FEDERATED CAUSAL INFERENCE ON MULTI-SITE OBSERVATIONAL DATA VIA PROPENSITY SCORE AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal inference typically assumes centralized access to individual-level data. Yet, in practice, data are often decentralized across multiple sites, making centralization infeasible due to privacy, logistical, or legal constraints. We address this problem by estimating the Average Treatment Effect (ATE) from decentralized observational data via a Federated Learning (FL) approach, allowing inference through the exchange of aggregate statistics rather than individual-level data. We propose a novel method to estimate propensity scores by computing a federated weighted average of local scores with Membership Weights (MW)—probabilities of site membership conditional on covariates—which can be flexibly estimated using parametric or non-parametric classification models. Unlike density ratio weights (DW) from the transportability and generalization literature, which either rely on strong modeling assumptions or cannot be implemented in FL, MW can be estimated using standard FL algorithms and are more robust, as they support flexible, non-parametric models—making them the preferred choice in multi-site settings with strict data-sharing constraints. The resulting propensity scores are used to construct Federated Inverse Propensity Weighting (Fed-IPW) and Augmented IPW (Fed-AIPW) estimators. Unlike meta-analysis methods, which fail when any site violates positivity, our approach leverages heterogeneity in treatment assignment across sites to improve overlap. We show that Fed-IPW and Fed-AIPW perform well under site-level heterogeneity in sample sizes, treatment mechanisms, and covariate distributions. Both theoretical analysis and experiments on simulated and real-world data highlight their advantages over meta-analysis and related methods.

1 INTRODUCTION

The Average Treatment Effect (ATE) is a key causal estimand used to quantify the effect of a treatment on an outcome and is commonly employed as the primary measure of efficacy in evaluating new therapies, including vaccines, before regulatory approval (Polack et al., 2020). In Randomized Clinical Trials (RCTs), treatment assignment is randomized, ensuring that the observed association between treatment and outcome reflects a causal effect. Under this design, the ATE can be consistently estimated using a simple Difference-in-Means (DM) estimator (Splawa-Neyman, 1990), which can be further refined through covariate adjustment to reduce variance (FDA, 2023; EMA, 2024; Lei & Ding, 2021). However, RCTs are often expensive, time-consuming, or infeasible. In such cases, estimating treatment effects from observational data becomes the only viable alternative (Hernán, 2018; Hernán & Robins, 2006). Although such real-world data is abundant, drawing causal inferences from it is challenging due to confounding covariates, rendering the unadjusted DM estimator biased (Grimes & Schulz, 2002). Adjusting for confounders is thus essential (VanderWeele, 2019). This can be done by predicting counterfactual outcomes before averaging the differences (the G-formula plug-in estimator, Robins, 1986). Another approach is to weight individuals according to their treatment probability, emulating a randomized trial. For instance, the Inverse Propensity Weighting (IPW) estimator (Rosenbaum & Rubin, 1983) relies on estimating the *propensity score*—the probability of treatment given covariates. Doubly robust estimators such as the Augmented IPW (AIPW) (Bang & Robins, 2005) combine weighting with outcome modeling to remain consistent as long as either model is correctly specified.

Larger datasets improve the precision of treatment effect estimates, especially for underrepresented subgroups. Yet real-world data is typically decentralized—spread across hospitals, companies, or countries—making aggregation difficult, particularly in healthcare where privacy regulations, data ownership, and governance issues impede centralization. Federated Learning (FL) (Kairouz et al., 2021) offers a solution to train models across distributed data without sharing individual-level data. While FL has been largely applied to prediction tasks, its extension to causal inference remains limited. This problem is especially challenging in observational studies, where differences in covariate distributions and treatment assignment mechanisms across sites create multiple sources of heterogeneity that must be addressed without sharing raw data, while achieving results comparable to centralized analyses—an issue that remains largely unsolved.

Contributions. We propose federated (A)IPW estimators for decentralized observational data, moving beyond the aggregation of local ATE estimates used in meta-analysis (Riley et al., 2023). At the core of our approach is a flexible, primarily non-parametric strategy for federating propensity scores. Unlike prior methods that fit a single global parametric model via parameter averaging (Xiong et al., 2023) or federated gradient descent (Guo et al., 2025), we construct a global propensity score as a mixture of locally estimated models. This involves (1) local estimation of propensity scores at each site—accommodating heterogeneity in treatment assignment and flexibility in model choice—and (2) aggregation into a global score using Membership Weights (MW), i.e., the probability of site membership given covariates. MW can be estimated in a federated manner using flexible, potentially non-parametric classification models, ensuring both robustness and communication efficiency. In contrast, transferring Density Ratio Weights (DW) from the transportability and generalization literature to our setting requires strong modeling assumptions, as they are otherwise incompatible with FL constraints. Using our federated propensity scores, we build the Federated IPW (Fed-IPW) and its augmented variant (Fed-AIPW), derive their variances, and show they achieve equal or lower variance than meta-analysis estimators.

Our approach is particularly advantageous when overlap between treatment groups is poor or absent within sites. In such scenarios, cross-site collaboration becomes crucial, as combining sites increases overall overlap and enables treatment effect estimation that may be infeasible locally. Indeed, when treatment assignment mechanisms differ substantially—such as when one site treats a subgroup absent elsewhere—the combined dataset achieves markedly greater overlap, allowing treatment effects to be estimated that would otherwise be poorly identified in isolation. Additionally, our framework naturally accommodates heterogeneity in sample sizes, treatment policies, covariate distributions, and violations of positivity. Numerical experiments on simulated and real data confirm our theoretical findings and highlight the method's practical benefits.

