\text{pool}}\right)^{-1}\right) &\preceq \frac{1}{n}\sum_{k=1}^{K}\frac{\mathbb{E}(n_k)}{n}\mathbb{E}\left(\left(\frac{1}{\mathbb{E}(n_k)}X'^{(w)\top}_{k} X^{(w)}_{k}\right)^{-1}\right)\\
&\preceq \sum_{k=1}^{K}\frac{\mathbb{E}(n_k)^2}{n^2}\mathbb{E}\left(\left(X'^{(w)\top}_{k} X'^{(w)}_{k}\right)^{-1}\right)\\
&\preceq \sum_{k=1}^{K}\frac{\mathbb{E}(n_k^2)}{n^2}\mathbb{E}\left(\left(X'^{(w)\top}_{k} X'^{(w)}_{k}\right)^{-1}\right) \qquad \text{as } \mathbb{E}(n_k)^2 \leq \mathbb{E}(n_k^2)
\end{aligned}$$

which leads to the conclusion of Proposition 2: $\mathbb{V}(\hat{\theta}^{(w)}_{\text{pool}}) = \mathbb{V}(\hat{\theta}^{(w)}_{\text{GD}}) = \mathbb{V}(\hat{\theta}^{(w)}_{\text{IVW}}) \preceq \mathbb{V}(\hat{\theta}^{(w)}_{\text{SW}})$.

B.2.4 Asymptotic variances of the outcome model parameters

Asymptotically, we have

$$\begin{aligned}
X'^{(w)\top}_{k} X'^{(w)}_{k} &= \frac{1}{n^{(w)}_k}\sum_{i=1}^{n^{(w)}_k} X'^{(w)\top}_{k,i} X'^{(w)}_{k,i}\\
&\xrightarrow{n^{(w)}_k\to\infty} A_k\\
\left(X'^{(w)\top}_{k} X'^{(w)}_{k}\right)^{-1} &\to A_k^{-1} \quad \text{by continuous mapping}
\end{aligned}$$

where $A_k = \begin{pmatrix} 1 & \mu_{k,1} & \dots & \mu_{k,p} \\ \mu_{k,1} & & & \\ \vdots & & \Sigma_k & \\ \mu_{k,p} & & & \end{pmatrix}$ and $\mu_{k,j}$ is the mean of the j-th covariate and $\Sigma_k = \mathbb{E}\left(X^{(w)\top}_{k} X^{(w)}_{k}\right)$ is the covariate matrix in study k.

Therefore we get the asymptotic variance of the local outcome model parameters:

$$\begin{aligned}
\mathbb{V}^{\infty}\left(\hat{\theta}^{(w)}_{k}\right) &= \frac{\sigma^2}{n^{(w)}_k}A_k^{-1}\\
\mathbb{V}^{\infty}\left(\hat{\beta}^{(w)}_{k}\right) &= \frac{\sigma^2}{n^{(w)}_k}\Sigma_k^{-1}
\end{aligned}$$

Under the graphical model in Figure 1, $\Sigma_k = \Sigma$ and $A_k = A$ which yields:

$$
\begin{aligned}
\text{Local OLS} \quad & \mathbb{V}^{\infty}\left(\hat{\theta}_k^{(w)}\right) = \sigma^2\left(X_k'^{(w)^\top} X_k'^{(w)}\right)^{-1} = \frac{\sigma^2}{n_k^{(w)}} A^{-1} \\
\text{Pool OLS} \quad & \mathbb{V}^{\infty}\left(\hat{\theta}_{\text{pool}}^{(w)}\right) = \sigma^2\left(X'^{(w)^\top} X'^{(w)}\right)^{-1} = \frac{\sigma^2}{n^{(w)}} A^{-1} \\
\text{GD} \quad & \mathbb{V}^{\infty}\left(\hat{\theta}_{\text{GD}}^{(w)}\right) = \sigma^2\left(X'^{(w)^\top} X'^{(w)}\right)^{-1} = \frac{\sigma^2}{n^{(w)}} A^{-1} \\
\text{1S-SW} \quad & \mathbb{V}^{\infty}\left(\hat{\theta}_{\text{SW}}^{(w)}\right) = \mathbb{E}\left(\mathbb{V}^{\infty}\left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \hat{\theta}_k^{(w)} \mid H\right)\right) + 0 \\
& = \mathbb{E}\left(\sum_{k=1}^{K} \frac{{n_k^{(w)}}^2}{{n^{(w)}}^2} \mathbb{V}^{\infty}\left(\hat{\theta}_k^{(w)} \mid H\right)\right) \\
& = \mathbb{E}\left(\sum_{k=1}^{K} \left(\frac{n_k^{(w)}}{n^{(w)}}\right)^2 \frac{\sigma^2}{n_k^{(w)}} A^{-1}\right) \\
& = \frac{\sigma^2}{{n^{(w)}}^2} A^{-1} \sum_{k=1}^{K} \mathbb{E}(n_k^{(w)}) \\
& = \frac{\sigma^2}{n^{(w)}} A^{-1} \\
\text{1S-IVW} \quad & \mathbb{V}^{\infty}\left(\hat{\theta}_{\text{IVW}}^{(w)}\right) = \mathbb{V}^{\infty}\left(\frac{\sum_{k=1}^{K}\left(\mathbb{V}^{\infty}(\hat{\theta}_k^{(w)})^{-1}\hat{\theta}_k^{(w)}\right)}{\sum_{k=1}^{K}\mathbb{V}^{\infty}(\hat{\theta}_k^{(w)})^{-1}}\right) \\
& = \mathbb{V}\left(\frac{\sum_{k=1}^{K} \frac{1}{\sigma^2} X_k'^{(w)^\top} X_k'^{(w)} \times \left(X_k'^{(w)^\top} X_k'^{(w)}\right)^{-1} X_k'^{(w)^\top} y_k^{(w)}}{\sum_{k=1}^{K} \frac{1}{\sigma^2} X_k'^{(w)^\top} X_k'^{(w)}}\right) \\
& = \mathbb{V}\left(\frac{\sum_{k=1}^{K} X_k'^{(w)^\top} y_k^{(w)}}{\sum_{k=1}^{K} X_k'^{(w)^\top} X_k'^{(w)}}\right) \\
& = \frac{\sum_{k=1}^{K} X_k'^{(w)^\top} \mathbb{V}(y_k^{(w)}) X_k'^{(w)}}{\left(\sum_{k=1}^{K} X_k'^{(w)^\top} X_k'^{(w)}\right)^2} \\
& = \frac{\sigma^2}{\sum_{k=1}^{K} X_k'^{(w)^\top} X_k'^{(w)}} \\
& = \frac{\sigma^2}{n^{(w)}} A^{-1}
\end{aligned}
$$

So under this model, for $w \in \{0,1\}$, $\mathbb{V}^{\infty}\left(\hat{\theta}_{\text{GD}}^{(w)}\right) = \mathbb{V}^{\infty}\left(\hat{\theta}_{\text{SW}}^{(w)}\right) = \mathbb{V}^{\infty}\left(\hat{\theta}_{\text{IVW}}^{(w)}\right) = \frac{\sigma^2}{n^{(w)}} A^{-1}$. These results, along with the associated communication costs, are summarized in Table 4.

B.2.5 Bias of the ATE estimators

We now prove that under the graphical model in Figure 1 all the estimators are unbiased.

First, notice that for any random variable $U \in \mathbb{R}^{n \times d+1}, w \in \{0,1\}$,

$$
\begin{aligned}
\mathbb{E}(U\hat{\theta}^{(w)}) &= \mathbb{E}\left(U(X'^{(w)^\top} X'^{(w)})^{-1} X'^{(w)} Y^{(w)}\right) \\
&= \mathbb{E}\left(U(X'^{(w)^\top} X'^{(w)})^{-1} X'^{(w)} (X'^{(w)}\theta^{(w)} + \varepsilon)\right) \\
&= \mathbb{E}(U)\theta^{(w)}
\end{aligned}
$$

Estimator	Notation	Condition	$\mathbb{V}^\infty$	Com. rounds	Com. cost
Local	$\hat{\theta}_k^{(w)}$	Condition 1	$\frac{\sigma^2}{n_k^{(w)}} A^{-1}$	0	0
One-Shot SW	$\hat{\theta}_{\text{1S-SW}}^{(w)}$ (Eq. 9)	Condition 1	$\frac{\sigma^2}{n^{(w)}} A^{-1}$	1	$O(d+1)$
One-Shot IVW	$\hat{\theta}_{\text{1S-IVW}}^{(w)}$ (Eq. 10)	Condition 1	$\frac{\sigma^2}{n^{(w)}} A^{-1}$	1	$O(d^2)$
GD-Federated	$\hat{\theta}_{\text{GD}}^{(w)}$ (Alg. D)	Condition 2	$\frac{\sigma^2}{n^{(w)}} A^{-1}$	T	$O(Td)$
Pool	$\hat{\theta}_{\text{pool}}^{(w)}$	Condition 2	$\frac{\sigma^2}{n^{(w)}} A^{-1}$	—	—

Table 4: Properties of the (unbiased) estimators of the outcome model parameters in the homogeneous setting: asymptotic variance, number of communication rounds and total communication cost (in number of floats per study).

Then,

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_k) &= \mathbb{E}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\left(X'_{k,i}\hat{\theta}_k^{(1)} - X'_{k,i}\hat{\theta}_k^{(0)}\right)\right) && \text{Defined in (6)} \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\mathbb{E}\left(X'_{k,i}\hat{\theta}_k^{(1)}\right) - \mathbb{E}\left(X'_{k,i}\hat{\theta}_k^{(0)}\right)\right) \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\mathbb{E}\left(X'_{k,i}\right)\left(\theta^{(1)} - \theta^{(0)}\right)\right) \\
&= \mathbb{E}\left(X'_i\right)\left(\theta^{(1)} - \theta^{(0)}\right) = \tau && \text{Defined in (3)}
\end{aligned}$$

Similarly, conditioning on H to account for the variability in the n_k random (binomial) terms:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{\text{pool}} \mid H) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X'_i\hat{\theta}_{\text{pool}}^{(1)} - X'_i\hat{\theta}_{\text{pool}}\right) \mid H\right) && \text{Defined in (4)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left(X'_i\hat{\theta}_{\text{pool}}^{(1)} \mid H\right) - \mathbb{E}\left(X'_i\hat{\theta}_{\text{pool}}^{(0)} \mid H\right)\right) \\
&= \frac{1}{n}\sum_{k=1}^{K} n_k \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbb{E}(X'_{k,i}\hat{\theta}_{\text{1S-IVW}}^{(1)} \mid H) - \mathbb{E}(X'_{k,i}\hat{\theta}_{\text{1S-IVW}}^{(0)} \mid H)) = \mathbb{E}(\hat{\tau}_{\text{1S-IVW}} \mid H) && \text{Def. 7 + Th. 1} \\
&= \frac{1}{n}\sum_{k=1}^{K} n_k \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbb{E}(X'_{k,i}\hat{\theta}_{\text{GD}}^{(1)} \mid H) - \mathbb{E}(X'_{k,i}\hat{\theta}_{\text{GD}}^{(0)} \mid H)) = \mathbb{E}(\hat{\tau}_{\text{GD}} \mid H) && \text{Definition 8} \\
&= \mathbb{E}\left(X'_i \mid H\right)\left(\theta^{(1)} - \theta^{(0)}\right) = \tau
\end{aligned}$$

And,

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{\text{1S-SW}} \mid H) &= \frac{1}{n}\sum_{k=1}^{K} n_k \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbb{E}(X'_{k,i}\hat{\theta}_{\text{1S-SW}}^{(1)}) - \mathbb{E}(X'_{k,i}\hat{\theta}_{\text{1S-SW}}^{(0)})) && \text{Definition 6} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left(X'_{k,i}\sum_{l=1}^{K}\frac{n_l}{n}\hat{\theta}_l^{(1)}\right) - \mathbb{E}\left(X'_{k,i}\sum_{l=1}^{K}\frac{n_l}{n}\hat{\theta}_l^{(0)}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{K}\frac{n_l}{n}\mathbb{E}\left(X'_{k,i}\hat{\theta}_l^{(1)} \mid H\right) - \sum_{l=1}^{K}\frac{n_l}{n}\mathbb{E}\left(X'_{k,i}\hat{\theta}_l^{(0)} \mid H\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=1}^{K}\frac{n_l}{n}\mathbb{E}\left(X'_{k,i} \mid H\right)\left(\theta^{(1)} - \theta^{(0)}\right) \mid H\right) \\
&= \mathbb{E}\left(X'_i \mid H\right)\left(\theta^{(1)} - \theta^{(0)}\right) = \tau
\end{aligned}$$

Noticing that none of the expectations above depend on the n_k and that $X \perp\!\!\!\perp H$, we can remove the conditioning over H, so $\mathbb{E}(\hat{\tau}_{\text{pool}}) = \mathbb{E}(\hat{\tau}_{\text{GD}}) = \mathbb{E}(\hat{\tau}_{\text{1S-IVW}}) = \mathbb{E}(\hat{\tau}_{\text{1S-SW}}) = \tau$.

