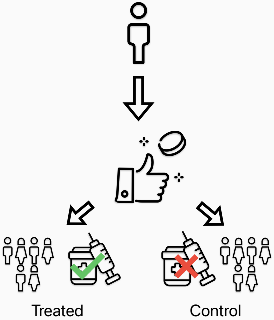
Federated Causal Inference: ATE Estimation from Multi-Site Observational Data via Propensity Score Aggregation

Rémi Khellaf, Aurélien Bellet and Julie Josse (INRIA, Montpellier)

Federated causal inference

Goal of causal inference: measure the effect of a treatment on an outcome

Randomized Controlled Trials (RCTs):

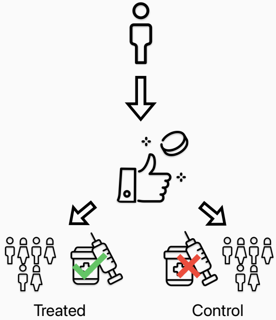


+ : direct causal association
− : limited scope (eligibility criteria), small sample sizes, not always feasible

Federated causal inference

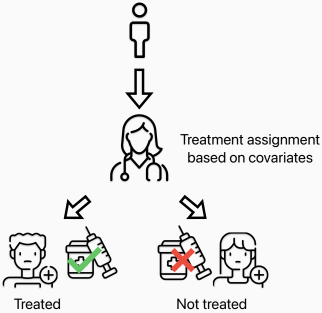
Goal of causal inference: measure the effect of a treatment on an outcome

Randomized Controlled Trials (RCTs):



+ : direct causal association
− : limited scope (eligibility criteria), small sample sizes, not always feasible

Observational Data:

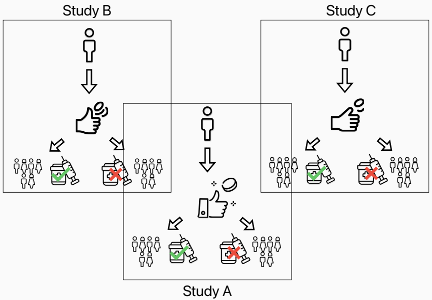


+ : abundant, large scope, always available
− : naturally scattered across sites (e.g., hospitals), confounding factors

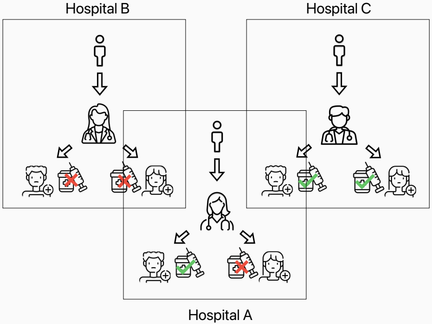
Federated causal inference

Multi-source causal inference: higher validity and generalization

Randomized Controlled Trials (RCTs):



Observational Data:



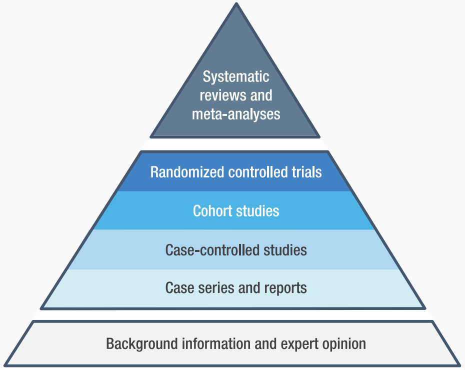
+ : direct causal association
− : limited scope (eligibility criteria), small sample sizes, not always feasible

+ : abundant, large scope, always available
− : naturally scattered across sites (e.g., hospitals), confounding factors

Classic approach: Meta-analysis

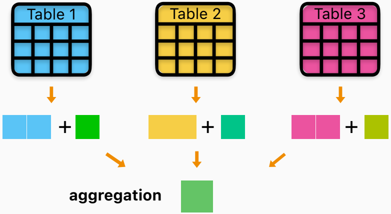
Meta-analysis (MA) combines effects from multiple studies

It is at the top of the evidence hierarchy



Classic approach: Meta-analysis

Meta-analysis (MA) combines effects from multiple studies on:

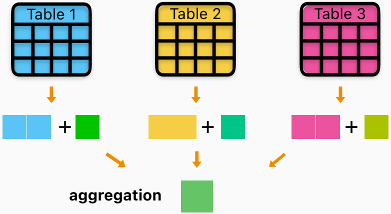


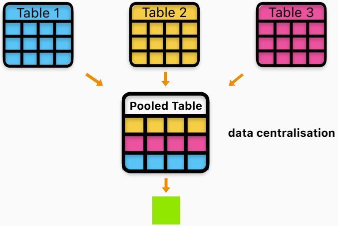
Aggregated Data (AD):

- Studies report summary statistics + effect sizes which are aggregated into a single one.
- **Limitation**: Prone to **ecological bias**.

Classic approach: Meta-analysis

Meta-analysis (MA) combines effects from multiple studies on:





Aggregated Data (AD):

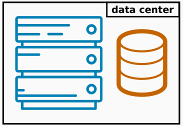
- Studies report summary statistics + effect sizes which are aggregated into a single one.
- **Limitation**: Prone to **ecological bias**.