Related work. Federated causal inference is still nascent. Khellaf et al. (2025) estimate federated G-formula ATE estimates across multiple RCTs by fitting parametric outcome models at each site via FL, achieving lower variance than DM. In observational settings, Vo et al. (2022) propose a federated Bayesian approach using Gaussian processes with a shared covariance kernel, but it requires sharing the first four moments of the data, limiting scalability and efficiency. Guo et al. (2025) learn a global propensity score via consensus voting over parametric parameters, retaining only sites meeting a shared specification. In contrast, our method assumes no common propensity score: each site may fit its own model.

To address site heterogeneity, Xiong et al. (2023) use a logistic propensity model with shared and site-specific parameters, federating only the common ones. Yin et al. (2025) fit a global model adjusting for covariates and site membership but limit heterogeneity to a site-specific scalar. By contrast, our method makes no structural assumptions, enabling fully nonparametric estimation with heterogeneous local models and relaxing the need for local overlap at each site.

A related body of work focuses on generalizing causal findings from multi-site source populations to a target population. Han et al. (2025) use density ratio weighting of local ATEs to adjust for covariate shift but assume homogeneous nuisance functions across sites and rely on meta-analysis of aggregate statistics. Guo et al. (2024) extend this idea by applying density ratio weights to aggregate local propensity scores to construct a target-specific score, which requires density estimation *within each treatment arm at each site*—demanding large sample sizes per arm for stable estimates. In both cases, non-parametric density ratio estimation is infeasible under FL constraints, as it requires sharing raw data or detailed covariate representations (e.g., kernel evaluations or high-dimensional histograms). In contrast, our MW-based approach leverages flexible parametric or non-parametric

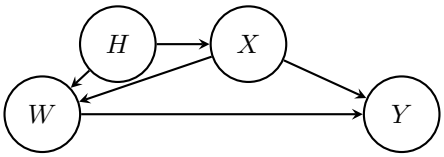


Figure 1: Graphical model for multi-site observational data.

supervised models (e.g., logistic regression, neural networks, gradient-boosted trees) for which federated training is already established, operating without sharing raw data, potentially at lower sample complexity, with full support for heterogeneous nuisance functions, and allowing the natural inclusion of external control arms.

2 PRELIMINARIES

2.1 ATE ESTIMATORS FROM CENTRALIZED MULTI-SITE OBSERVATIONAL DATA

In this section, we recall the key components of ATE estimation in a centralized multi-site setting. Following the potential outcomes framework (Rubin, 1974; Splawa-Neyman, 1990), we consider random variables $(X, H, W, Y(1), Y(0))$, where $X \in \mathbb{R}^d$ represents patient covariates, $H \in [K]$ indicates site membership, $W \in \{0, 1\}$ denotes the binary treatment, and $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, respectively. We assume that the Stable Unit Treatment Values Assumption (SUTVA) holds, so that the observed outcome is $Y = WY(1) + (1 - W)Y(0)$, and that the potential outcomes are uniformly bounded. In the centralized setting, we observe $n = \sum_{k=1}^{K} n_k$ observations of independently and identically distributed (i.i.d.) tuples $(H_i, X_i, W_i, Y_i)_{i \in 1,\dots,n} \sim \mathcal{P}^{\otimes n}$, with $n_k = \sum_{i=1}^{n} \mathbb{1}_{\{H_i = k\}}$ the number of observations in site k. We aim to estimate the ATE defined as the risk difference $\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid H]]$, where the expectation is taken over the population $\mathcal{P}$. To be able to identify the ATE, we assume unconfoundedness (standard in causal inference) and further consider Assumption 2, which is specific to the multi-site setting.

Assumption 1 (Unconfoundedness). $(Y(0), Y(1)) \perp\!\!\!\perp W \mid X$.

Assumption 2 (Ignorability on sites). $(Y(0), Y(1)) \perp\!\!\!\perp H \mid X$.

Robertson et al. (2021) refer to Assumption 2 as the *no center-outcome association* condition. It can be stated as a testable null hypothesis requiring that, for every treatment level w and any pair of centers (k, k'), $\mathbb{E}[Y \mid X, W = w, H = k] = \mathbb{E}[Y \mid X, W = w, H = k']$. Combined with Assumption 1, this ensures that X forms a sufficient set of covariates for confounding adjustment. Our setting is depicted in the graphical model in Figure 1, highlighting that we remove any direct effect of the site on the outcome.

We define $\mu_w(x) = \mathbb{E}[Y \mid X = x, W = w]$ for $w \in \{0, 1\}$, and let $\tau(x) = \mu_1(x) - \mu_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ be the Conditional Average Treatment Effect (CATE). The oracle propensity score is denoted by $e(x) = \mathbb{P}(W \mid X = x)$, and we consider the weak (global) overlap assumption (Wager, 2024).

Assumption 3 (Global overlap). $\mathcal{O}_{\text{global}} = \mathbb{E}\left[(e(X)(1 - e(X)))^{-1}\right] < +\infty$.