B.2.6 Asymptotic variances of the ATE estimators

We recall that $\rho_k = \mathbb{P}(H_i = k) = \mathbb{E}\left(\frac{n_k}{n}\right)$ and that $p = \mathbb{P}(W_i = 1) = \sum_{k=1}^K \mathbb{P}(H_i = k)\mathbb{P}(W_i = 1|H_i = k) = \sum_{k=1}^K \rho_k p_k$.

Proof of Section 3.1.2 : we prove that the SW ($\omega_k^{\text{SW}} = \frac{n_k}{\sum_{k=1}^K n_k}$) and IVW ($\omega_k^{\text{IVW}} = \frac{\mathbb{V}^\infty(\hat{\tau}_k^{\text{fed}})^{-1}}{\sum_{k=1}^K \mathbb{V}^\infty(\hat{\tau}_k^{\text{fed}})^{-1}}$) weights are asymptotically equivalent for federated local estimators ($\hat{\tau}_k^{\text{1S-SW}}$, $\hat{\tau}_k^{\text{1S-IVW}}$ both defined in Eq. 11, and $\hat{\tau}_k^{\text{GD}}$ defined in Eq. 14). We denote by $\hat{\tau}_k^{\text{fed}}$ any of these estimators.

Then:

$$\begin{aligned}
\omega_k^{\text{IVW}} &= \frac{\left(\frac{\sigma^2}{n_k}\left(\frac{n}{n^{(1)}} + \frac{n}{n^{(0)}}\right) + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2\right)^{-1}}{\sum_{k=1}^K \left(\frac{\sigma^2}{n_k}\left(\frac{n}{n^{(1)}} + \frac{n}{n^{(0)}}\right) + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2\right)^{-1}} \\
&= \frac{n_k\left(\sigma^2\left(\frac{n}{n^{(1)}} + \frac{n}{n^{(0)}}\right) + \|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2\right)^{-1}}{\sum_{k=1}^K n_k\left(\sigma^2\left(\frac{n}{n^{(1)}} + \frac{n}{n^{(0)}}\right) + \|\beta^{(1)} - \beta^{(0)}\|_\Sigma^2\right)^{-1}} \\
&= \frac{n_k}{\sum_{k=1}^K n_k} \\
&= \omega_k^{\text{SW}}
\end{aligned}$$

Therefore, asymptotically, aggregating the local federated estimates of the ATE with SW or IVW is the same.

Asymptotic Variance of local ATE estimator.

First, recall from Appendix B.2.4 that if $\mathbb{E}(X^\top \varepsilon \mid H) = 0$ and Condition 1 holds, the OLS estimator $\hat{\theta}_k^{(w)}$ is consistent and asymptotically normal:

$$\begin{aligned}
\hat{\theta}_k^{(w)} &\xrightarrow{P} \theta_k^{(w)} \\
\sqrt{n_k^{(w)}}(\hat{\theta}_k^{(w)} - \theta_k^{(w)}) &\xrightarrow{d} \mathcal{N}(0, \sigma^2 A^{-1})
\end{aligned} \tag{17}$$

which gives:

$$\sqrt{n_k^{(w)}}(\hat{c}_k^{(w)} - c_k^{(w)}) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{18}$$

$$\sqrt{n_k^{(w)}}(\hat{\beta}_k^{(w)} - \beta_k^{(w)}) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \Sigma_k^{-1}) \tag{19}$$

In particular, we have that $\hat{c}_k^{(1)}, \hat{c}_k^{(0)}, \hat{\beta}_k^{(1)}, \hat{\beta}_k^{(0)}, \overline{X_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{k,i}$ are all asymptotically independent.

Then we have that:

$$\begin{aligned}
\hat{\tau}_k - \tau &= \hat{c}_k^{(1)} - c^{(1)} - (\hat{c}_k^{(0)} - c^{(0)}) + \overline{X_k}(\hat{\beta}_k^{(1)} - \hat{\beta}^{(0)}) - \mathbb{E}(X_k)(\beta_k^{(1)} - \beta^{(0)}) \\
&= \hat{c}_k^{(1)} - c^{(1)} - (\hat{c}_k^{(0)} - c^{(0)}) + \overline{X_k}\left((\hat{\beta}_k^{(1)} - \beta^{(1)}) - (\hat{\beta}_k^{(0)} - \beta^{(0)})\right) \\
&\quad - \mathbb{E}(X_k)(\beta_k^{(1)} - \beta^{(0)}) - \overline{X_k}(\beta_k^{(1)} - \beta^{(0)}) \\
&= \underbrace{\hat{c}_k^{(1)} - c^{(1)}}_{A_1} - \underbrace{(\hat{c}_k^{(0)} - c^{(0)})}_{A_0} + \underbrace{\overline{X_k}\left((\hat{\beta}_k^{(1)} - \beta^{(1)}) - (\hat{\beta}_k^{(0)} - \beta^{(0)})\right)}_{B} \\
&\quad + \underbrace{(\overline{X_k} - \mathbb{E}(X_k))(\beta^{(1)} - \beta^{(0)})}_{C}
\end{aligned} \tag{20}$$

- For $w \in \{0,1\}$, $A_w \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2}{n_k^{(w)}})$ from Equation (18)
- Let $M > 0$ be a real number. Then,

$$\mathbb{P}\left(|n_k B| > M\right) = \mathbb{P}\Big(|\underbrace{\sqrt{n_k}\left(\overline{X}_k - \mathbb{E}(X_k)\right)}_{\xrightarrow{d}\mathcal{N}(0,\Sigma_k)} \underbrace{\sqrt{n_k}(\hat{\beta}_k^{(1)} - \beta_k^{(1)})}_{\xrightarrow{d}\mathcal{N}(0,\frac{\sigma^2}{n_k^{(1)}})} - \underbrace{\sqrt{n_k}(\hat{\beta}_k^{(0)} + \beta^{(0)})}_{\xrightarrow{d}\mathcal{N}(0,\frac{\sigma^2}{n_k^{(0)}})}| > M\Big) \xrightarrow{n_k \to \infty} 0$$

  so that $\mathbb{P}(\left|\frac{B}{1/n_k} - 0\right| > M) < \varepsilon$, meaning that $B = \mathcal{O}_P(1/n_k)$ by definition.

- C: from the central limit theorem, we have that $\sqrt{n_k}\left(\overline{X}_k - \mathbb{E}(X_k)\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_k)$ and $\beta_k^{(1)} - \beta_k^{(0)}$ is a constant vector. However, for any p-multivariate $\sqrt{n}Z \sim \mathcal{N}(0, \Sigma_k)$ random variable and D a constant vector of size $p \times 1$, we have $\sqrt{n}ZD \sim \mathcal{N}(0, D^\top \Sigma D)$.
  Therefore, $\sqrt{n_k}C \xrightarrow{d} \mathcal{N}(0, \|\beta^{(1)} - \beta^{(0)}|^2_{\Sigma_k})$.

Finally, we have that:

$$\begin{aligned}
\mathbb{V}^\infty(\hat{\tau}_k) &= \frac{1}{n_k}\mathbb{V}^\infty(\sqrt{n_k}(\hat{\tau}_k - \tau)) \\
&= \frac{1}{n_k}\left(\mathbb{V}^\infty(\sqrt{n_k}A_1) + \mathbb{V}^\infty(\sqrt{n_k}A_0) + \mathbb{V}^\infty(\sqrt{n_k}B) + \mathbb{V}^\infty(\sqrt{n_k}C)\right) \\
&= \frac{1}{n_k}\left(\frac{n_k}{n_k^{(1)}}\sigma^2 + \frac{n_k}{n_k^{(0)}}\sigma^2 + 0 + \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma_k}\right) \\
&= \sigma^2\left(\frac{1}{n_k^{(1)}} + \frac{1}{n_k^{(0)}}\right) + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma_k} \\
&= \frac{\sigma^2}{n_k}\left(\frac{1}{p_k} + \frac{1}{1-p_k}\right) + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma_k}
\end{aligned} \tag{21}$$

with $p_k = \mathbb{P}(W_i = 1 | H_i = k)$.

Using Central Limit Theorem:

$$\boxed{\sqrt{n_k}(\hat{\tau}_k - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{p_k(1-p_k)} + \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma_k}\right)} \tag{22}$$

Proof of Table 1:

From Equation (22) we have that in a Bernoulli trial (and denoting $[H = k] = \{H_i = k\}_{i=1}^n$):

$$\textbf{Local ATE aVar} \quad \mathbb{V}^\infty\left(\hat{\tau}_k | H = k\right) = \frac{\sigma^2}{n_k p_k (1-p_k)} + \frac{1}{n_k}\|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma_k}$$

Let's apply this result to the pooled dataset $\mathcal{Z}$ and considering it a Bernoulli trial with treatment probability p, and denoting $H = \{H_i\}_{i=1}^n$:

$$\mathbb{V}^\infty\left(\hat{\tau}_{\text{pool}} | H\right) = \frac{\sigma^2}{np(1-p)} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}$$

Finally, because H is not associated with the outcome, we have that $\mathbb{V}^\infty\left(\hat{\tau}_{\text{pool}}\right) = \mathbb{V}^\infty\left(\hat{\tau}_{\text{pool}} | H\right)$, which allows us to conclude:

$$\textbf{Pooled ATE aVar} \quad \mathbb{V}^{\infty}\left(\hat{\tau}_{\text{pool}}\right) = \frac{\sigma^2}{np(1-p)} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|_{\Sigma}^2$$

which proves Eq. 5 (even in the non-centered covariates case).