Individual Patient Data (IPD):

- Studies' data are pooled together before causal analysis.
- **Limitation**: Harder to share individual data

Enabling individual patient data analysis with federated learning

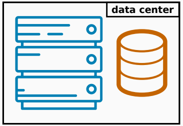
IPD cannot always be pooled altogether



- Data may be too sensitive to share: personal data regulations (GDPR, HIPAA...), no consent and release agreement during data collection
- Parties may have competitive concerns (e.g., pharmaceutical companies performing costly RCTs)

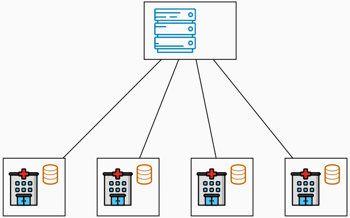
Enabling individual patient data analysis with federated learning

IPD cannot always be pooled altogether



- Data may be too sensitive to share: personal data regulations (GDPR, HIPAA...), no consent and release agreement during data collection
- Parties may have competitive concerns (e.g., pharmaceutical companies performing costly RCTs)

Federated Learning enables IPD analysis without pooling



- Client-server architecture enabling collaborative learning without sharing individual data
- Recent framework with strong theoretical guarantees [Kairouz et al., 2021]
- Encompasses privacy (e.g., differential privacy) and security concerns (e.g., adversarial attacks)

Going beyond meta-analysis with federated causal inference

Our work bridges causal inference and federated learning [Kairouz et al., 2021] to better estimate average treatment effects from decentralized data sources

1. We consider several estimators with varying communication costs
2. We study their statistical performance under various types of data heterogeneity
3. We validate on numerical experiments and provide guidelines for practitioners

---

[1] **R.K.**, A. Bellet, and J. Josse. "Federated Causal Inference: Multi-Centric ATE Estimation beyond Meta-Analysis." AISTATS (2024).

[2] **R.K.**, A. Bellet, and J. Josse. "Federated Causal Inference from Multi-Site Observational Data via Propensity Score Aggregation." Arxiv (2025).

Going beyond meta-analysis with federated causal inference

Our work bridges causal inference and federated learning [Kairouz et al., 2021] to better estimate average treatment effects from decentralized data sources

1. We consider several estimators with varying communication costs
2. We study their statistical performance under various types of data heterogeneity
3. We validate on numerical experiments and provide guidelines for practitioners

Multiple RCTs[1]: compares meta-analysis, one-shot and multi-shot FL

---

[1] **R.K.**, A. Bellet, and J. Josse. "Federated Causal Inference: Multi-Centric ATE Estimation beyond Meta-Analysis." AISTATS (2024).

[2] **R.K.**, A. Bellet, and J. Josse. "Federated Causal Inference from Multi-Site Observational Data via Propensity Score Aggregation." Arxiv (2025).

Going beyond meta-analysis with federated causal inference

Our work bridges causal inference and federated learning [Kairouz et al., 2021] to better estimate average treatment effects from decentralized data sources

1. We consider several estimators with varying communication costs
2. We study their statistical performance under various types of data heterogeneity
3. We validate on numerical experiments and provide guidelines for practitioners

Multiple RCTs[1]: compares meta-analysis, one-shot and multi-shot FL

Multiple sites with observational data[2]: focuses on the federation of heterogeneous propensity scores to estimate the ATE

[1] **R.K.**, A. Bellet, and J. Josse. "Federated Causal Inference: Multi-Centric ATE Estimation beyond Meta-Analysis." AISTATS (2024).

[2] **R.K.**, A. Bellet, and J. Josse. "Federated Causal Inference from Multi-Site Observational Data via Propensity Score Aggregation." Arxiv (2025).

Related work in Federated Causal Inference

- **Multicentric framework**: IPD meta-analysis offers clear advantages over AD, especially when local datasets are small[3,4]

[3] Riley, Richard D., et al. "Two-stage or not two-stage? That is the question for IPD meta-analysis projects." Research synthesis methods 14.6 (2023)

[4] Robertson, Sarah E., et al. "Center-specific causal inference with multicenter trials: reinterpreting trial evidence in the context of each participating center." arXiv (2021)

Related work in Federated Causal Inference

- **Multicentric framework**: IPD meta-analysis offers clear advantages over AD, especially when local datasets are small
- **Federation of model parameters**: outcome and propensity score models can be federated[3,4], but it is unclear how the subsequent ATE estimators compare to meta-analysis on AD.