Assumption 3 is crucial for propensity score-based estimators, as it states that every region of the covariate space has a non-zero probability of receiving both treatments. A lower value of $\mathcal{O}_{\text{global}}$ indicates that these probabilities lie further away from 0 and 1, which corresponds to better overlap. For further insights on overlap, see Li et al. (2018a;b), and for a "misoverlap" metric, refer to Clivio et al. (2024).

With Assumptions 1, 2 and 3, the ATE is identifiable as $\tau = \mathbb{E}\left[\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)}\right]$ (see Appendix A.1). Throughout the paper, we denote oracle ATE estimators, which assume knowledge of the nuisance components e, μ_0, μ_1, by a superscript *. We define the *Oracle multi-site centralized estimators* as

follows:

$$\hat{\tau}^*_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \tau_{\text{IPW}}(X_i; e), \qquad \hat{\tau}^*_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \tau_{\text{AIPW}}(X_i; e, \mu_1, \mu_0), \tag{1}$$

where $\tau_{\text{IPW}}(X_i; e) = \frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}$ and $\tau_{\text{AIPW}}(X_i; e, \mu_1, \mu_0) = \mu_1(X_i) - \mu_0(X_i) + \frac{W_i(Y_i-\mu_1(X_i))}{e(X_i)} - \frac{(1-W_i)(Y_i-\mu_0(X_i))}{1-e(X_i)}$. These oracle estimators are unbiased and asymptotically normal.

Theorem 1. *Under Assumptions 1, 2 and 3, we have* $\sqrt{n}(\hat{\tau} - \tau) \to \mathcal{N}(0, V)$ *with*

$$\begin{cases} V_{\text{IPW}} = \mathbb{E}\left[\frac{Y(1)^2}{e(X)}\right] + \mathbb{E}\left[\frac{Y(0)^2}{1-e(X)}\right] - \tau^2, \\ V_{\text{AIPW}} = \mathbb{V}\left[\tau(X)\right] + \mathbb{E}\left[\left(\frac{(Y-\mu_1(X))^2}{e(X)}\right)\right] + \mathbb{E}\left[\left(\frac{(Y-\mu_0(X))^2}{1-e(X)}\right)^2\right]. \end{cases}$$

The above asymptotic variances align with those in the single-site setting (Hirano et al., 2003), as detailed in Appendix A.2. However, in practice, the propensity score and outcome models are typically unknown and must be estimated from data. This creates a challenge in the decentralized setting, where centralizing data to compute μ_1, μ_0, and e is not feasible. Therefore, the estimators in Definition 1 need to be adapted to this setting. Importantly, the (non-oracle) AIPW estimator is inherently *doubly robust*, remaining consistent as long as either the outcome or the propensity score model is correctly specified (Chernozhukov et al., 2018).

2.2 Meta-Analysis Estimators

We now turn to a decentralized setting in which the K sites cannot share individual-level data. A natural baseline for estimating the ATE across sites is a two-stage meta-analysis approach (Burke et al., 2017), wherein each site independently estimates the relevant nuisance parameters and communicates only the resulting ATE estimates for aggregation. In this setting, we need the following assumption.

Assumption 4 (Local overlap). $\forall k \in [K], \mathcal{O}_k = \mathbb{E}\big[(e(X)(1-e(X)))^{-1} \mid H = k\big] < +\infty$.

Assumption 4 is much stronger than global overlap (Assumption 3), as it must hold at every site. Denoting by $e_k(x) = \mathbb{P}(W = 1 \mid X = x, H = k)$ the oracle local propensity score at site k, we can define the oracle meta-analysis estimators as follows:

$$\hat{\tau}^{\text{meta}^*}_{\text{IPW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{(k)}_{\text{IPW}}, \qquad \hat{\tau}^{\text{meta}^*}_{\text{AIPW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{(k)}_{\text{AIPW}}, \tag{2}$$

where $\hat{\tau}^{(k)}_{\text{IPW}} = \frac{1}{n_k}\sum_{i=1}^{n_k} \tau_{\text{IPW}}(X_i; e_k)$ and $\hat{\tau}^{(k)}_{\text{AIPW}} = \frac{1}{n_k}\sum_{i=1}^{n_k} \tau_{\text{AIPW}}(X_i; e_k, \mu_1, \mu_0)$ are the local estimators at site k. While alternative aggregation weights—such as the inverse variance of local estimates—can be considered, they produce, in our setting, biased estimates of the global ATE $\tau = \sum_{k=1}^{K} \rho_k \tau_k$, where $\rho_k = \mathbb{P}(H = k)$ and $\tau_k = \mathbb{E}[Y(1) - Y(0) \mid H = k]$ is the local ATE. This bias appears whenever the τ_k differ, which commonly occurs when covariate distributions vary across sites and treatment effects are heterogeneous (i.e., depend on covariates), see (Berenfeld et al., 2025).