To compute the asymptotic variance of the local federated outcome parameters ATE estimator, we first compute the asymptotic variance of the federated-outcome model parameters estimated individual treatment effect $\hat{\tau}_{k,i}^{\text{fed}} = X'_{k,i}\hat{\theta}_{\text{fed}}^{(1)} - X'_{k,i}\hat{\theta}_{\text{fed}}^{(0)}$. To do this, remark that:

$$\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) &= \mathbb{V}^{\infty}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\tau}_{k,i}^{\text{pool}}\right) \\
&= \mathbb{V}^{\infty}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\tau}_{k,i}^{\text{fed}}\right) && \text{since } \hat{\theta}_{\text{fed}}^{(w)} \to \hat{\theta}_{\text{pool}}^{(w)} \\
&= \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbb{V}^{\infty}(\hat{\tau}_{k,i}^{\text{fed}}) + \frac{1}{n^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{n_k}\sum_{j\neq i}^{n_l}\text{Cov}^{\infty}\left(\hat{\tau}_{k,i}^{\text{fed}}, \hat{\tau}_{l,j}^{\text{fed}}\right)
\end{aligned}$$

Let us show that the $\text{Cov}(\hat{\tau}_{k,i}^{\text{fed}}, \hat{\tau}_{l,j}^{\text{fed}})$ are asymptotically null for all $(k,i) \neq (l,j)$. Denote by $\text{Cov}^{\infty}(a,b)$ the asymptotic covariance of a and b random variables. We further write $\hat{Y}_i^{(w)} = X'_i\hat{\theta}_{\text{fed}}^{(w)}$, and remark that asymptotically, $\hat{Y}_i^{(w)} = X'_i\hat{\theta}_{\text{pool}}^{(w)}$, then:

$$\begin{aligned}
\text{Cov}^{\infty}\left(\hat{\tau}_{k,i}^{\text{fed}}, \hat{\tau}_{l,j}^{\text{fed}}\right) &= \text{Cov}^{\infty}\left(X'_{k,i}\left(\hat{\theta}_{\text{fed}}^{(1)} - \hat{\theta}_{\text{fed}}^{(0)}\right), X'_{l,j}\left(\hat{\theta}_{\text{fed}}^{(1)} - \hat{\theta}_{\text{fed}}^{(0)}\right)\right) \\
&= \text{Cov}^{\infty}\left(\hat{Y}_{k,i}(1) - \hat{Y}_{k,i}(0), \hat{Y}_{l,j}(1) - \hat{Y}_{l,j}(0)\right) \\
&= \text{Cov}^{\infty}\left(\hat{Y}_{k,i}(1), \hat{Y}_{l,j}(1)\right) - \text{Cov}^{\infty}\left(\hat{Y}_{k,i}(1), \hat{Y}_{l,j}(0)\right) \\
&\quad - \text{Cov}^{\infty}\left(\hat{Y}_{k,i}(0), \hat{Y}_{l,j}(1)\right) + \text{Cov}^{\infty}\left(\hat{Y}_{k,i}(0), \hat{Y}_{l,j}(0)\right) \\
\text{Cov}^{\infty}\left(\hat{Y}_{k,i}^{(w_a)}, \hat{Y}_{l,j}^{(w_b)}\right) &= \mathbb{E}\left[\left(\hat{Y}_{k,i}^{(w_a)} - \mathbb{E}(\hat{Y}_{k,i}^{(w_a)})\right)\left(\hat{Y}_{l,j}^{(w_b)} - \mathbb{E}(\hat{Y}_{l,j}^{(w_b)})\right)\right] && \forall w_a, w_b \in \{0,1\} \\
&= \mathbb{E}\left[\left(\hat{Y}_{k,i}^{(w_a)} - Y_{k,i}^{(w_a)}\right)\left(\hat{Y}_{l,j}^{(w_b)} - Y_{l,j}^{(w_b)}\right)\right] && \text{unbiased estimators} \\
&= \mathbb{E}\left[\varepsilon_{k,i}^{(w_a)}\varepsilon_{l,j}^{(w_b)}\right] && \text{residuals} \\
&= \mathbb{E}\left[\varepsilon_{k,i}^{(w_a)}\right]\mathbb{E}\left[\varepsilon_{l,j}^{(w_b)}\right] && \text{independent errors} \\
&= 0 && \text{centered noise}
\end{aligned}$$

So that $\text{Cov}^{\infty}\left(\hat{\tau}_i^{\text{fed}}, \hat{\tau}_j^{\text{fed}} | H_i, H_j\right) = 0$, which yields that $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbb{V}^{\infty}(\hat{\tau}_{k,i}^{\text{fed}} | H_i = k)$. Finally, since the individuals follow the same distribution across studies within the graphical model in Figure 1, the $\hat{\tau}_{k,i}^{\text{fed}}$ are *i.i.d.* across studies, so that their asymptotic variances are equal:

$$\mathbb{V}^{\infty}(\hat{\tau}_{k,i}^{\text{fed}} | H_i = k) = n\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \sigma^2\left(\frac{1}{p(1-p)}\right) + \|\beta^{(1)} - \beta^{(0)}\|_{\Sigma}^2$$

Therefore,

$$\begin{aligned}
\textbf{Federated Local ATE} \quad \mathbb{V}^{\infty}(\hat{\tau}_k^{\text{fed}}|H=k) &= \mathbb{V}^{\infty}(\frac{1}{n_k}\sum_{i=1}^{n_k}\hat{\tau}_{k,i}^{\text{fed}}|H=k) \\
&= \frac{1}{n_k^2}\sum_{i=1}^{n_k}\mathbb{V}^{\infty}(\hat{\tau}_{k,i}^{\text{fed}}|H=k) + \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j\neq i}\underbrace{\text{Cov}^{\infty}\left(\hat{\tau}_{k,i}^{\text{fed}},\hat{\tau}_{k,j}^{\text{fed}}|H=k\right)}_{=0} \\
&= \frac{n}{n_k}\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) \\
&= \frac{\sigma^2}{n_k}\left(\frac{1}{p(1-p)}\right) + \frac{1}{n_k}\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2
\end{aligned}$$

For the Meta estimators we use the law of total variance:

$$\begin{aligned}
\textbf{Meta-SW ATE} \quad \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}}) &= \mathbb{V}^{\infty}(\sum_{k=1}^{K}\frac{n_k}{n}\hat{\tau}_k|H) \\
&= \mathbb{E}\left(\sum_{k=1}^{K}\left(\frac{n_k}{n}\right)^2\mathbb{V}^{\infty}(\hat{\tau}_k|H)\right) + \mathbb{V}^{\infty}\left(\mathbb{E}\left(\sum_{k=1}^{K}\frac{n_k}{n}\hat{\tau}_k|H\right)\right) \\
&= \sum_{k=1}^{K}\mathbb{E}\left[\left(\frac{n_k}{n}\right)^2\left(\frac{\sigma^2}{n_k p_k(1-p_k)} + \frac{1}{n_k}\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2\right)\right] + 0 \\
&= \frac{\sigma^2}{n}\sum_{k=1}^{K}\mathbb{E}\left[\frac{n_k}{n}\right]\frac{1}{p_k(1-p_k)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2 \\
&= \frac{\sigma^2}{n}\sum_{k=1}^{K}\frac{\rho_k}{p_k(1-p_k)} + \frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2
\end{aligned}$$

$$\begin{aligned}
\textbf{Meta-IVW ATE} \quad \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) &= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\frac{\sum_{k=1}^{K}\left(\mathbb{V}^{\infty}(\hat{\tau}_k|H=k)^{-1}\hat{\tau}_k\right)}{\sum_{k=1}^{K}\mathbb{V}^{\infty}(\hat{\tau}_k|H=k)^{-1}}|H\right)\right) + 0 \\
&= \mathbb{E}\left(\frac{\sum_{k=1}^{K}\mathbb{V}^{\infty}(\hat{\tau}_k|H=k)^{-2}\mathbb{V}^{\infty}(\hat{\tau}_k|H=k)}{\left(\sum_{k=1}^{K}\mathbb{V}^{\infty}(\hat{\tau}_k|H=k)^{-1}\right)^2}\right) && \hat{\tau}_k \perp\!\!\!\perp \hat{\tau}_l \\
&= \mathbb{E}\left(\left(\sum_{k=1}^{K}\mathbb{V}^{\infty}(\hat{\tau}_k|H=k)^{-1}\right)^{-1}\right) \\
&= \mathbb{E}\left(\left(\sum_{k=1}^{K}\left(\frac{\sigma^2}{n_k p_k(1-p_k)} + \frac{1}{n_k}\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2\right)^{-1}\right)^{-1}\right) \\
&= \mathbb{E}\left(\left(\sum_{k=1}^{K}\frac{n_k}{\frac{\sigma^2}{p_k(1-p_k)} + \|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2}\right)^{-1}\right) \\
&= \frac{1}{n}\left(\sum_{k=1}^{K}\frac{\rho_k}{\frac{\sigma^2}{p_k(1-p_k)} + \|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2}\right)^{-1} && \text{(LLN)}
\end{aligned}$$

where (LLN) refers to the law of large numbers, stating that given $n_k \sim \text{Binomial}(n, \rho_k)$ and $\mathbb{E}[n_k] = n\rho_k$, for large n, we have $\mathbb{E}\left(\frac{1}{n_k}\right) \approx \frac{1}{\mathbb{E}[n_k]} = \frac{1}{n\rho_k}$.

We denote by $\hat{\tau}^{\text{fed}}_{\text{SW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{\text{fed}}_k$:

$$\begin{aligned}
\textbf{SS Weighted Fed. ATE} \quad \mathbb{V}^{\infty}(\hat{\tau}^{\text{fed}}_{\text{SW}}) &= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{\text{fed}}_k \mid H\right)\right) + 0 \\
&= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_k} \sum_{i=1}^{n_k} \left(X'_{k,i} \hat{\theta}^{(1)}_{\text{fed}} - X'_{k,i} \hat{\theta}^{(0)}_{\text{fed}}\right) \mid H\right)\right) \\
&= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\frac{1}{n} \sum_{i=1}^{n} \left(X'_i \hat{\theta}^{(1)}_{\text{fed}} - X'_i \hat{\theta}^{(0)}_{\text{fed}}\right) \mid H\right)\right) \\
&= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\frac{1}{n} \sum_{i=1}^{n} \left(X'_i \hat{\theta}^{(1)}_{\text{pool}} - X'_i \hat{\theta}^{(0)}_{\text{pool}}\right) \mid H\right)\right) \\
&= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\hat{\tau}_{\text{pool}} \mid H\right)\right) \\
&= \mathbb{V}^{\infty}\left(\hat{\tau}_{\text{pool}}\right) \\
&= \frac{\sigma^2}{n} \frac{1}{p(1-p)} + \frac{1}{n} \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}
\end{aligned}$$

We denote by $\hat{\tau}^{\text{fed}}_{\text{IVW}} = \frac{\sum_{k=1}^{K} \left(\mathbb{V}^{\infty}(\hat{\tau}^{\text{fed}}_k)^{-1} \hat{\tau}^{\text{fed}}_k\right)}{\sum_{k=1}^{K} \mathbb{V}^{\infty}(\hat{\tau}^{\text{fed}}_k)^{-1}}$:

$$\begin{aligned}
\textbf{IV Weighted Fed. ATE} \quad \mathbb{V}^{\infty}(\hat{\tau}_{\text{IVW-agg}}) &= \mathbb{E}\big(\mathbb{V}^{\infty}\big(\frac{\sum_{k=1}^{K} \left(\mathbb{V}^{\infty}(\hat{\tau}^{\text{fed}}_k)^{-1} \hat{\tau}^{\text{fed}}_k\right)}{\sum_{k=1}^{K} \mathbb{V}^{\infty}(\hat{\tau}^{\text{fed}}_k)^{-1}}\big)\big) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\text{SW-agg}}) \qquad \text{from Section 3.1.2} \\
&= \frac{\sigma^2}{n} \frac{1}{p(1-p)} + \frac{1}{n} \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}
\end{aligned}$$

B.2.7 Comparison of asymptotic variances - General Case

Proof of Theorem 2: Under the graphical model in Figure 1:

- $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-SW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-IVW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{GD}})$ (Table 1)
- From Proposition 1 we have: $\mathbb{V}^{\infty}(\hat{\tau}_{\text{meta-IVW}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{SW}})$.
- Proving that $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}})$ is equivalent to proving the following inequality:

$$\begin{aligned}
\frac{1}{\sum_{k=1}^{K} \left(\frac{\sigma^2}{n_k p_k (1-p_k)} + \frac{1}{n_k} \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}\right)^{-1}} &\geq \frac{\sigma^2}{np(1-p)} + \frac{1}{n} \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma} \\
\iff \quad \frac{1}{D} &\geq \frac{\sigma^2}{p(1-p)} + a
\end{aligned}$$

with $D = \sum_{k=1}^{K} \frac{n_k}{n} \frac{x_k}{\sigma^2 + x_k a}$, $x_k = p_k(1-p_k)$ and $a = \|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}$.