[3] Xiong, Ruoxuan, et al. "Federated causal inference in heterogeneous observational data." Statistics in Medicine (2023)

[4] Vo, Thanh Vinh, and Tze-Yun Leong. "Federated Causal Inference from Observational Data." arXiv (2023)

Related work in Federated Causal Inference

- **Multicentric framework**: IPD meta-analysis offers clear advantages over AD, especially when local datasets are small
- **Federation of model parameters**: outcome and propensity score models can be federated, but it is unclear how the subsequent ATE estimators compare to meta-analysis on AD.
- **Generalization**: transferring ATE estimates from multiple source sites to a target domain can be done with density ratio weighting method[3]. Their approach resembles meta-analysis, relying on aggregate statistics rather than individual-level data

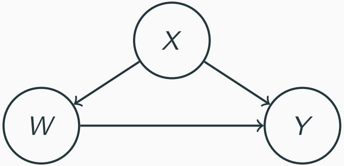
[3] Han, Larry, et al. "Federated adaptive causal estimation (face) of target treatment effects." Journal of the American Statistical Association (2025)

Problem Setting: Observational Data

Reminder: Classic centralized framework with observational data

- Goal: estimate effect of treatment W on outcome Y given covariates X
- Average Treatment Effect (ATE) measured as a risk difference $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$
- Confounded: account for either $\mathbb{P}(W_i = 1 \mid X_i) = e(X_i)$ or $\mathbb{E}(Y_i \mid W_i, X_i) = \mu_{W_i}(X_i)$

Obs.	Covariates			Treatment	Outcome	Potential Outcomes	
i	X_1	X_2	X_3	W	Y	$Y^{(1)}$	$Y^{(0)}$
1	2.3	1.5	M	1	3.2	3.2	??
2	2.2	3.1	F	0	2.8	??	2.8
3	3.5	2.0	F	1	2.1	2.1	??
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	3.7	2.0	F	0	2.8	??	2.8
n	2.5	1.7	M	1	3.2	3.2	??



Reminder: Classic centralized framework with observational data

- Goal: estimate effect of treatment W on outcome Y given covariates X
- Average Treatment Effect (ATE) measured as a risk difference $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$
- Confounded: account for either $\mathbb{P}(W_i = 1 \mid X_i) = e(X_i)$ or $\mathbb{E}(Y_i \mid W_i, X_i) = \mu_{W_i}(X_i)$

Classic (oracle) centralized ATE estimators

Inverse Propensity Weighting (IPW):

$$\hat{\tau}^{*}_{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right)$$

Reminder: Classic centralized framework with observational data

- Goal: estimate effect of treatment W on outcome Y given covariates X
- Average Treatment Effect (ATE) measured as a risk difference $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$
- Confounded: account for either $\mathbb{P}(W_i = 1 \mid X_i) = e(X_i)$ or $\mathbb{E}(Y_i \mid W_i, X_i) = \mu_{W_i}(X_i)$

Classic (oracle) centralized ATE estimators

Inverse Propensity Weighting (IPW):

$$\hat{\tau}^*_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)$$

Augmented IPW (AIPW):

$$\hat{\tau}^*_{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{W_i\left(Y_i - \mu_1(X_i)\right)}{e(X_i)} - \frac{(1-W_i)\left(Y_i - \mu_0(X_i)\right)}{1-e(X_i)} + \mu_1(X_i) - \mu_0(X_i)\Big)$$

Reminder: Classic centralized framework with observational data

- Goal: estimate effect of treatment W on outcome Y given covariates X
- Average Treatment Effect (ATE) measured as a risk difference $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$
- Confounded: account for either $\mathbb{P}(W_i = 1 \mid X_i) = e(X_i)$ or $\mathbb{E}(Y_i \mid W_i, X_i) = \mu_{W_i}(X_i)$

Classic (oracle) centralized ATE estimators

Inverse Propensity Weighting (IPW):

$$\hat{\tau}^*_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)$$

Augmented IPW (AIPW):

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

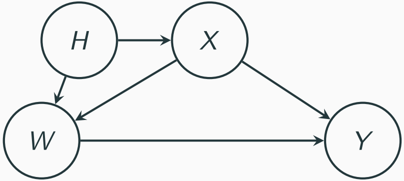
Assumptions for consistency:

- **Unconfoundedness**: $Y(0), Y(1) \perp\!\!\!\perp W \mid X$
- **Consistency**: $Y(w) = Y_i \mid W_i = w, X_i$
- **Bounded outcomes**
- **Overlap**: $\exists \eta > 0, \forall X_i \in \mathcal{X}, \eta < e(X_i) < 1 - \eta$

Our setting: multi-site decentralized observational data

- We consider K decentralized and potentially heterogeneous sites
- The goal is to estimate the ATE: $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right) = \sum_{k=1}^{K} \mathbb{P}(H = k)\tau_k$

Source	Obs.	Covariates			Treatment	Outcomes
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2



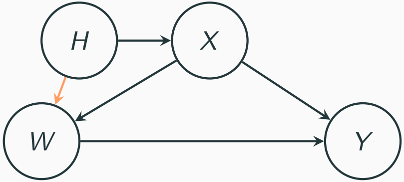
Our setting: multi-site decentralized observational data

- We consider K decentralized and potentially heterogeneous sites
- The goal is to estimate the ATE: $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right) = \sum_{k=1}^{K} \mathbb{P}(H = k)\tau_k$

Source	Obs.	Covariates			Treatment	Outcomes
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

Heterogeneity in **treatment allocations**

$e_k = \mathbb{P}(W_i \mid X_i, H_i = k)$



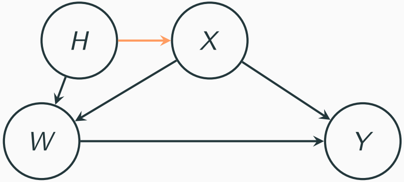
$\neq \{e_k\}_k$

Our setting: multi-site decentralized observational data

- We consider K decentralized and potentially heterogeneous sites
- The goal is to estimate the ATE: $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right) = \sum_{k=1}^{K} \mathbb{P}(H = k)\tau_k$