Theorem 2. *Under Assumptions 1, 2 and 4, the oracle meta-analysis estimators are unbiased for the ATE with asymptotic variances*

$$V^{\text{meta}^*}_{\text{IPW}} = \sum_{k=1}^{K} \rho_k V^{(k)}_{\text{IPW}} + \mathbb{V}\left[\tau_H\right], \qquad V^{\text{meta}^*}_{\text{AIPW}} = \sum_{k=1}^{K} \rho_k V^{(k)}_{\text{AIPW}} + \mathbb{V}\left[\tau_H\right],$$

with within-site variance

$$\begin{cases} V^{(k)}_{\text{IPW}} = \mathbb{E}\left[\frac{Y(1)^2}{e_k(X)} \mid H = k\right] + \mathbb{E}\left[\frac{Y(0)^2}{1-e_k(X)} \mid H = k\right] - \tau_k^2 \\ V^{(k)}_{\text{AIPW}} = \mathbb{V}\left[\tau(X) \mid H = k\right] + \mathbb{E}\left[\left(\frac{(Y-\mu_1(X))^2}{e_k(X)}\right)^2 \mid H = k\right] + \mathbb{E}\left[\left(\frac{(Y-\mu_0(X))^2}{1-e_k(X)}\right) \mid H = k\right], \end{cases}$$

and $\mathbb{V}\left[\tau_H\right] = \mathbb{V}\left[\mathbb{E}\left[Y(1) - Y(0) \mid H\right]\right]$ *the between-sites variance of the local ATEs.*

This result is proved in Appendix A.3. A key limitation of meta-analysis estimators is their reliance on Assumption 4, which is fragile and often violated—for instance, when a site applies a deterministic treatment policy (treating all patients or only a subgroup). In such cases, these estimators are ill-defined, yielding biased ATE estimates. To address this, we propose a federated approach that constructs the global propensity score e as a weighted combination of local scores e_k, enabling valid inference even without local overlap.

3 Federated Estimators via Propensity Score Aggregation

3.1 Oracle Federated Estimators

As discussed before, existing federated causal inference methods often rely on restrictive assumptions—such as a common propensity score across sites (Guo et al., 2025), site differences limited to intercept shifts (Yin et al., 2025), or predefined shared structures (Xiong et al., 2023). In practice, treatment assignment frequently varies across sites due to differences in norms, resources, or clinical practices. To properly account for this heterogeneity, the global propensity score must be expressed as a *weighted combination* of the site-specific scores (see Appendix A.4). A first choice of weights are the density ratio weights (DW) ω_k^{DW}:

$$e(x) = \sum_{k=1}^{K} \underbrace{\rho_k \frac{f_k(x)}{f(x)}}_{=\omega_k^{\mathrm{DW}}(x)} e_k(x), \tag{3}$$

where f_k and f are the covariate densities locally and globally. Similar weights are used in transportability methods that reweight data to match a target population (Han et al., 2023; 2025; Guo et al., 2024), though here the objective is to recover the global propensity score across the super-population defined by the K participating sites. Unfortunately, DW estimation in a federated setting requires modeling the f_k's, which entails strong distributional assumptions and becomes challenging in high dimensions.

Instead, we propose the *Membership Weights (MW)* ω_k^{MW}:

$$e(x) = \sum_{k=1}^{K} \underbrace{\mathbb{P}(H = k \mid X = x)}_{=\omega_k^{\mathrm{MW}}(x)} e_k(x), \tag{4}$$

which represent the probability of site membership given the covariates. Unlike DW, MW do not require explicit density modeling and can be estimated directly via federated parametric (e.g., logistic regression, neural networks) or non-parametric (e.g., gradient-boosted trees) classification models, providing a flexible, communication-efficient alternative and making it the preferred choice for federated settings. We refer to Section 3.2 for more details on the federated estimation of MW and its advantages over DW.

Equations 3 and 4 enable combining locally estimated propensity scores e_k into a global propensity score using globally learned weights $\omega_k(x)$. Building on this decomposition, we define our oracle Federated IPW and AIPW estimators (Fed-(A)IPW). We define the *Oracle federated estimators* as follows:

$$\hat{\tau}^{\mathrm{fed}^*}_{\mathrm{IPW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{\mathrm{fed}(k)}_{\mathrm{IPW}}, \qquad \hat{\tau}^{\mathrm{fed}^*}_{\mathrm{AIPW}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\tau}^{\mathrm{fed}(k)}_{\mathrm{AIPW}}, \tag{5}$$

where $\hat{\tau}^{\mathrm{fed}(k)}_{\mathrm{IPW}} = \frac{1}{n_k} \sum_{i=1}^{n_k} \tau_{\mathrm{IPW}}(X_i; e)$ and $\hat{\tau}^{\mathrm{fed}(k)}_{\mathrm{AIPW}} = \frac{1}{n_k} \sum_{i=1}^{n_k} \tau_{\mathrm{AIPW}}(X_i; e, \mu_1, \mu_0)$ rely on the global propensity score $e(X) = \sum_{k=1}^{K} \omega_k(X) e_k(X)$.

Theorem 3 (proved in Appendix A.5) establishes that, in the oracle setting, Fed-(A)IPW estimators attain the same efficiency as their centralized counterparts.

Theorem 3. *Under Assumptions 1, 2, and 3, the oracle federated estimators (Equation 5) are identical to the oracle centralized estimators (Equation 1).*

Theorem 4 (proved in Appendix A.6) further shows that even when local overlap (Assumption 4) holds, federated estimators have lower variance than meta-analysis estimators.