$x \mapsto \frac{x}{\sigma^2 + xa}$ is concave. Therefore, by Jensen's inequality $D \leq \frac{\overline{x}}{\sigma^2 + \overline{x}a}$, with $\overline{x} = \sum_{k=1}^{K} \frac{n_k}{n} x_k$. We then have:

$$\frac{1}{D} \geq \frac{\sigma^2 + \overline{x}a}{\overline{x}} = \frac{\sigma^2}{\overline{x}} + a = \frac{\sigma^2}{\sum_{k=1}^{K} \frac{n_k}{n} p_k(1-p_k)} + a$$

Since $p \mapsto p(1-p)$ is also concave, by Jensen's inequality again we have:

$$\sum_{k=1}^{K} \frac{n_k}{n} p_k(1-p_k) \leq \left(\sum_{k=1}^{K} \frac{n_k}{n} p_k\right) \left(\sum_{k=1}^{K} \frac{n_k}{n} (1-p_k)\right) = p(1-p)$$

So that $\frac{1}{D} \geq \frac{\sigma^2}{p(1-p)} + a$, which ends the proof.

In conclusion,

$$\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}})$$

which concludes into $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{GD}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-SW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-IVW}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}})$.
Illustrating examples of this property:

1. $\mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) - \mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) \geq 0$ increases as the treatment probabilities $\{p_k\}_k$ become more distinct. For example, with $K = 2$ studies with balanced datasets ($n_1 = n_2 = n/2$), having $p_1 = 0.99$ and $p_2 = 0.01$ yields $\mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}})$ to be 8 times larger than $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}})$ (where we chose $\sigma^2 = 1$ and $\|\beta^{(1)} - \beta^{(0)}\|_{\Sigma}^2 = 10$).

2. Similarly, $\mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}}) - \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) \geq 0$ grows with the $\{p_k(1-p_k)\}_k$ terms being more and more different from one another. For $K = 2$ and $n_1 = n_2 = n/2$, having $p_1 = 0.99$ and $p_2 = 0.5$ gives a 2.5 larger asymptotic variance of the Meta-SW than that of the Meta-IVW estimator.

B.2.8 Comparison of asymptotic variances - Special Case

Proof that the estimators of the ATE have equal asymptotic variances when one RCT is conducted over K studies: Under this setting, we modify the graphical model in Figure 1 and remove the edge between H and W:

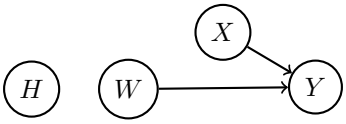


Figure 7: Graphical Model of One RCT Conducted Over K Studies

Under this graphical model (Figure 7), using the variances in Table 1 and as $\forall k, p_k = p$, we have:
First, $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-SW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-IVW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{GD}})$.
Then,

$$\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}}) &= \frac{\sigma^2}{n} \sum_{k=1}^{K} \frac{\mathbb{E}(n_k)}{n} \frac{1}{p_k(1-p_k)} + \frac{1}{n} \left\| \beta^{(1)} - \beta^{(0)} \right\|_{\Sigma}^2 \\
&= \frac{\sigma^2}{n} \frac{1}{p(1-p)} \sum_{k=1}^{K} \frac{\mathbb{E}(n_k)}{n} + \frac{1}{n} \left\| \beta^{(1)} - \beta^{(0)} \right\|_{\Sigma}^2 \\
&= \frac{\sigma^2}{n} \frac{1}{p(1-p)} + \frac{1}{n} \left\| \beta^{(1)} - \beta^{(0)} \right\|_{\Sigma}^2 \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}})
\end{aligned}$$

And

$$\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) &= \frac{1}{n}\left(\sum_{k=1}^{K}\frac{\rho_k}{\frac{\sigma^2}{p_k(1-p_k)}+\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2}\right)^{-1} \\
&= \left(\sum_{k=1}^{K}\left(\frac{\sigma^2}{\mathbb{E}(n_k)}\frac{1}{p_k(1-p_k)}+\frac{1}{\mathbb{E}(n_k)}\left\|\beta^{(1)}-\beta^{(0)}\right\|_{\Sigma}^2\right)^{-1}\right)^{-1} \\
&= \left(\sum_{k=1}^{K}\left(\frac{\sigma^2}{\mathbb{E}(n_k)}\frac{1}{p(1-p)}+\frac{1}{\mathbb{E}(n_k)}\left\|\beta^{(1)}-\beta^{(0)}\right\|_{\Sigma}^2\right)^{-1}\right)^{-1} \\
&= \left(\sum_{k=1}^{K}\left(\frac{1}{\mathbb{E}(n_k)}\left(\frac{\sigma^2}{p(1-p)}+\left\|\beta^{(1)}-\beta^{(0)}\right\|_{\Sigma}^2\right)\right)^{-1}\right)^{-1} \\
&= \frac{\sigma^2}{n}\frac{1}{p(1-p)}+\frac{1}{n}\left\|\beta^{(1)}-\beta^{(0)}\right\|_{\Sigma}^2 \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}})
\end{aligned}$$

which gives $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{GD}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-IVW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-SW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) = \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}})$.

B.3 Heterogeneous Settings

B.3.1 Distributional Shift in Covariates

In this part, we consider the graphical model in Figure 2, and model 2.

Unbiasedness and asymptotic variance of Meta-SW under covariate heterogeneity:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{\text{Meta-SW}}) &= \mathbb{E}_H\left(\mathbb{E}\left(\sum_{k=1}^{K}\frac{n_k}{n}\hat{\tau}_k|H\right)\right) \\
&= \sum_k \mathbb{E}_H\left(\frac{n_k}{n}\right)\mathbb{E}(\hat{\tau}_k) \\
&= \sum_k \rho_k\tau_k && \text{under Condition 1 and model (2)} \\
&= \tau
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}}) &= \mathbb{E}\left(\mathbb{V}^{\infty}\left(\hat{\tau}_{\text{Meta-SW}}\right)\right)+\mathbb{V}^{\infty}\left(\mathbb{E}\left(\hat{\tau}_{\text{Meta-SW}}\right)\right) \\
&= \sum_k \mathbb{E}\left(\left(\frac{n_k}{n}\right)^2\mathbb{V}^{\infty}(\hat{\tau}_k)\right) \\
&= \sum_{k=1}^{K}\frac{\mathbb{E}(n_k)}{n^2}\left(\frac{\sigma^2}{p_k(1-p_k)}+\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma_k}^2\right) \\
&= \frac{1}{n}\sum_{k=1}^{K}\rho_k\frac{\sigma^2}{p_k(1-p_k)}+\frac{1}{n}\sum_{k=1}^{K}\rho_k\left(\beta^{(1)}-\beta^{(0)}\right)^{\top}\Sigma_k\left(\beta^{(1)}-\beta^{(0)}\right) \\
&= \frac{1}{n}\sum_{k=1}^{K}\rho_k\frac{\sigma^2}{p_k(1-p_k)}+\frac{1}{n}\left(\beta^{(1)}-\beta^{(0)}\right)^{\top}\sum_{k=1}^{K}\rho_k\Sigma_k\left(\beta^{(1)}-\beta^{(0)}\right) \\
&= \frac{\sigma^2}{n}\sum_{k=1}^{K}\frac{\rho_k}{p_k(1-p_k)}+\frac{1}{n}\|\beta^{(1)}-\beta^{(0)}\|_{\Sigma}^2
\end{aligned}$$

Unbiasedness and asymptotic variance of GD under covariate heterogeneity: The computation of the GD estimator's bias and asymptotic variance are direct using exactly the same proof as in Appendix B.2.6 since the assumptions over X_{pool} are still met in the distributional shift setting.

Proof of Proposition 2:
We have $\hat{c}_k^{(w)} = \bar{Y}_k^{(w)} - \bar{X}_k^{(w)}\hat{\beta}^{(w)}$ so that:

$$\begin{aligned}
\mathbb{V}(\hat{c}_k^{(w)} \mid X_k^{(w)}) &= \mathbb{V}(\bar{Y}_k^{(w)} - \bar{X}_k^{(w)}\hat{\beta}^{(w)} \mid X_k^{(w)}) \\
&= \mathbb{V}(\bar{Y}_k^{(w)}) + \mathbb{V}(\bar{X}_k^{(w)}\hat{\beta}^{(w)} \mid X_k^{(w)}) - 2\text{Cov}\left(\bar{Y}_k^{(w)} \mid X_k^{(w)}, \bar{X}_k^{(w)}\hat{\beta}^{(w)} \mid X_k^{(w)}\right) \\
&= \frac{\sigma^2}{n_k^{(w)}} + \bar{X_k^{(w)}}^\top \mathbb{V}(\hat{\beta}^{(w)} \mid X_k^{(w)})\bar{X}_k^{(w)} - 2\text{Cov}\left(\bar{Y}_k^{(w)} \mid X_k^{(w)}, \bar{X}_k^{(w)}\hat{\beta}^{(w)} \mid X_k^{(w)}\right)
\end{aligned}$$

Given $X_k^{(w)}$ we have:

$$\begin{aligned}
\text{Cov}\left(\bar{Y}_k^{(w)}, \hat{\beta}_k\right) &= \text{Cov}\left(\frac{1}{n_k^{(w)}}\sum_{i=1}^{n_k^{(w)}} Y_{k,i}, \frac{\sum_{j=1}^{n_k^{(w)}}(X_{k,j}^{(w)} - \bar{X}_k^{(w)})Y_{k,j}}{\sum_{i=1}^{n_k^{(w)}}(X_{k,i} - \bar{X}_k^{(w)})^2}\right) \\
&= \frac{1}{n_k^{(w)}\sum_{i=1}^{n_k^{(w)}}(X_{k,i}^{(w)} - \bar{X}_k^{(w)})^2}\text{Cov}\left(\sum_{i=1}^{n_k^{(w)}} Y_{k,i}, \sum_{j=1}^{n_k^{(w)}}(X_{k,j} - \bar{X}_k^{(w)})Y_j\right) \\
&= \frac{1}{n_k^{(w)}\sum_{i=1}^{n_k^{(w)}}(X_{k,i} - \bar{X}_k^{(w)})^2}\sum_{i=1}^{n_k^{(w)}}\sum_{j=1}^{n_k^{(w)}}(X_{k,j}^{(w)} - \bar{X}_k^{(w)})\text{Cov}\left(Y_{k,i}, Y_{k,j}\right) \\
&= \frac{1}{n_k^{(w)}\sum_{i=1}^{n_k^{(w)}}(X_{k,i}^{(w)} - \bar{X}_k^{(w)})^2}\sum_{j=1}^{n_k^{(w)}}(X_{k,j}^{(w)} - \bar{X}_k^{(w)})\sigma^2 \\
&= 0
\end{aligned}$$

Therefore, $\mathbb{V}(\hat{c}_k^{(w)} \mid X_k^{(w)}) = \sigma^2\left(\frac{1}{n_k^{(w)}} + (\bar{X}_k^{(w)})^\top({X_k^{(w)}}^\top X_k^{(w)})^{-1}\bar{X}_k^{(w)}\right)$, which yields in a random design over the $X_k^{(w)}$:

$$\mathbb{V}^\infty(\hat{c}_k^{(w)}) = \sigma^2\left(\frac{1}{n_k^{(w)}} + \mathbb{E}\left((\bar{X}_k^{(w)})^\top({X_k^{(w)}}^\top X_k^{(w)})^{-1}\bar{X}_k^{(w)}\right)\right)$$

where

$$\begin{aligned}
\mathbb{E}\left((\bar{X}_k^{(w)})^\top({X_k^{(w)}}^\top X_k^{(w)})^{-1}\bar{X}_k^{(w)}\right) &= \mu_k^\top\Sigma_k^{-1}\mu_k + \mathbb{E}\left(\frac{1}{n_k^{(w)}}(\bar{X}_k^{(w)} - \mu_k)^\top\Sigma_k^{-1}(\bar{X}_k^{(w)} - \mu_k)\right) \\
&= \mu_k^\top\Sigma_k^{-1}\mu_k + \frac{1}{n_k^{(w)}}\text{Tr}\left(\Sigma_k^{-1}\sum_{i=1}^{n_k^{(w)}}\mathbb{E}\left((X_{ki}^{(w)} - \mu_k)(X_{ki}^{(w)} - \mu_k)^\top\right)\right) \\
&= \mu_k^\top\Sigma_k^{-1}\mu_k + \frac{1}{n_k^{(w)}}\text{Tr}\left(\Sigma_k^{-1}\sum_{i=1}^{n_k^{(w)}}\Sigma_k\right) \\
&= \mu_k^\top\Sigma_k^{-1}\mu_k + \frac{1}{n_k^{(w)}}\text{Tr}\left(\Sigma_k^{-1}\right)
\end{aligned}$$

so that $\mathbb{V}^\infty(\hat{c}_k^{(w)}) = \sigma^2\left(\frac{1}{n_k^{(w)}} + \mu_k^\top\Sigma_k^{-1}\mu_k + \frac{1}{n_k^{(w)}}\text{Tr}\left(\Sigma_k^{-1}\right)\right)$.