Source	Obs.	Covariates			Treatment	Outcomes
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

Heterogeneity in **covariates distribution**



$$X \mid H = k \nsim X \mid H = k'$$

Our setting: multi-site decentralized observational data

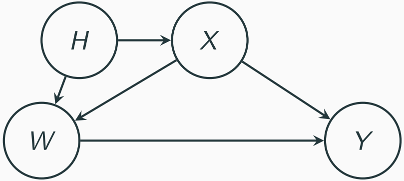
- We consider K decentralized and potentially heterogeneous sites
- The goal is to estimate the ATE: $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right) = \sum_{k=1}^{K} \mathbb{P}(H = k)\tau_k$

Source	Obs.	Covariates			Treatment	Outcomes
H	i	X_1	X_2	X_3	W	Y
1	1	2.3	1.5	M	1	3.2
1	2	2.2	3.1	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	4.5	5.0	F	1	4.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	1	3.7	2.0	F	0	2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	n_K	2.5	1.7	M	0	3.2

Constraint: cannot pool data

$\Rightarrow$ no access to e, μ_1, μ_0

$\Rightarrow$ cannot compute (A)IPW estimators



How to estimate τ without access to individual-level data?

Oracle Multi-Site ATE Estimators

Meta-Analysis

Baseline estimators: oracle meta-analysis

A **meta-analysis** estimator is a **weighted average of local estimates** $\{\hat{\tau}_k\}_k$, which are computed with local nuisance functions $e_k(X_i), \mu_{1,k}(X_i), \mu_{0,k}(X_i)$

$$\hat{\tau}^{\mathrm{meta}} = \sum_{k=1}^{K} \rho_k \hat{\tau}_k$$

with $\rho_k = \mathbb{P}(H_i = k) \approx \frac{n_k}{n}$ and

$$\hat{\tau}_k = \begin{cases} \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\mu_{1,k}(X_i) - \mu_{0,k}(X_i) + \frac{W_i(Y_i - \mu_{1,k}(X_i))}{e_k(X_i)} - \frac{(1 - W_i)(Y_i - \mu_{0,k}(X_i))}{1 - e_k(X_i)} \right) & \text{(AIPW)} \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{W_i Y_i}{e_k(X_i)} - \frac{(1 - W_i) Y_i}{1 - e_k(X_i)} \right) & \text{(IPW)} \end{cases}$$

Baseline estimators: oracle meta-analysis

A **meta-analysis** estimator is a **weighted average of local estimates** $\{\hat{\tau}_k\}_k$, which are computed with local nuisance functions $e_k(X_i), \mu_{1,k}(X_i), \mu_{0,k}(X_i)$

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

(Asymptotically consistent) $\hat{\tau}^{\text{meta}} \underset{n\to\infty}{\overset{p}{\to}} \tau$ if all local estimates are asymptotically consistent, i.e., $\forall k \in [K], \hat{\tau}_k \underset{n_k\to\infty}{\overset{p}{\to}} \tau_k$, which requires at each site k:

- Unconfoundedness, consistency, bounded potential outcomes
- Local overlap: $\exists \eta, \forall x \in \mathcal{X} \eta < e_k(x) < 1 - \eta$

$\Rightarrow$ forbids the inclusion of sites with no (un)treated individuals for some X_i (e.g. external control arms, systematic treatment rule...)

Oracle Multi-Site ATE Estimators

Federated Estimators

Federated estimators: introduction

- Principle: **decompartmentalize** the estimation of the causal effect, i.e., leverage individual-level data **without sharing raw data**

Federated estimators: introduction

- Principle: **decompartmentalize** the estimation of the causal effect, i.e., leverage individual-level data **without sharing raw data**
- We assume site ignorability: $(Y(0), Y(1)) \perp\!\!\!\perp H \mid X$

  $\Rightarrow$ common conditional outcome models $\{\mu_1, \mu_0\}$ across sites

  $\Rightarrow$ no centre effect: H is not a confounder, so learning $e(X)$ suffices to deconfound.

  - Can be relaxed with parametric modelling of the effect of H on Y and/or learning $e(X; H)$.

Federated estimators: introduction

- Principle: **decompartmentalize** the estimation of the causal effect, i.e., leverage individual-level data **without sharing raw data**
- We assume site ignorability: $(Y(0), Y(1)) \perp\!\!\!\perp H \mid X$
  - $\Rightarrow$ common conditional outcome models $\{\mu_1, \mu_0\}$ across sites
  - $\Rightarrow$ no centre effect: H is not a confounder, so learning $e(X)$ suffices to deconfound.
  - Can be relaxed with parametric modelling of the effect of H on Y and/or learning $e(X; H)$.
- We do not assume common treatment assignments $\{e_k\}_k$ across sites
  - highly flexible framework, can handle all kinds of heterogeneity in treatment allocations (not just intercept shift)
  - realistic setting: e.g., different hospitals may have different treatment protocols
  - if ready to make the assumption of homogeneity in $\{e_k\}_k$, $e(X)$ can be learned directly with a federated SGD algorithm