Theorem 4. *Under Assumptions 1, 2 and 4, we have:*

$$\mathbb{V}[\hat{\tau}^*_{\mathrm{IPW}}] = \mathbb{V}[\hat{\tau}^{\mathrm{fed}^*}_{\mathrm{IPW}}] \leq \mathbb{V}[\hat{\tau}^{\mathrm{meta}^*}_{\mathrm{IPW}}], \qquad \mathbb{V}[\hat{\tau}^*_{\mathrm{AIPW}}] = \mathbb{V}[\hat{\tau}^{\mathrm{fed}^*}_{\mathrm{AIPW}}] \leq \mathbb{V}[\hat{\tau}^{\mathrm{meta}^*}_{\mathrm{AIPW}}],$$

with equality when the local propensity scores are identical across sites.

This variance reduction arises for two reasons. First, decomposing e as a weighted sum of e_k's marginalizes over $H \mid X$, eliminating unnecessary adjustment for site membership and thereby reducing variance. Second, our federated approach improves overlap compared with meta-analysis, as formalized below.

Theorem 5 (Overlap improvement)**.** $0 \leq \mathcal{O}_{\mathrm{global}} \leq \sum_{k=1}^{K} \rho_k \mathcal{O}_k$.

Theorem 5 (proved in Appendix A.7) shows that global overlap is always at least as good as the worst local overlap. Even when local overlap holds, sites with poor overlap benefit from the federated approach because the global score $e(x)$ is more bounded away from 0 and 1 than the local scores $\{e_k(x)\}_{k\in[K]}$. Notably, sites with poor individual overlap can even improve the overall overlap of the federated population, as illustrated in the following example.

Example. Let $K = 2$ with $X_i = 1$ in both sites, $\mathbb{P}(H_i \mid X_i) = 0.5$, $e_1(X_i) = 0.99$ and $e_2(X_i) = 0.01$, leading to $e(X_i) = \sum_{k=1}^{2} 0.5 \times e_k(X_i) = 0.5$. Local overlaps are poor, $\mathcal{O}_1 = \mathcal{O}_2 = (0.99 \times 0.01)^{-1} \approx 101$, whereas the global overlap is $\mathcal{O}_{\mathrm{global}} = (0.5 \times 0.5)^{-1} = 4$—the optimal value achieved in a randomized trial with 50% treatment probability. This illustrates how heterogeneity in treatment assignments can enhance global overlap and enable more robust causal inference.

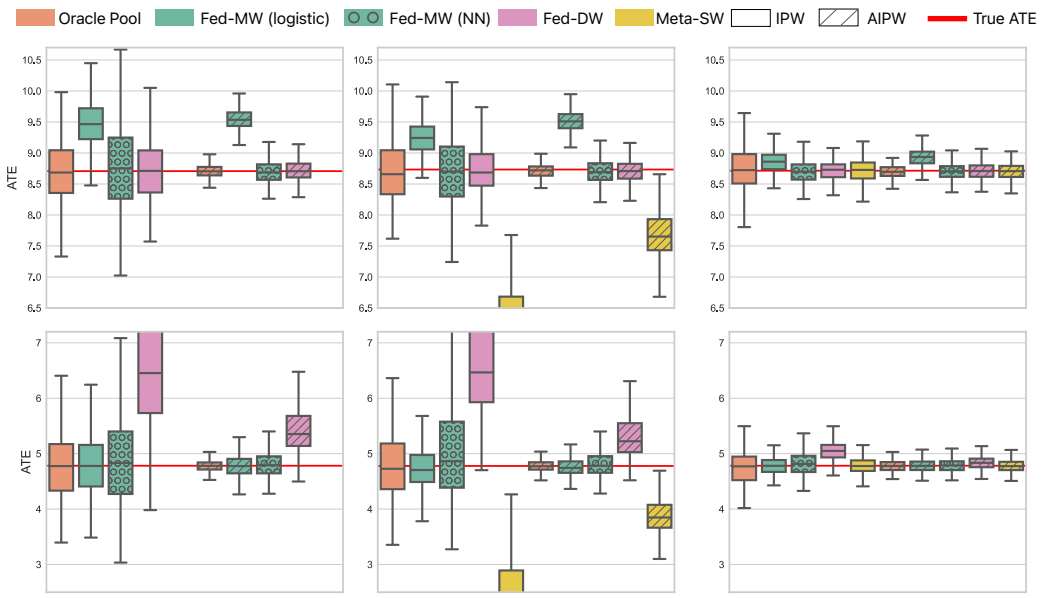
3.2 Federated estimation

We now move beyond oracle estimators and describe how to implement our Fed-(A)IPW estimators in a practical federated learning setting. Constructing the global score propensity score requires two steps, which can be executed in parallel: each site k estimates and shares a local propensity score $\hat{e}_k(x)$; and the sites collaboratively estimate federated weights $\{\hat{\omega}_k(x)\}_{k\in[K]}$. Fed-AIPW adds a third step to train outcome models $\hat{\mu}_0, \hat{\mu}_1$ via federated learning. We detail how to estimate $\{\hat{e}_k(x), \hat{\omega}_k(x)\}_k$ and $\hat{\mu}_0, \hat{\mu}_1$ below.

Local propensity scores. Each e_k can be estimated using any probabilistic binary classifier, either parametric (e.g., logistic regression or neural networks) or non-parametric ((e.g., generalized random forests, Lee et al., 2010). A key advantage of our approach is flexibility: sites can use different estimation methods tailored to local data or computational constraints. It also does *not* require Assumption 4: local scores may approach 0 or 1 provided the *global* score remains bounded away from these extremes.This, in particular, enables the integration of external control arms (FDA, 2023; EMA, 2023), where some sites have $e_k(X) = 0$ for all control patients yet still contribute to the global analysis.