Then, we get $\mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{pool}}) = \sigma^2 \left(\frac{1}{n^{(w)}} + \mu^\top \Sigma^{-1} \mu + \frac{1}{n^{(w)}} \text{Tr}\left(\Sigma^{-1}\right)\right)$ and

$$\begin{aligned}\mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{1S-SW}}) &= \sigma^2 \sum_{k=1}^{K} \left(\frac{n_k^{(w)}}{n^{(w)}}\right)^2 \left(\frac{1}{n_k^{(w)}} + \mu_k^\top \Sigma_k^{-1} \mu_k + \frac{1}{n_k^{(w)}} \text{Tr}\left(\Sigma_k^{-1}\right)\right) \\ &= \sigma^2 \left(\frac{1}{n^{(w)}} + \sum_{k=1}^{K} \left(\frac{n_k^{(w)}}{n^{(w)}}\right)^2 \left(\mu_k^\top \Sigma_k^{-1} \mu_k + \frac{1}{n_k^{(w)}} \text{Tr}\left(\Sigma_k^{-1}\right)\right)\right)\end{aligned}$$

Finally,

$$\begin{aligned}\frac{1}{\sigma^2}\left(\mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{pool}}) - \mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{1S-SW}})\right) &= \mu^\top \Sigma^{-1} \mu - \sum_{k=1}^{K} \left(\frac{n_k^{(w)}}{n^{(w)}}\right)^2 \mu_k^\top \Sigma_k^{-1} \mu_k \\ &= \left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \mu_k^\top\right) \left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \left(\Sigma_k + (\mu_k - \mu)(\mu_k - \mu)^\top\right)\right)^{-1} \\ &\quad \times \left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \mu_k\right) - \sum_{k=1}^{K} \left(\frac{n_k^{(w)}}{n^{(w)}}\right)^2 \mu_k^\top \Sigma_k^{-1} \mu_k\end{aligned}$$

With Jensen's inequality applied to the convex function $f(x) = x^\top \Sigma_k^{-1} x$ for any positive definite matrix Σ_k with positive and summing to 1 weights, we have:

$$\left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \mu_k\right)^\top \left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \Sigma_k\right)^{-1} \left(\sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \mu_k\right) \leq \sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \mu_k^\top \Sigma_k^{-1} \mu_k.$$

Therefore,

$$\mu^\top \Sigma^{-1} \mu - \sum_{k=1}^{K} \left(\frac{n_k^{(w)}}{n^{(w)}}\right)^2 \mu_k^\top \Sigma_k^{-1} \mu_k \leq \sum_{k=1}^{K} \frac{n_k^{(w)}}{n^{(w)}} \left(1 - \frac{n_k^{(w)}}{n^{(w)}}\right) \mu_k^\top \Sigma_k^{-1} \mu_k \leq 0$$

After applying the law of total variance over H, we get $\mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{pool}}) \leq \mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{1S-SW}})$, implying $\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-SW}})$.

In particular, notice that:

- If the studies have equal means ($\forall k, \mu_k = \mu$), then $\mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{pool}}) = \mathbb{V}^{\infty}(\hat{c}^{(w)}_{\text{1S-SW}})$, even if the $\{\Sigma_k\}_k$ are different.

- If a small number of studies have very distinct means from the rest of the studies, the difference in variances between the Pool and 1S-SW estimators will be large.

B.4 Study-Effects

In this part, we consider the graphical model in Figure 3 and model (16).

Proof of meta estimators' unbiasedness and asymptotic variance (Section 4.2):
First, let us compute the intercepts for the treated and control groups in study k under model (16).

First, $a_k^{(w)} = c^{(w)} + h_k$ is the intercept in model (16), so that $a_k^{(w)} = \mathbb{E}(Y_k^{(w)} - X_k \beta^{(w)} - \varepsilon_k^{(w)}) = \mathbb{E}(Y_k^{(w)}) - \mathbb{E}(X_k)\beta^{(w)}$, which yields a **locally estimated** intercept $\hat{a}_k^{(w)} = \overline{Y_k}^{(w)} - \overline{X_k}^{(w)} \hat{\beta}_k^{(w)}$ in study k for treatment group w.

Then, we can compute the Meta estimator with weights ω_k:

$$\begin{aligned}
\hat{\tau}_{\text{Meta}} &= \sum_{k=1}^{K} \omega_k \hat{\tau}_k \\
&= \sum_{k=1}^{K} \omega_k \left(\hat{a}_k^{(1)} - \hat{a}_k^{(0)} + \overline{X_k}(\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) \right) \\
&= \sum_{k=1}^{K} \omega_k \left((a_k^{(1)} + \hat{\epsilon}^{(1)}) - (a_k^{(0)} + \hat{\epsilon}^{(0)}) + \overline{X_k}(\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) \right) \\
&= \sum_{k=1}^{K} \omega_k \left((c^{(1)} + h_k + \hat{\epsilon}^{(1)}) - (c^{(0)} + h_k + \hat{\epsilon}^{(0)}) + \overline{X_k}(\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) \right) \\
&= \sum_{k=1}^{K} \omega_k \left(\hat{c}^{(1)} - \hat{c}^{(0)} + \overline{X_k}(\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) \right)
\end{aligned}$$

where $\hat{c}^{(w)} = c^{(w)} + \hat{\epsilon}^{(w)}$ and $\hat{\epsilon}^{(w)}$ is the estimation bias of the OLS, with expectancy 0.

Therefore,

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{\text{Meta}}) &= \sum_{k=1}^{K} \omega_k \mathbb{E}\left(\hat{a}_k^{(1)} - \hat{a}_k^{(0)} + \overline{X_k}(\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) \right) \\
&= \sum_{k=1}^{K} \omega_k \mathbb{E}\left(\hat{c}^{(1)} - \hat{c}^{(0)} + \overline{X_k}(\hat{\beta}^{(1)} - \hat{\beta}^{(0)}) \right) \\
&= \sum_{k=1}^{K} \omega_k \mathbb{E}(\hat{\tau}_k) \\
&= \sum_{k=1}^{K} \omega_k \tau_k
\end{aligned}$$

which leads the Meta estimators to be unbiased and have the same asymptotic variance as in Table 1.

B.4.1 Unadjusted Federated Estimators

We define the pooled outcome model parameters estimator as $\hat{\theta}_{\text{pool}}^{(w)} = \arg\min_\theta \sum_{i=1}^{n} \left(Y_i^{(w)} - X_i' \theta \right)^2$, and define $\hat{a}_{\text{pool}}^{(w)}$ the estimated intercept in the pooled data without the membership variable for group w, *i.e.* we have $\hat{\theta}^{(w)} = \{\hat{a}_{\text{pool}}^{(w)}, \hat{\beta}_{\text{pool}}^{(w)}\}$ and the estimated intercept in the pooled data is:

$$\begin{aligned}
\hat{a}_{\text{pool}}^{(w)} &= \overline{Y^{(w)}} - \overline{X'^{(w)}} \hat{\theta}_{\text{pool}}^{(w)} \\
&= \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} Y_i(w) - \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} X_i'^{(w)} \hat{\theta}_{\text{pool}}^{(w)} \\
&= \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} \left(c^{(w)} + h_{k,i} + X_{k,i}' \theta^{(w)} + \varepsilon_i(w) \right) - \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} X_i'^{(w)} \hat{\theta}_{\text{pool}}^{(w)} \qquad \text{with } h_{k,i} = h_k \mathbb{1}_{\{H_i = k\}} \\
&= c^{(w)} + \overline{h_k^{(w)}} + \overline{\varepsilon_i(w)} + \overline{X^{(w)}} \left(\beta^{(w)} - \hat{\beta}_{\text{pool}}^{(w)} \right)
\end{aligned}$$

with:

- $\overline{h_k^{(w)}} = \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} h_{k,i} = \frac{1}{n^{(w)}} \sum_{k=1}^{n^{(w)}} n_k^{(w)} h_k$ is the average effect of study k in group w.
- $\overline{\varepsilon_i(w)} = \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} \varepsilon_i(w)$ is the average error in group w.
- $\overline{X^{(w)}} = \frac{1}{n^{(w)}} \sum_{i=1}^{n^{(w)}} X_i^{(w)}$ is the average covariate in group w.

Therefore, the estimate of the intercept in the pooled dataset in presence of study-effects has expectancy:

$$\begin{aligned}
\mathbb{E}(\hat{a}_{\text{pool}}^{(w)}) &= c^{(w)} + \mathbb{E}\left(\overline{h_k^{(w)}}\right) \\
&= c^{(w)} + \mathbb{E}\left(\frac{1}{n^{(w)}} \sum_{k=1}^{K} n_k^{(w)} h_k\right) \\
&= c^{(w)} + \sum_{k=1}^{K} \mathbb{E}\left(\frac{n_k^{(w)}}{n^{(w)}}\right) h_k
\end{aligned}$$

Then the unadjusted pooled estimator is:

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

We now prove that this unadjusted estimator is biased when the p_k probabilities are not equal across studies. Then, we have:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{\text{pool}}) &= \mathbb{E}\left(\hat{a}_{\text{pool}}^{(1)} - \hat{a}_{\text{pool}}^{(0)} + \overline{X}(\hat{\beta}_{\text{pool}}^{(1)} - \hat{\beta}_{\text{pool}}^{(0)})\right) \\
&= c^{(1)} - c^{(0)} + \sum_{k=1}^{K} \mathbb{E}\left(\frac{n_k^{(1)}}{n^{(1)}}\right) h_k - \sum_{k=1}^{K} \mathbb{E}\left(\frac{n_k^{(0)}}{n^{(0)}}\right) h_k \\
&= \tau + \sum_{k=1}^{K} \mathbb{E}\left(\frac{n_k^{(1)}}{n^{(1)}} - \frac{n_k^{(0)}}{n^{(0)}}\right) h_k \\
&= \tau + \underbrace{\sum_{k=1}^{K} \left(p_k \mathbb{E}\left(\frac{n_k}{\sum_{k=1}^{K} n_k p_k}\right) - (1-p_k)\mathbb{E}\left(\frac{n_k}{\sum_{k=1}^{K} n_k(1-p_k)}\right)\right) h_k}_{\text{bias}}
\end{aligned}$$

Then consider the two following cases:

- If equal treatment assignment ($p_k = p$ for all k):
  Then, the bias is equal to $\sum_{k=1}^{K} p\mathbb{E}\left(\frac{n_k}{pn}\right) h_k - (1-p)\mathbb{E}\left(\frac{n_k}{(1-p)n}\right) h_k = 0$ so that $\mathbb{E}(\hat{\tau}_{\text{pool}}) = \tau$.
- If unequal treatment assignment ($p_k \neq p$ for all k):
  Then, the bias is $\neq 0$ so that $\mathbb{E}(\hat{\tau}_{\text{pool}}) \neq \tau$.