Federated estimators: propensity score decomposition

μ_1, μ_0 are common across sites $\Rightarrow$ can be learned with a federated SGD algorithm (see later)

Federated estimators: propensity score decomposition

μ_1, μ_0 are common across sites $\Rightarrow$ can be learned with a federated SGD algorithm (see later)

The propensity scores are heterogeneous across sites $\Rightarrow$ directly learning a global e is inefficient $\Rightarrow$ other learning strategies must be considered

Federated estimators: propensity score decomposition

μ_1, μ_0 are common across sites $\Rightarrow$ can be learned with a federated SGD algorithm (see later)

The propensity scores are heterogeneous across sites $\Rightarrow$ directly learning a global e is inefficient $\Rightarrow$ other learning strategies must be considered

Our method:
e in the pooled dataset decomposes as a weighted sum of the local ones:

$$e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$$

Federated estimators: propensity score decomposition

μ_1, μ_0 are common across sites $\Rightarrow$ can be learned with a federated SGD algorithm (see later)

The propensity scores are heterogeneous across sites $\Rightarrow$ directly learning a global e is inefficient $\Rightarrow$ other learning strategies must be considered

Our method:
e in the pooled dataset decomposes as a weighted sum of the local ones:

$$e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$$

$\Rightarrow$ learn **federation weights** $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ and **local propensity scores** $e_k(x) = \mathbb{P}(W_i = 1 \mid X_i = x, H_i = k)$

Federated estimators: oracle form

A federated estimator of the ATE is a **weighted average of local estimates** $\{\hat{\tau}_k^{\mathrm{fed}}\}_k$, which are computed with **global nuisance functions** e, μ_1, μ_0

$$\hat{\tau}^{\mathrm{fed}} = \sum_{k=1}^{K} \rho_k \hat{\tau}_k^{\mathrm{fed}}$$

with $\rho_k = \mathbb{P}(H_i = k) \approx \frac{n_k}{n}$ and

$$\hat{\tau}_k = \begin{cases} \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\mu_1(X_i) - \mu_0(X_i) + \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} \right) & \text{(AIPW)} \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right) & \text{(IPW)} \end{cases}$$

Federated estimators: oracle form

A federated estimator of the ATE is a **weighted average of local estimates** $\{\hat{\tau}_k^{\text{fed}}\}_k$, which are computed with **global nuisance functions** e, μ_1, μ_0

$$\hat{\tau}^{\text{fed}} = \sum_{k=1}^{K} \rho_k \hat{\tau}_k^{\text{fed}}$$

(Asymptotically consistent) $\hat{\tau}^{\text{fed}} \xrightarrow[n \to \infty]{p} \tau$ if globally hold:

- Unconfoundedness, consistency, bounded potential outcomes
- Global overlap: $\exists \eta, \forall x \in \mathcal{X}, \eta < e(x) < 1 - \eta$

$\Rightarrow$ allows the inclusion of sites with no (un)treated individuals for some X_i, as long as other sites cover them

Theoretical results and numerical illustrations

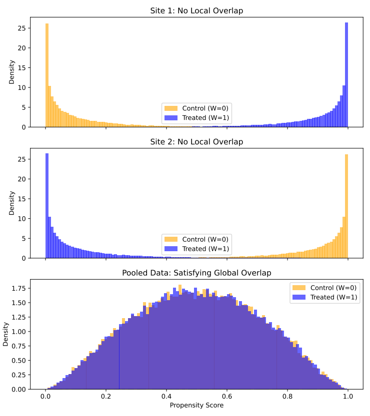
Assuming global overlap:

- $\hat{\tau}^{\text{fed}^*} = \hat{\tau}^{\text{pool}^*}$

Theoretical results and numerical illustrations

Assuming global overlap:

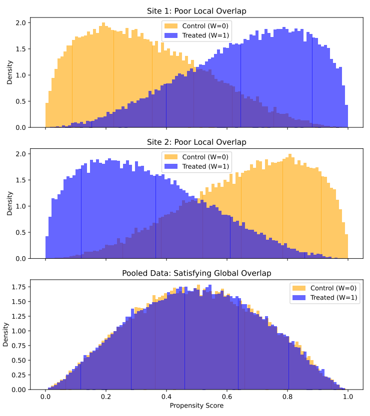
- $\hat{\tau}^{\text{fed}^*} = \hat{\tau}^{\text{pool}^*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}^*}$, only $\hat{\tau}^{\text{fed}^*}$



Theoretical results and numerical illustrations

Assuming global overlap:

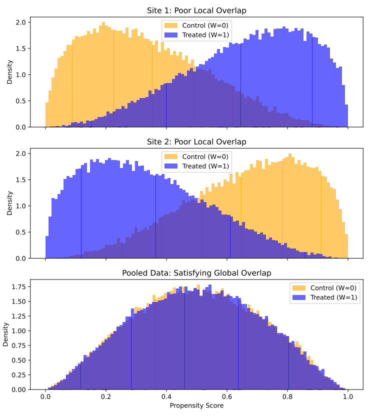
- $\hat{\tau}^{\text{fed}*} = \hat{\tau}^{\text{pool}*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}*}$, only $\hat{\tau}^{\text{fed}*}$
- If local overlap **at every site**:



Theoretical results and numerical illustrations

Assuming global overlap:

- $\hat{\tau}^{\text{fed}*} = \hat{\tau}^{\text{pool}*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}*}$, only $\hat{\tau}^{\text{fed}*}$
- If local overlap **at every site**:
  - $\hat{\tau}^{\text{meta}*}$ can be computed too

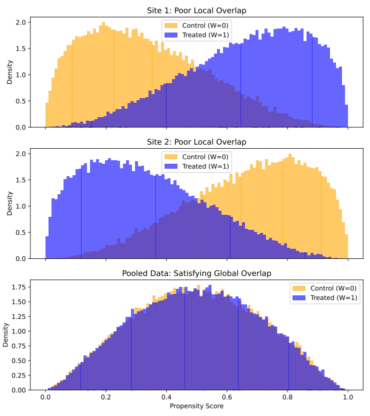


Theoretical results and numerical illustrations

Assuming global overlap:

- $\hat{\tau}^{\text{fed}^*} = \hat{\tau}^{\text{pool}^*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}^*}$, only $\hat{\tau}^{\text{fed}^*}$
- If local overlap **at every site**:
    - $\hat{\tau}^{\text{meta}^*}$ can be computed too
    - The **global overlap is always "better"** than the worst local ones: $\mathcal{O}_{\text{global}} \leq \sum_{k=1}^{K} \rho_k \mathcal{O}_k$

Overlap at site k: $\mathcal{O}_k = \mathbb{E}_{X \sim P_k}\left[1/(e_k(X_i)(1 - e_k(X_i)))\right]$
Global overlap: $\mathcal{O}_{\text{global}} = \mathbb{E}_{X \sim P}\left[1/(e(X_i)(1 - e(X_i)))\right]$



Theoretical results and numerical illustrations

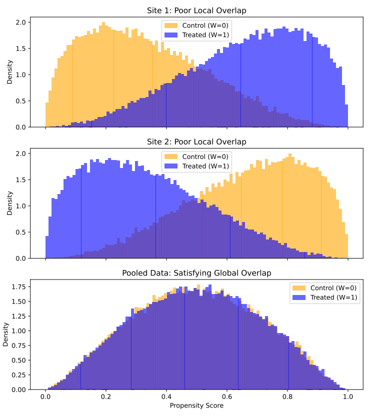
Assuming global overlap:

- $\hat{\tau}^{\text{fed}*} = \hat{\tau}^{\text{pool}*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}*}$, only $\hat{\tau}^{\text{fed}*}$
- If local overlap **at every site**:
  - $\hat{\tau}^{\text{meta}*}$ can be computed too
  - The **global overlap is always "better"** than the worst local ones: $\mathcal{O}_{\text{global}} \leq \sum_{k=1}^{K} \rho_k \mathcal{O}_k$

$\Rightarrow$ Improved stability of $\hat{\tau}^{\text{fed}*}$ over $\hat{\tau}^{\text{meta}*}$

Overlap at site k: $\mathcal{O}_k = \mathbb{E}_{X \sim P_k}\left[1/(e_k(X_i)(1 - e_k(X_i)))\right]$

Global overlap: $\mathcal{O}_{\text{global}} = \mathbb{E}_{X \sim P}\left[1/(e(X_i)(1 - e(X_i)))\right]$



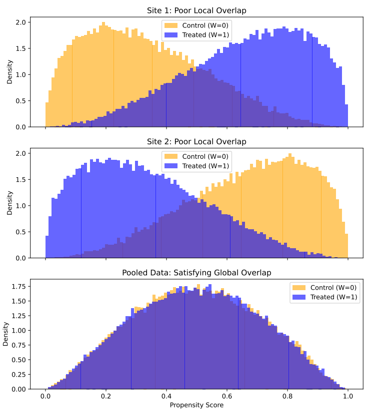
Theoretical results and numerical illustrations

Assuming global overlap:

- $\hat{\tau}^{\text{fed}^*} = \hat{\tau}^{\text{pool}^*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}^*}$, only $\hat{\tau}^{\text{fed}^*}$
- If local overlap **at every site**:
  - $\hat{\tau}^{\text{meta}^*}$ can be computed too
  - The **global overlap is always "better"** than the worst local ones: $\mathcal{O}_{\text{global}} \leq \sum_{k=1}^{K} \rho_k \mathcal{O}_k$

  $\Rightarrow$ Improved stability of $\hat{\tau}^{\text{fed}^*}$ over $\hat{\tau}^{\text{meta}^*}$

$\Rightarrow \mathbb{V}(\hat{\tau}^{\text{pool}^*}) = \mathbb{V}(\hat{\tau}^{\text{fed}^*}) < \mathbb{V}(\hat{\tau}^{\text{meta}^*})$ if heterogeneous e_k's, equality if homogeneous



Theoretical results and numerical illustrations

Assuming global overlap:

- $\hat{\tau}^{\text{fed}^*} = \hat{\tau}^{\text{pool}^*}$
- If no local overlap **in at least one site**: cannot compute $\hat{\tau}^{\text{meta}^*}$, only $\hat{\tau}^{\text{fed}^*}$
- If local overlap **at every site**:
  - $\hat{\tau}^{\text{meta}^*}$ can be computed too
  - The **global overlap is always "better"** than the worst local ones: $\mathcal{O}_{\text{global}} \leq \sum_{k=1}^{K} \rho_k \mathcal{O}_k$

  $\Rightarrow$ Improved stability of $\hat{\tau}^{\text{fed}^*}$ over $\hat{\tau}^{\text{meta}^*}$

$\Rightarrow \mathbb{V}(\hat{\tau}^{\text{pool}^*}) = \mathbb{V}(\hat{\tau}^{\text{fed}^*}) < \mathbb{V}(\hat{\tau}^{\text{meta}^*})$ if heterogeneous e_k's, equality if homogeneous

$\Rightarrow$ Federated estimators should always be preferred over meta-analysis when no communication constraints.

Federated Estimators

Propensity Score Estimation in Practice

Learning a global propensity score with federated learning

The propensity score in the pooled dataset decomposes as $e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$ with $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ the federation weights. Then, to estimate e:

Learning a global propensity score with federated learning

The propensity score in the pooled dataset decomposes as $e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$ with $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ the federation weights. Then, to estimate e:

- e_k's: locally estimated with any (non-)parametric method (logistic, generalized random forests [Athey et al., 2019], etc.) $\rightarrow$ flexible, handles treatment allocation heterogeneity

Learning a global propensity score with federated learning

The propensity score in the pooled dataset decomposes as $e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$ with $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ the federation weights. Then, to estimate e:

- e_k's: locally estimated with any (non-)parametric method (logistic, generalized random forests [Athey et al., 2019], etc.) $\rightarrow$ flexible, handles treatment allocation heterogeneity
- $\omega_k(x)$'s: two approaches

Learning a global propensity score with federated learning

The propensity score in the pooled dataset decomposes as $e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$ with $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ the federation weights. Then, to estimate e:

- e_k's: locally estimated with any (non-)parametric method (logistic, generalized random forests [Athey et al., 2019], etc.) $\rightarrow$ flexible, handles treatment allocation heterogeneity
- $\omega_k(x)$'s: two approaches
  - **Membership Weights** (MW): $H \mid X$

$$\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$$

$\rightarrow$ estimate with a federated probabilitic classifier (logistic regression, neural networks...)

Learning a global propensity score with federated learning

The propensity score in the pooled dataset decomposes as $e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$ with $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ the federation weights. Then, to estimate e:

- e_k's: locally estimated with any (non-)parametric method (logistic, generalized random forests [Athey et al., 2019], etc.) $\rightarrow$ flexible, handles treatment allocation heterogeneity
- $\omega_k(x)$'s: two approaches
  - **Membership Weights** (MW): $H \mid X$

    $$\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$$

    $\rightarrow$ estimate with a federated probabilitic classifier (logistic regression, neural networks...)
  - **Density Ratio Weights** (DW): $X \mid H$

    $$\omega_k(x) = \mathbb{P}(H_i = k) \frac{\mathbb{P}(X_i = x \mid H_i = k)}{\mathbb{P}(X_i = x)} = \rho_k \frac{f_k(x)}{f(x)}$$

    $\rightarrow$ estimate f_k by parametric density estimation (e.g., Gaussian Mixture Models) at site k and global density by $f(x) = \sum_{k=1}^{K} \rho_k f_k(x)$ with $\rho_k = \mathbb{P}(H_i = k)$

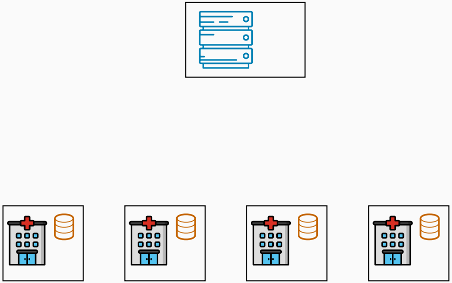
Learning a global propensity score with federated learning

The propensity score in the pooled dataset decomposes as $e(x) = \sum_{k=1}^{K} \omega_k(x) e_k(x)$ with $\omega_k(x) = \mathbb{P}(H_i = k \mid X_i = x)$ the federation weights. Then, to estimate e:

	MW	**DW**
	$\mathbb{P}(H_i = k \mid X_i)$	$\frac{\rho_k f_k(X_i)}{f(X_i)}$
Flexible / non-parametric	✓	✗
Comm. rounds	T	1
Comm. cost	$O(TKd)$	$O(Kd^2)$
Scales to high d	✓	✗

Fed-MW estimation with FedAvg

$$\hat{\omega}_k(x) = \mathbb{P}_{\hat{\theta}}(H_i = k \mid X_i = x)$$



Algorithm FedAvg (server-side)

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

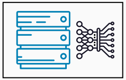
Algorithm $\textsc{ClientUpdate}(k, \theta)$

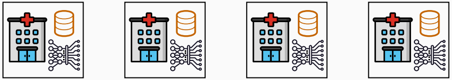
$\theta^{(k)} \leftarrow \theta$
for local step $e = 1$ to E **do**
  $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$
  compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$
  update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)};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