Federated weights: density ratio vs. membership. Parametric density ratio weights $\omega_k^{\mathrm{DW}}(x) = \rho_k \frac{f_k(x)}{f(x)}$ can be implemented in a one-shot fashion by sharing local density parameters with the server, which then reconstructs the global mixture and computes the weights. A common choice is to assume parametric covariate distributions (say, Gaussian), estimate $(\hat{\mu}_k, \hat{\Sigma}_k)$ locally, and transmit them once to the server. This requires to communicate $O(Kd^2)$ parameters and is highly sensitive to model misspecification—an issue that becomes critical in high dimensions. Nonparametric density estimation would relax these assumptions but is statistically inefficient and does not yet have practical federated implementations.

In contrast, our membership weights $\omega_k^{\mathrm{MW}}(x) = \mathbb{P}(H = k \mid X = x)$ can be learned with any probabilistic multiclass classifier trained federatively—for example, logistic regression for simplicity and interpretability, or neural networks to capture complex nonlinearities. Such models are readily supported by modern FL algorithms and software libraries. Using the standard FedAvg algorithm (McMahan et al., 2017) requires exchanging TKP floats (training rounds $\times$ sites $\times$ model parameters), which is feasible for models of practical size and a large number of sites. Non-parametric



(a) No local overlap (b) Poor local overlap (c) Good local overlap

Figure 2: Synthetic data: DGP A (top) and DGP B (bottom).

classifiers like random forests (Hauschild et al., 2022) and gradient-boosted trees (Li et al., 2020) have also been adapted to the federated setting.

Outcome models. To construct the doubly robust Fed-AIPW estimator, μ_0, μ_1 are trained federatively, as in Khellaf et al. (2025). As for MW, standard FL algorithms and librairies can be used to train a wide range of parametric and non-parametric supervised learning models.

Remark (Missing values). Our approach can naturally accommodate missing data. Local propensity scores can be estimated consistently with logistic regression or random forests (Jiang et al., 2020; Josse et al., 2024). For membership weights, constant imputation combined with federated random forests provides a consistent solution (Le Morvan et al., 2021), whereas density ratio weights would require adapting more complex federated EM algorithms (Dieuleveut et al., 2021; Marfoq et al., 2021). Finally, doubly robust estimators with missing data can be obtained via non-parametric federated outcome models (Mayer et al., 2020).

4 Experiments

Synthetic data. We consider $K = 3$ sites and $d = 10$ covariates. Two data-generating processes (DGPs) are used. In *DGP A*, each site k independently samples $n_k = 650$ individuals from a site-specific multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$. In *DGP B*, a total of $n = 4000$ individuals are first drawn from a bimodal Gaussian mixture and then assigned to sites according to a multinomial logistic model based on their covariates. We vary within-site overlap to mimic different practical scenarios: *No local overlap* ($\mathcal{O}_2 = +\infty$, the second site has no treated individuals), *Poor local overlap* ($\mathcal{O}_2 \approx 10^7$), and *Good local overlap* ($\mathcal{O}_2 \approx 4.6$). The outcome models μ_1, μ_0 are shared across sites and specified as polynomial functions with interactions. For comparison, we also generate data consistent with the setting in Xiong et al. (2023) (Figure 3). All results are averaged over 800 simulation runs; full details are provided in Appendix C.

We evaluate our proposed **Fed-IPW** and **Fed-AIPW** using the **MW** weights estimated either via federated multinomial logistic regression (well specified in *DGP B*) or via a two-layer Neural Network (NN). We compare these methods against several competitors: **Fed-IPW** and **Fed-AIPW** with the alternative **DW** weights based on Gaussian density estimation (well specified in *DGP A*); the **Centralized Oracle** (Def. 1); meta-analysis IPW/AIPW with sample-size weighting **(Meta-**

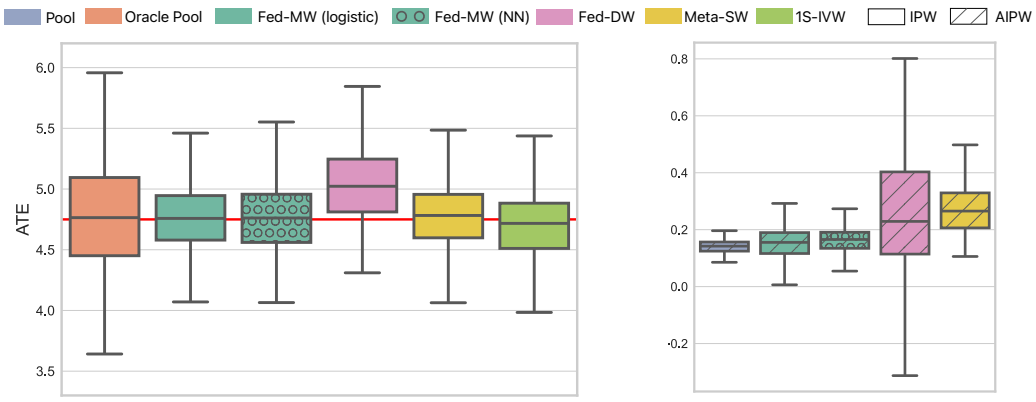


Figure 3: Comparison to Xiong et al. (2023) (IPW).