Therefore, the Pool estimator is biased when the treatment probabilities are not equal among studies. This happens because the variable H acts as a confounder between the treatment and the outcome in cases of study-effects. We need to account for it by adding a membership variable H in the dataset, which will allow the model to estimate the study-effects.

B.4.2 With membership variable in the dataset

Proof of unbiasedness of the adjusted GD estimator: Denoting $\beta_H = (h_1, \ldots, h_K)^\top$ the coefficients of the membership variables, model (16) can be written as:

$$\begin{aligned}
Y_{k,i}(w) &= c^{(w)} + h_k + X_{k,i}\beta^{(w)} + \varepsilon_i(w) \\
&= c^{(w)} + H_{k,i}\beta_H + X_{k,i}\beta^{(w)} + \varepsilon_i(w)
\end{aligned} \tag{23}$$

Under Equation (23), the (adjusted) Pooled estimator then estimates the coefficients of the variables H as a substitute for the study-effects. This technique does not allow to estimate distinctly the intercepts $c^{(w)}$ and the effects of the studies $\{h_k\}_k$, as it relies on a relative rescaling of the intercepts of each study with respect to a choosen study of reference. In practice, choosing study 1 as the reference study by not including H_1 in the dataset offers the advantages of avoiding the underdeterminancy of the solutions of $\{(c^{(w)}, h_k)\}_k$ in Equation (23), and is easy to implement. In any case, this is without loss of generality.

Results from OLS estimation yield that the estimated intercept is the mean of the outcomes in the reference study (we arbitrarily choose study 1 to be the reference study), and the estimated coefficients of the membership variables are the differences between the means of the outcomes in the reference study and the other studies, which writes as:

$$\hat{c}^{(w)}_{\text{pool}} = \frac{1}{n_1^{(w)}} \sum_{i=1}^{n_1^{(w)}} Y_{1,i}(w)$$
$$\hat{h}_k = \frac{1}{n_k^{(w)}} \sum_{i=1}^{n_k^{(w)}} Y_{k,i}(w) - \hat{c}^{(w)}_{\text{pool}} \quad \forall k \in [\![2, K]\!]$$

where $\hat{h}_k^{(w)}$ is the estimated effect of study k on the outcomes in group w. In our model, we have that $\mathbb{E}(\hat{c}^{(w)}_{\text{pool}}) = Y_1(w)$ so that:

$$\begin{aligned}
\mathbb{E}\left(\hat{h}_k^{(w)}\right) &= \mathbb{E}\left(Y_k(w)\right) - \mathbb{E}\left(Y_1(w)\right) \\
&= \mathbb{E}\left(c^{(w)} + X_{k,i}\beta^{(w)} + h_k + \varepsilon_i(w)\right) - \mathbb{E}\left(c^{(w)} - X_{1,i}\beta^{(w)} - h_1 - \varepsilon_i(w)\right) \\
&= h_k
\end{aligned}$$

Finally, the adjusted pool estimator is:

$$\begin{aligned}
\hat{\tau}_{\text{pool}\circ} &= \frac{1}{n} \sum_{i=1}^{n} \left(\hat{c}^{(1)}_{\text{pool}} - \hat{c}^{(0)}_{\text{pool}} + X_i(\hat{\beta}^{(1)}_{\text{pool}} - \hat{\beta}^{(0)}_{\text{pool}}) + \sum_{k=1}^{K} H_{k,i}(\hat{\beta}^{(1)}_H - \hat{\beta}^{(0)}_H) \right) \\
&= \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\tilde{X}'_{k,i}(\hat{\theta}^{(1)}_{\text{pool}\circ} - \hat{\theta}^{(0)}_{\text{pool}\circ}) \right)
\end{aligned}$$

with the augmented design matrix $\tilde{X}'_{k,i} = (1, X_{k,i}, H_{2,i}, \dots, H_{K,i})$, $\hat{\beta}^{(w)}_H = (\hat{h}_1^{(w)}, \dots, \hat{h}_K^{(w)})^\top$ the estimated effects of the studies on the outcomes in group w and $\hat{\theta}^{(w)}_{\text{pool}\circ} = (\hat{c}^{(w)}_{\text{pool}}, \hat{\beta}^{(w)}_{\text{pool}}, \hat{\beta}^{(w)}_H)$. We then have $\mathbb{E}(\hat{\beta}^{(1)}_H) = \mathbb{E}(\hat{\beta}^{(0)}_H) = \beta_H$.

Finally, we get an unbiased estimator:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}_{\text{pool}\circ}) &= \mathbb{E}\left(\hat{c}^{(1)}_{\text{pool}} - \hat{c}^{(0)}_{\text{pool}} + \overline{X}(\hat{\beta}^{(1)}_{\text{pool}} - \hat{\beta}^{(0)}_{\text{pool}}) + \sum_{k=1}^{K} H_k(\hat{\beta}^{(1)}_H - \hat{\beta}^{(0)}_H) \right) \\
&= \tau + \sum_{k=1}^{K} \mathbb{E}\left(H_k(\hat{\beta}^{(1)}_H - \hat{\beta}^{(0)}_H) \right) \\
&= \tau + \sum_{k=1}^{K} \mathbb{E}\left(H_k \mathbb{E}(\hat{\beta}^{(1)}_H - \hat{\beta}^{(0)}_H \mid H_k) \right) \\
&= \tau + \sum_{k=1}^{K} \mathbb{E}\left(H_k(\beta_H - \beta_H) \right) \\
&= \tau
\end{aligned}$$

The asymptotic variance of the adjusted Pool estimator is given by (using Appendix B.2.6): $\mathbb{V}^\infty(\hat{\tau}_{\text{pool}\circ}) = \frac{\sigma^2}{p(1-p)} + \frac{1}{n}\|\tilde{\beta}^{(1)} - \tilde{\beta}^{(0)}\|^2_{\tilde{\Sigma}}$ with $\tilde{\beta}^{(w)} = (\beta^{(w)}, \beta_H)$ and $\tilde{\Sigma} = \mathbb{V}(\tilde{X})$. Furthermore, remark that the block covariance-

variance matrix is of the form $\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & \text{diag}(p_1, \dots, p_K) \end{pmatrix}$ because the study-effects are independent of the covariates X. Therefore, the asymptotic variance of the adjusted Pooled estimator is:

$$\mathbb{V}^{\infty}(\hat{\tau}_{\text{poolo}}) = \frac{\sigma^2}{p(1-p)} + \frac{1}{n}\|\beta^{(1)} - \beta^{(0)}\|^2_{\Sigma}$$

Note that the asymptotic variance of the adjusted Pool estimator in this study-effect model is insensitive to these effects, which means that the heterogeneity among the studies does not affect this estimator.

Therefore, we can build the federated Gradient Descent with H variables ("$\hat{\tau}_{\text{GDo}}$") by allowing the studies to add $K-1$ columns to their local datasets, with one of them containing only ones, unique to their dataset, and under convergence of Alg. D, we get $\hat{\tau}_{\text{GDo}} = \hat{\tau}_{\text{poolo}}$.

C SIMULATIONS

C.1 Simulation parameters

We generate data as follows:

1. Generate $X_{k,i} \sim \mathcal{N}(\mu_k, \Sigma_k)$ for individual $i \in [\![1, n_k]\!]$, study $k \in [\![1, K]\!]$ and d covariates. Add a constant covariate 1 to each $X_{k,i}$ to account for the intercept.
2. Generate $W_{k,i} \sim \mathcal{B}(p_k)$.
3. Generate $\varepsilon_{k,i} \sim \mathcal{N}(0, \sigma^2)$ (homoscedasticity).
4. Build $Y_{k,i}(w) = c^{(w)} + h_k + X_{k,i}\beta_k^{(w)} + \varepsilon_{k,i}$, with $W_{k,i} = w$.

- Studies:
  - $K = 5$ studies
  - Sample sizes: *Large* settings: $\{n_k = 20d\}_k$ so $n = 20Kd$; *Small* settings: $\{n_k = 5d\}_k$ so that $n = 4Kd$
- Model:
  - $\sigma^2 = 1$
  - $\{\theta_k^{(w)}\}_k = \{(c^{(w)} + h_k, \beta_1^{(w)}, \dots, \beta_d^{(w)})\}$ with $c^{(1)} = -1.85, c^{(0)} = -2$,
    $\beta^{(1)} = (-1.75, -1.5, -1.25, -1.0, -0.75, -0.5, -0.25, 0.0, 0.25, 0.5)$,
    $\beta^{(0)} = (-1.8, -1.6, -1.4, -1.2, -1, -0.8, -0.6, -0.4, -0.2, 0)$,
    $h_k = 0$ (no study-effect by default for model (2))
- Covariates:
  - Dimension $d = 10$
  - $\left\{ \left(\mu_k = (\underbrace{1, \dots, 1}_{\lfloor d/2 \rfloor}, \underbrace{-1, \dots, -1}_{\lfloor d/2 \rfloor}), \Sigma_k = 0.5 I_d + 0.5 J_d\right) \right\}_k$, $J \in \mathbb{1}^{d,d}$ the matrix of ones

  which yields $\tau = c^{(1)} - c^{(0)} + \mu^{\top}\left(\beta^{(1)} - \beta^{(0)}\right) = -1.1$
- Treatment assignment:
  - RCT: $W_{k,i} \sim \mathcal{B}(p_k)$ with $\{p_k = 0.5\}_k$;

For the Meta-IVW estimator, we estimate the asymptotic variance of the local ATE estimates, *i.e.* $\hat{\tau}_{\text{Meta-IVW}} = \frac{\sum_{k=1}^{K} \widehat{\mathbb{V}^{\infty}(\hat{\tau}_k)}^{-1} \hat{\tau}_k}{\sum_{k=1}^{K} \widehat{\mathbb{V}^{\infty}(\hat{\tau}_k)}^{-1}}$ with:

- $\widehat{\mathbb{V}^{\infty}(\hat{\tau}_k)} = \frac{\hat{\sigma}_k^2}{n_k}\left(\frac{1}{\widehat{p_k(1-p_k)}}\right) + \frac{1}{n_k}\|\hat{\beta}_k^{(1)} - \hat{\beta}_k^{(0)}\|^2_{\hat{\Sigma}_k}$

- Sample variance of the residuals
  $\hat{\sigma}_k^2 = \frac{1}{n_k - d - 1} \sum_{i=1}^{n_k} \left(Y_{k,i} - X_{k,i}(\hat{\beta}_k^{(1)} \mathbb{1}_{[W_i=1]} + \hat{\beta}_k^{(0)} \mathbb{1}_{[W_i=0]})\right)^2$
- $\hat{p}_k = \frac{n_k^{(1)}}{n_k}$ and the sample covariance matrix $\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_{k,i} - \overline{X}_k)(X_{k,i} - \overline{X}_k)^\top$

C.2 Additional scenarios

C.2.1 Balanced datasets

In this scenario, we generate data in the *homogeneous* setting according to the graphical model in Figure 1 and outcome model 2 with $K = 5$ studies and $d = 10$ covariates dimension. We consider a balanced setting where all studies have equal sample size $n_k = 100d$ in the *Large* regime and $n_k = 5d$ in the *Small* sample size regime. The treatment assignment is the same for all studies in this first scenario, with $\forall k, W_{k,i} \sim \mathcal{B}(p)$ and $p = 0.5$ for all k. With the theoretical considerations on the hyperparameters discussed in Appendix D, we set them to the following values for the full batch Gradient Descent algorithm: $T = 1000, E = 1, B = n, \eta = 0.001$ for the *Small* setting.

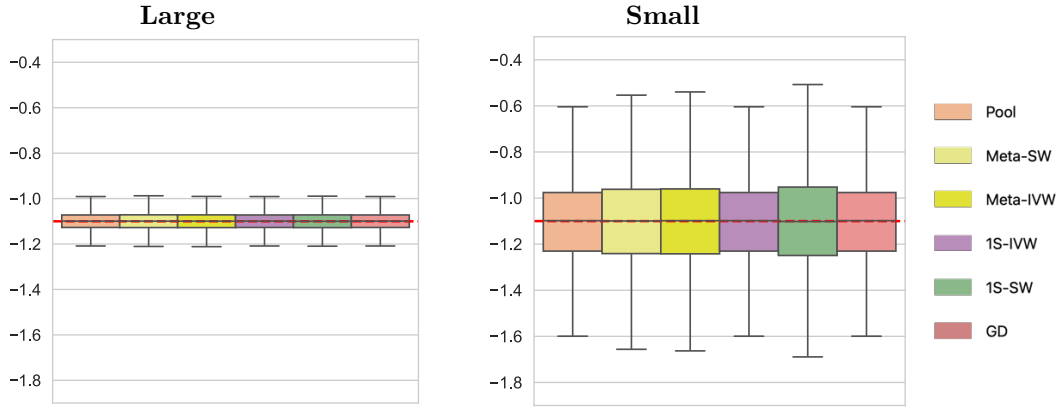


Figure 8: Estimation of the ATE in the balanced and homogeneous setting for RCT studies.

Figure 8 represents the distribution over 2000 simulations of the estimated ATEs using the Pool estimator on the concatenated data as well as the Meta estimators with both weighting strategies (sample weighting and inverse variance weighting), the One-Shot estimators (IVW and SW) and Gradient Descent estimator (GD).

In large sample size regime, it highlights that in the balanced and homogeneous setting, (left panel *Large*), all estimators are unbiased and have the same variance, as expected in the special case of one RCT conducted over K studies (Section 3.2).

In the small sample size regime (right panel *Small*), the Metas and One-Shot SW have larger variances than the One-Shot IVW and GD estimators which are both equal to the pooled data one as expected given Theorem 1.

C.2.2 Imbalanced datasets

We consider a case of imbalance in the sample sizes of the studies, where one study has more observations than the others. For the *Large* setting, $n_1 = 400d$ and $n_2 = ... = n_5 = 25d$, leading to the same total number of observations of $n = 5000$ as in Figure 8, whereas in the *Small* case $n_1 = 13d$ and $n_2 = ... = n_5 = 3d$ resulting in $n = 250$, similarly to the balanced scenario.

Figure 9 shows that in the *Large* case, the partition of the pooled data has no impact on the estimation of the ATE for all estimators as long as Condition 1 holds. The boxplots are similar to the ones obtained in the balanced case (Figure 8), as expected again with the "One RCT" scenario.

However, in the *Small* case, the variances of the Meta estimators and 1S-SW are greater than in the balanced setting due to the local estimates $\hat{\tau}_2, \dots \hat{\tau}_5$ being obtained after performing two local OLS regressions on each of their treatment arms on very small datasets.

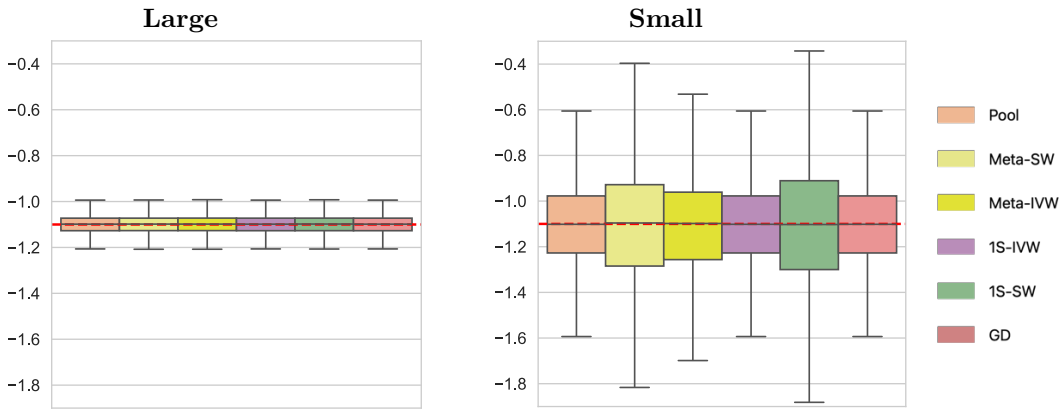


Figure 9: Estimation of the ATE in the imbalanced and homogeneous setting for RCT studies.

C.3 Shift in covariates distributions

We now consider an *heterogeneous* setting where the data are generated according to the graphical model in Figure 2 under model (2), where individuals follow different distributions according to the study they belong to. We consider the case where the means and covariance matrices $\{(\mu_k, \Sigma_k)\}_k$ are different from one study to another, with values:

Means	Covariances
$\mu_1 = (\underbrace{1, \dots, 1}_{\lfloor d/2 \rfloor}, \underbrace{-1, \dots, -1}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_1 = I_d + 0.5 - 0.5 I_d$
$\mu_2 = (\underbrace{-1, \dots, -1}_{\lfloor d/2 \rfloor}, \underbrace{1, \dots, 1}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_2 = 20 \cdot \Sigma_1 + 0.5 I_d - 0.5$
$\mu_3 = (0, \dots, 0)^\top$	$\Sigma_3 = 0.02 \cdot \Sigma_1 + 0.7 I_d$
$\mu_4 = (\underbrace{0.5, \dots, 0.5}_{\lfloor d/2 \rfloor}, \underbrace{-1, \dots, -1}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_4 = 1 \cdot \Sigma_1 + 0.5 I_d - 0.15$
$\mu_5 = (\underbrace{1.2, \dots, 1.2}_{\lfloor d/2 \rfloor}, \underbrace{0.8, \dots, 0.8}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_5 = 1.5 \cdot \Sigma_1 + 0.5 I_d - 0.15$

Note that in this scenario, the Meta-IVW is not relevant as explained in Section 4. In this setting, the targeted ATE is $\tau = \sum_{k=1}^{K} \rho_k \left(c^{(1)} - c^{(0)} + \mu_k^\top \left(\beta^{(1)} - \beta^{(0)} \right) \right) \approx 0.45$.

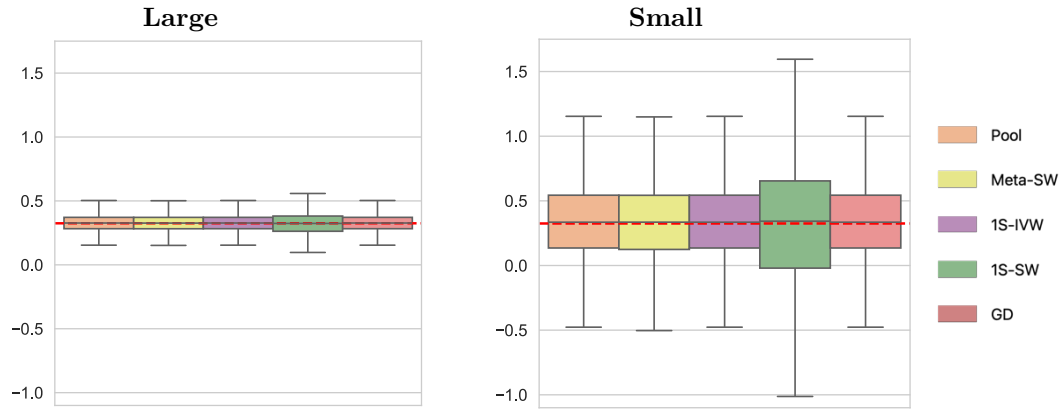


Figure 10: Estimation of the ATE in the heterogeneous setting for RCT studies.

Figure 10 illustrates the fact that when studies have different covariate means, the variance of the One-Shot

SW estimator is enlarged compared to homogeneous settings (Proposition 2), both in *Small* and *Large* sample sizes regimes. Notice that the Meta-SW estimator achieves better performance than in the homogeneous setting since its weights n_k/n are good estimates of ρ_k under Condition 1 of the true weights of the local ATEs, as $\tau = \sum_{k=1}^{K} \rho_k \tau_k$.

C.4 Study-effects

We now generate data according to model (16) under the graphical model in Figure 3, with study-effects equal to $(h_1, h_2, h_3, h_4, h_5) = (1, .2, -1, 30, 2)$, and with the other parameters set to their default values (Appendix C.1).

C.4.1 No adjustment

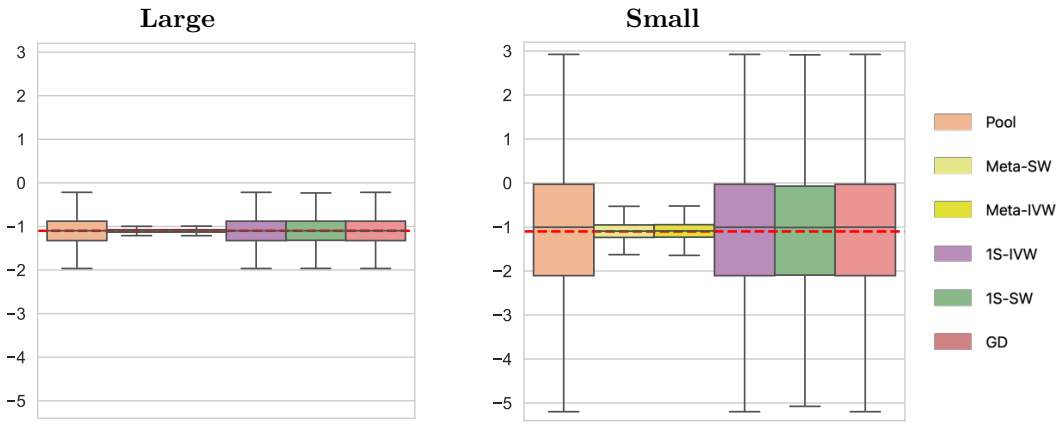


Figure 11: Estimation of the ATE in a homogeneous balanced setting with study-effects for RCTs.

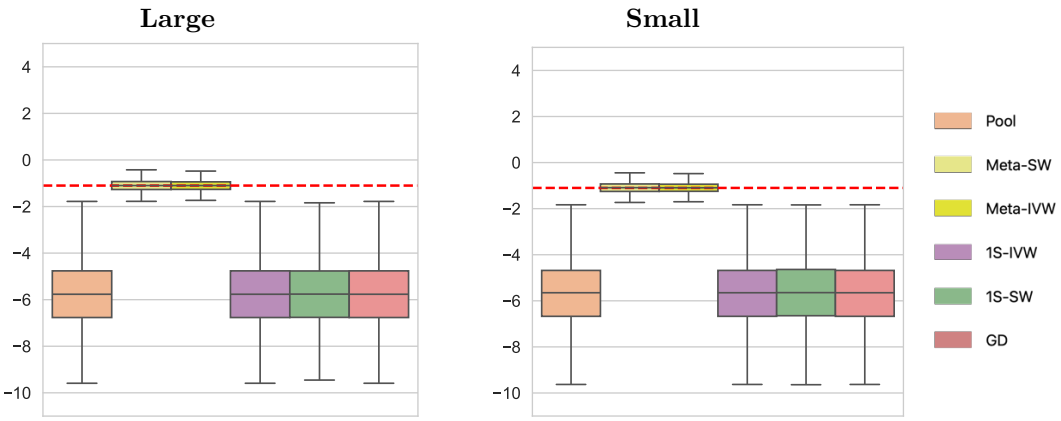


Figure 12: Estimation of the ATE in a homogeneous setting with study-effects and different treatment allocation schemes for RCTs.

Figures 11 and 12 present the distribution of all estimators without any modification, respectively in the "One RCT" scenario, and in the different treatment probabilities p_k setting. The estimators are not adjusted to take into account the presence of study-effects, leading to the presence of a confounder between the treatment and the outcome variables (Assumption c), violating the identifiability assumption.

This case illustrates the advantage of Meta estimators over other estimators, which require in this setting less a priori knowledge on the underlying model in order to be used.