Figure 4: AIPW estimates on Traumabase.

SW) (Def. 2); and the one-shot inverse-variance weighted IPW estimator **(1S-IVW)** of Xiong et al. (2023), evaluated under favorable conditions with shared propensity-score parameters across sites. For all estimators, propensity scores are fit via logistic regression and outcome models via linear regression, implying misspecification in the latter.

Figures 2a–2c (top: *DGP A*; bottom: *DGP B*) summarize the results under the three overlap regimes. Before discussing each setting, we highlight several general observations. First, Fed-IPW-DW is unbiased only under *DGP A*, where the Gaussian specification holds, and exhibits bias under *DGP B*. By contrast, Fed-IPW-MW adapts across data-generating processes: with logistic membership weights it is unbiased in *DGP B*, and with a more flexible neural network classifier it is unbiased for both *DGP A* and *DGP A*, reflecting the greater modeling flexibility of MW relative to DW's restrictive density specification. Second, Fed-AIPW enjoys its doubly robust property and remains unbiased across all overlap levels by double robustness; despite misspecified linear outcome models, accurate propensity estimation ensures consistency. Finally, Fed-IPW typically attains lower variance than the centralized oracle IPW, consistent with well-documented efficiency gains from using estimated propensity scores (Hirano et al., 2003).

In the *No local overlap* setting (Figure 2a), meta-analysis estimators are undefined because one site has no treated individuals. Our federated estimators remain unbiased under both DGPs, as Assumption 3 holds (global overlap $\mathcal{O}_{\text{global}} \approx 6.22$). In the *Poor local overlap* setting (Figure 2b), site 2 exhibits weak overlap, leading to bias and instability in meta-analysis estimators, including Meta-AIPW, as both the propensity scores and local outcome models are inaccurate. This issue is mitigated in the global dataset (see Figure 5 in the Appendix), allowing our federated estimators to remain reliable. In the *Good local overlap* setting (Figure 2c), all methods are unbiased, but federated estimators achieve the smallest variance.

Figure 3 considers a setting where all local propensity scores share a common subset of 5 of 10 logistic regression coefficients with data generated from *DGP B*. This setup matches the assumptions of the 1S-IVW method of Xiong et al. (2023), which relies on prior knowledge of the shared parameters to aggregate them—an assumption not required by our method. Fed-MW remains unbiased and attains the lowest variance, matching that of 1S-IVW, even without access to the additional information about shared parameters.

Additional simulations provided in Appendix E examine scenarios with more sites (Table 4), nonparametric estimation (Table 5), and local propensity model misspecification (Table 6), further confirming the robustness of our approach.

Real data. We analyze the multi-site Traumabase cohort (Mayer et al., 2020; Colnet et al., 2024) to estimate the effect of tranexamic acid on mortality across $K = 14$ centers. The local datasets are highly imbalanced in site sizes (106 to 2,092 patients) and treatment arms (e.g., site 11: 4 treated vs. 121 controls); there are 17 covariates and $n = 8{,}248$ patients in total, of whom 638 were treated. Covariates are standardized federatively by sharing site means and variances. We focus on AIPW estimators: local propensities e_k are estimated via logistic regression; outcome

models μ_1, μ_0 are trained with FedAvg logistic regression (5,000 rounds, 1 local epoch, step size $\eta = 0.1$); MW are learned with a FedAvg neural network (one hidden layer, 128 units), while DW are based on Gaussian density estimation. Competitors include Meta-SW (same local models, computed only on sites with enough treated units—where the number of treated observations exceeds the covariate dimension) and a centralized AIPW benchmark obtained using the R package `grf`'s `probability_forest` function (Tibshirani et al., 2018) on the pooled data. Empirical confidence intervals are constructed from 150 bootstrap resamples.

Figure 4 shows that our federated estimators closely match the centralized benchmark and exhibit lower variance than Meta-SW, which in this application departs from the centralized estimate. Among federated methods, Fed-MW is closest to the nonparametric centralized estimator—reflecting its more flexible modeling of propensities and membership—whereas Fed-DW is less stable, likely due to noisy per-site covariance estimates for a 17-dimensional Gaussian ($\approx 17^2$ parameters) in small sites.

5 Conclusion, Limitations and Future Work

We propose a theoretically grounded framework for federated causal inference that leverages membership weights to construct valid pseudo-populations across silos. These weights, estimated via flexible parametric or nonparametric models, improve overlap and yield more stable ATE estimates, without sharing raw data. Our framework accommodates heterogeneous local propensity score estimation strategies, supports external control arms, and remains robust to even extreme local treatment–control imbalances. Although sufficiently large per-site datasets are still needed—particularly in high-dimensional settings—our approach is especially well suited to a moderate number of large silos, where FL most effectively increases effective sample size.

Promising avenues for future work include principled handling of site effects (i.e., relaxing Assumption 2) and extending our framework to CATE estimation. While the ATE remains central in econometrics, biomedicine, and public policy, considering CATE is important for moving towards personalization. Our work lays a foundation for federated CATE estimation: most learners (T-, S-, X-, and R-learners) rely on nuisance quantities such as propensity scores, and our federated estimation framework could be incorporated into these methods, although further work is needed to assess its formal properties and practical performance.