C.4.2 Adjusted

We now consider the same study-effects scenario when adjusting the estimators with the procedures described in Section 4.2: the Pool and GD estimators are computed with access to the membership variables, and the One-Shot estimators do not federate local intercepts. The computation of the Meta estimators remains unchanged.

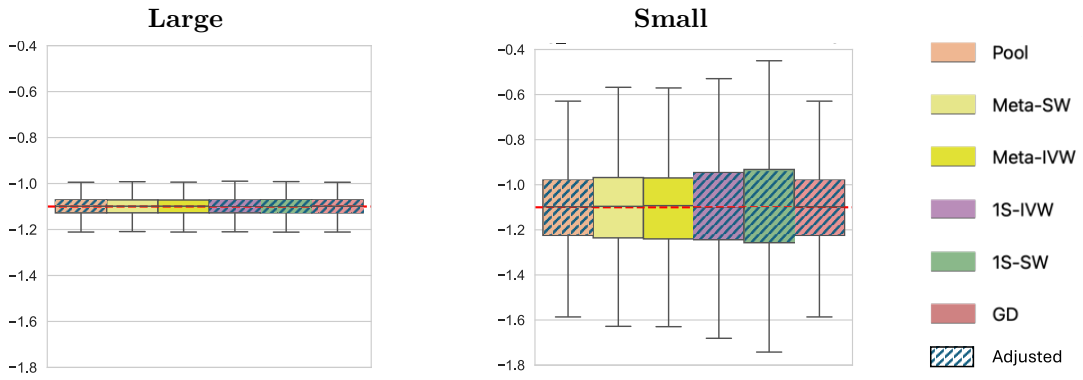


Figure 13: Estimation of the ATE in a homogeneous setting with study-effects for RCTs.

Figure 13 shows that all the estimators are now unbiased, and their variance are equal to the pooled data one in the *Large* sample size regime.

However, in the *Small* regime, the adjusted One-Shot IVW and SW are less efficient than others, as they suffer from the unshared intercepts' variances. The adjusted One-Shot IVW does not benefit from Theorem 1. Their variances are highly dependent on the partition of the data into K splits, unlike the Pool and GD estimators whose variances solely depend on the total amount of data. The meta estimators naturally handle the study-effects and their variances do not suffer much from their magnitude.

This scenario still highlights the advantage of the Meta estimators, which do not require a specific modelling of the study-effects. However, in this setting they have a higher variance than the (adjusted) GD and (adjusted) pool estimators, as in the simple homogeneous balanced setting (C.2.1).

C.5 Full heterogeneity: shifts in covariates distributions, study-effects and in treatment allocation

We now combine the different scenarios of Appendix C.3, Appendix C.4 with study-effects (specific intercept per study), distribution of the covariates that are different from one study to the other (different means and covariance matrices) and different probabilities of being treated by study, corresponding to the graphical model in Figure 5. The simulation parameters used in Figures 4b are displayed in Appendix C.6:

Means	Covariances	Study-Effects	Treatment probabilities
$\mu_1 = (\underbrace{1,\dots,1}_{\lfloor d/2 \rfloor}, \underbrace{-1,\dots,-1}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_1 = I_d + 0.5 - \frac{1}{2} I_d$	$h_1 = 1$	$p_1 = 0.75$
$\mu_2 = (\underbrace{-1,\dots,-1}_{\lfloor d/2 \rfloor}, \underbrace{1,\dots,1}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_2 = 20 \cdot \Sigma_1 + 0.5 I_d - 0.5$	$h_2 = 0.2$	$p_2 = 0.75$
$\mu_3 = (0,\dots,0)^\top$	$\Sigma_3 = 0.02 \cdot \Sigma_1 + 0.7 I_d$	$h_3 = -1$	$p_3 = 0.75$
$\mu_4 = (\underbrace{0.5,\dots,0.5}_{\lfloor d/2 \rfloor}, \underbrace{-1,\dots,-1}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_4 = 1 \cdot \Sigma_1 + 0.5 I_d - 0.15$	$h_4 = 30$	$p_4 = 0.25$
$\mu_5 = (\underbrace{1.2,\dots,1.2}_{\lfloor d/2 \rfloor}, \underbrace{0.8,\dots,0.8}_{\lfloor d/2 \rfloor})^\top$	$\Sigma_5 = 1.5 \cdot \Sigma_1 + 0.5 I_d - 0.15$	$h_5 = 2$	$p_4 = 0.25$

C.6 G-Formula OLS covariate adjustment in non-linearity

In this setting, we simulate one dataset where $Y_i(1) = 0x_1^2 + \frac{-1}{2}x_2^2 + \frac{1}{2}x_3^2 + \frac{3}{2}x_4^2 + x_3 * x_4$ and $Y_i(0) = -0.35x_1^2 + 0x_2^2 + \frac{1}{2}x_3^2 + \frac{3}{2}x_4^2 + x_1 * x_2$.

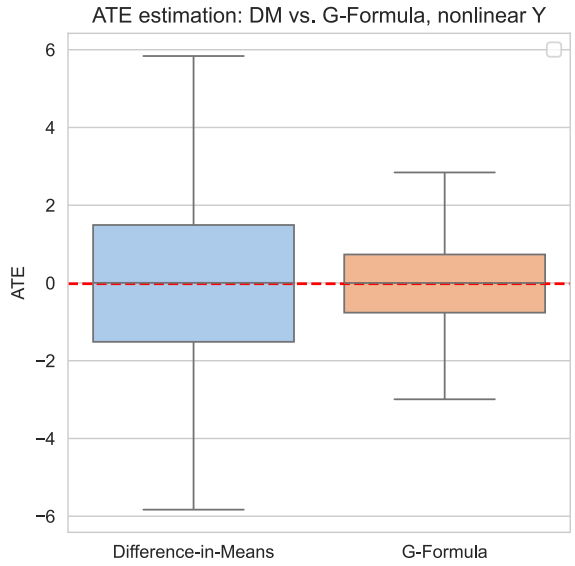


Table 5: ATE Estimation: DM vs. G-Formula with OLS adjustment on covariates under non-linear outcome modeling

The pooled-data OLS G-Formula has reduced variance compared to the simple Difference-in-Means, even if the outcome model is non-linear, illustrating (U.S. Food and Drug Administration, 2023; European Medicines Agency, 2024; Tsiatis et al., 2008; Benkeser et al., 2021; Lin, 2013; Wager, 2020; Lei and Ding, 2021; Van Lancker et al., 2024).

D Algorithms

We describe in Algorithm 1 the procedure to obtain the Gradient Descent estimator of outcome model parameters $\hat{\theta}_{\text{GD}}$:

We provide some details on the choices of learning rates discussed in Section 3.1.3:

- In the setting where $T = 1$ and $E \to \infty$, each study can choose a local learning rate $\eta_k \leq \frac{2}{L_k}$, where L_k is the smoothness constant of the local problem. This quantity is equal to the largest eigenvalue of the covariance matrix of study k, $(X_k^{(w)})^\top X_k^{(w)}$, so choosing a learning rate $\eta_k = \frac{2}{\lambda_{\max,k}}$ is a (conservative) choice which ensures the convergence to the local solution $\hat{\theta}_k^{(w)}$.
- When performing one local step per round ($E = 1$), one can ensure that $\hat{\theta}_{\text{GD}}^{(w)} \to \hat{\theta}_{\text{pool}}^{(w)}$ as $T \to \infty$ by choosing a learning rate $\eta < \frac{2}{L}$, where L is the smoothness constant of the global problem. Here, L corresponds to the highest eigenvalue $\lambda_{\max}$ of the pooled data $(X^{(w)})^\top X^{(w)}$. Its computation in the federated setting is not straightforward (although there exists some methods, e.g., based on distributed power method). A simple alternative that requires a single round of communication consists in noticing that $\lambda_{\max} \leq \sum_{k=1}^K \lambda_{\max,k}$,

Algorithm 1 Federated Averaging (FedAvg) algorithm to learn $\hat{\theta}^{(w)}_{\text{GD}}$

Input: K studies, E local steps, B batch size, η learning rate, T rounds of communication
Server executes:
  Initialize $\theta^{(w)}_0$
for each round $t = 0, 1, \dots T$ **do**
  for each study $k \in [\![1, K]\!]$ **in parallel do**
    $\theta^{k^{(w)}}_{t+1} \leftarrow \text{LocalUpdate}(k, \theta^{(w)}_t)$
  end for
  $\theta^{(w)}_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n^{(w)}_k}{n^{(w)}} \theta^{k^{(w)}}_{t+1}$ `// Federated Averaging`
end for
LocalUpdate$(k,\ \theta^{(w)}_k)$:
for each local step $e = 0, 1, \dots, E-1$ **do**
  $\mathcal{B}_k \leftarrow$ a random batch of B samples from $\mathcal{Z}_k$
  $\nabla \ell(\theta^{(w)}_k, \mathcal{B}_k) \leftarrow -\frac{2}{B} X^{\top}_{\mathcal{B}_k} (Y_{\mathcal{B}_k} - X_{\mathcal{B}_k}\ \theta^{(w)}_k)$ `// Compute gradient on the mini-batch`
  $\theta^{(w)}_k \leftarrow \theta^{(w)}_k - \eta \nabla \ell(\theta^{(w)}_k, \mathcal{B}_k)$ `// Gradient descent step`
end for
return $\hat{\theta}^{(w)}_{\text{GD}} \leftarrow \theta^{(w)}_k$

where $\lambda_{\max,k}$ is the highest eigenvalue of study k, to set the learning rate smaller than $\frac{2}{\sum_{k=1}^{K} \lambda_{\max,k}}$. In the homogeneous setting, since all sites have equal covariance matrices, $\lambda_{\max}$ can be approximated by a mere averaging of the local estimates of the highest eigenvalues, i.e. $\hat{\lambda}_{\max} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\lambda}_{\max,k}$ where $\hat{\lambda}_{\max,k}$ is learned with any eigenvalues estimation method locally.

For a choice of a learning rate as discussed above, setting:

$$T = \tilde{\Omega}\left(\frac{\lambda_{\max} E}{\lambda_{\min}} \log(1/\varepsilon) + \frac{\sqrt{\lambda_{\max}} \zeta E}{\lambda_{\min} \sqrt{\varepsilon}} \right),$$

gives $\|\theta^{(w)}_T - \hat{\theta}^{(w)}_{\text{pool}}\|^2 \leq \varepsilon$, where $\theta^{(w)}_T$ is the output of FedAvg after T communication rounds, $\lambda_{\max}, \lambda_{\min}$ are respectively the maximum and minimum eigenvalues of ${X^{(w)}_{\text{pool}}}^{\top} X^{(w)}_{\text{pool}}$ ($\lambda_{\min} > 0$ under Condition 2), and $\zeta^2 = \frac{1}{K} \sum_{k=1}^{K} \left\| \frac{1}{n_k} \sum_{i=1}^{n_k} X_{k,i} ({X^{(w)}_{k,i}}^{\top} \hat{\theta}^{(w)}_{\text{pool}} - Y^{(w)}_{k,i}) \right\|^2$, and the $\tilde{\Omega}$ notation hides constants and polylogarithmic factors (Koloskova et al., 2020).

Applying these hyperparameters choices to the simulation setting described in Appendix C, we get in the Homogeneous scenario (C.2.1) $\hat{\lambda}^{(w)}_{\max} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\lambda}^{(w)}_{\max,\text{k}} \approx 10.5$ for both treated and control groups, which yields a choice of learning rate $\eta < \min_{w \in \{0,1\}} (2/\hat{\lambda}^{(w)}_{\max}) \approx 0.19$. We chose to divide the upper bound by 10 to ensure stability of the algorithm, which is a common practice, and thus set η to 10^{-2}.

In the simulation, we most often used the following values of parameters for the full batch Gradient Descent algorithm: T ranging from 1000 to 4000, with increased values for heterogeneous settings, $E = 1, B = n, \eta = 10^{-2}$.