REFERENCES

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Clément Berenfeld, Ahmed Boughdiri, Bénédicte Colnet, Wouter AC van Amsterdam, Aurélien Bellet, Rémi Khellaf, Erwan Scornet, and Julie Josse. Causal meta-analysis: Rethinking the foundations of evidence-based medicine. *arXiv preprint arXiv:2505.20168*, 2025.

Danielle L Burke, Joie Ensor, and Richard D Riley. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in medicine*, 36(5):855–875, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Oscar Clivio, David Bruns-Smith, Avi Feller, and Christopher C Holmes. Towards principled representation learning to improve overlap in treatment effect estimation. In *9th Causal Inference Workshop at UAI 2024*, 2024.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Geneviève Robin. Federated expectation maximization with heterogeneity mitigation and variance reduction. *CoRR*, abs/2111.02083, 2021. URL https://arxiv.org/abs/2111.02083.

EMA. Reflection paper on establishing efficacy based on single arm trials submitted as pivotal evidence in a marketing authorisation, 2023. European Medicines Agency.

EMA. ICH E9 Statistical Principles for Clinical Trials: Scientific Guideline, 2024.

FDA. Adjusting for covariates in randomized clinical trials for drugs and biological products, 2023.

FDA. Considerations for the design and conduct of externally controlled trials for drug and biological products. https://www.fda.gov/media/164960/download, 2023. Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research and Oncology Center of Excellence, Food and Drug Administration.

David A Grimes and Kenneth F Schulz. Bias and causal associations in observational research. *The lancet*, 359(9302):248–252, 2002.

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

Zijian Guo, Xiudi Li, Larry Han, and Tianxi Cai. Robust inference for federated meta-learning. *Journal of the American Statistical Association*, pp. 1–16, 2025.

Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36:70453–70482, 2023.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, (just-accepted): 1–25, 2025.

Anne-Christin Hauschild, Marta Lemanczyk, Julian Matschinske, Tobias Frisch, Olga Zolotareva, Andreas Holzinger, Jan Baumbach, and Dominik Heider. Federated random forests can improve local performance of predictive models for various healthcare applications. *Bioinformatics*, 38 (8):2278–2286, 2022.

Miguel A Hernán. The c-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, 108(5):616–619, 2018.

Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Wei Jiang, Julie Josse, Marc Lavielle, TraumaBase Group, et al. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

Julie Josse, Jacob M Chen, Nicolas Prost, Gaël Varoquaux, and Erwan Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479, 2024.

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.

Rémi Khellaf, Aurélien Bellet, and Julie Josse. Federated Causal Inference: Multi-Study ATE Estimation beyond Meta-Analysis. In *AISTATS*, 2025.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

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

Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky and. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018a. doi: 10.1080/01621459.2016.1260466. URL https://doi.org/10.1080/01621459.2016.1260466.

Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*, 188(1):250–257, 09 2018b. ISSN 0002-9262. doi: 10.1093/aje/kwy201. URL https://doi.org/10.1093/aje/kwy201.

Qinbin Li, Zeyi Wen, and Bingsheng He. Practical federated gradient boosting decision trees. In *AAAI*, pp. 4642–4649, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Othmane Marfoq, Giovanni Neglia, AurÃ©lien Bellet, Laetitia Kameni, and Richard Vidal. Federated Multi-Task Learning under a Mixture of Distributions. In *NeurIPS*, 2021.

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*, pp. 1273–1282. PMLR, 2017.

Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, Satrajit Roychoudhury, Kenneth Koury, Ping Li, Warren V. Kalina, David Cooper, Robert W. Frenck, Laura L. Hammitt, Özlem Türeci, Haylene Nell, Axel Schaefer, Serhat Ünal, Dina B. Tresnan, Susan Mather, Philip R. Dormitzer, Uğur Şahin, Kathrin U. Jansen, and William C. Gruber. Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615, 2020. doi: 10.1056/NEJMoa2034577. URL https://www.nejm.org/doi/full/10.1056/NEJMoa2034577.

Richard D Riley, Joie Ensor, Miriam Hattle, Katerina Papadimitropoulou, and Tim P Morris. Two-stage or not two-stage? that is the question for ipd meta-analysis projects. *Research Synthesis Methods*, 14(6):903–910, 2023.

Sarah E Robertson, Jon A Steingrimsson, Nina R Joyce, Elizabeth A Stuart, and Issa J Dahabreh. Center-specific causal inference with multicenter trials: reinterpreting trial evidence in the context of each participating center. *arXiv preprint arXiv:2104.05905*, 2021.

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.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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.

Jerzy Splawa-Neyman. 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.

Julie Tibshirani, Susan Athey, Rina Friedberg, Vitor Hadad, David Hirshberg, Luke Miner, Erik Sverdrup, Stefan Wager, Marvin Wright, and Maintainer Julie Tibshirani. Package 'grf'. *Comprehensive R Archive Network*, 2018.

Tyler J VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34: 211–219, 2019.

Thanh Vinh Vo, Young Lee, Trong Nghia Hoang, and Tze-Yun Leong. Bayesian federated estimation of causal effects from observational data. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2024–2034. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/vo22a.html.

Stefan Wager. Causal inference: A statistical learning approach, 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.

Changchang Yin, Hong-You Chen, Wei-Lun Chao, and Ping Zhang. Federated inverse probability treatment weighting for individual treatment effect estimation. *arXiv preprint arXiv:2503.04946*, 2025.