Fed-MW estimation with FedAvg

$$\hat{\omega}_k(x) = \mathbb{P}_{\hat{\theta}}(H_i = k \mid X_i = x)$$

initialize model





Algorithm FedAvg (server-side)

initialize global model parameters θ_0

for each round $t = 1$ to T **do**

  for each client $k \in K$ in parallel **do**

    $\theta_k \leftarrow \textsc{ClientUpdate}(k, \theta)$

  $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

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

$\theta^{(k)} \leftarrow \theta$

for local step $e = 1$ to E **do**

  $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$

  compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$

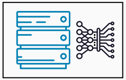
  update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$

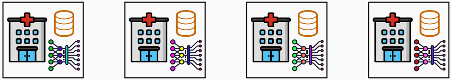
return $\theta^{(k)}$ to server

Fed-MW estimation with FedAvg

$$\hat{\omega}_k(x) = \mathbb{P}_{\hat{\theta}}(H_i = k \mid X_i = x)$$

each party makes an update
using its local dataset





Algorithm FedAvg (server-side)

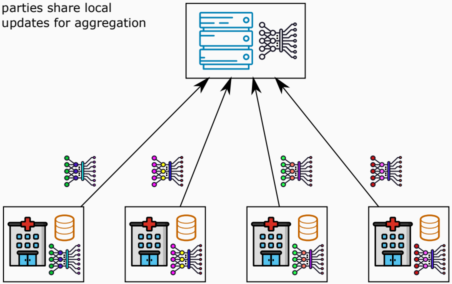
initialize global model parameters θ_0
for each round $t = 1$ to T **do**
  for each client $k \in K$ in parallel **do**
    $\theta_k \leftarrow \textsc{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

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

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

Fed-MW estimation with FedAvg

$$\hat{\omega}_k(x) = \mathbb{P}_{\hat{\theta}}(H_i = k \mid X_i = x)$$



Algorithm FedAvg (server-side)

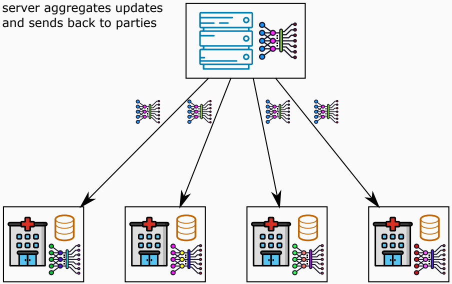
initialize global model parameters θ_0
for each round $t = 1$ to T **do**
  for each client $k \in K$ in parallel **do**
    $\theta_k \leftarrow \textsc{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

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

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

Fed-MW estimation with FedAvg

$$\hat{\omega}_k(x) = \mathbb{P}_{\hat{\theta}}(H_i = k \mid X_i = x)$$



Algorithm FedAvg (server-side)

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

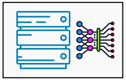
Algorithm $\textsc{ClientUpdate}(k, \theta)$

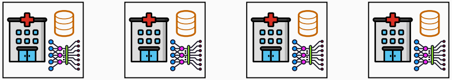
$\theta^{(k)} \leftarrow \theta$
for local step $e = 1$ to E **do**
  $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$
  compute $\nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{B}_k)$
  update $\theta^{(k)} \leftarrow \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)};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

Fed-MW estimation with FedAvg

$$\hat{\omega}_k(x) = \mathbb{P}_{\hat{\theta}}(H_i = k \mid X_i = x)$$

parties update their copy of the model and iterate





Algorithm FedAvg (server-side)

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

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

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

Fed-MW estimation with FedAvg

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

Same principle to estimate global outcome models μ_1, μ_0

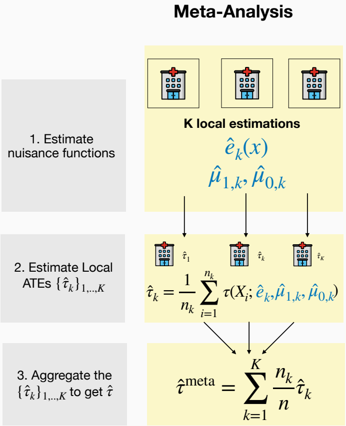
Algorithm FedAvg (server-side)

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

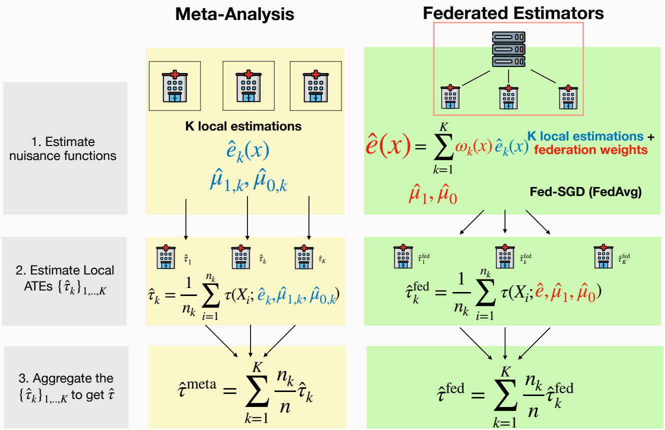
Algorithm $\textsc{ClientUpdate}(k, \theta)$

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

Multi-site estimators: summary

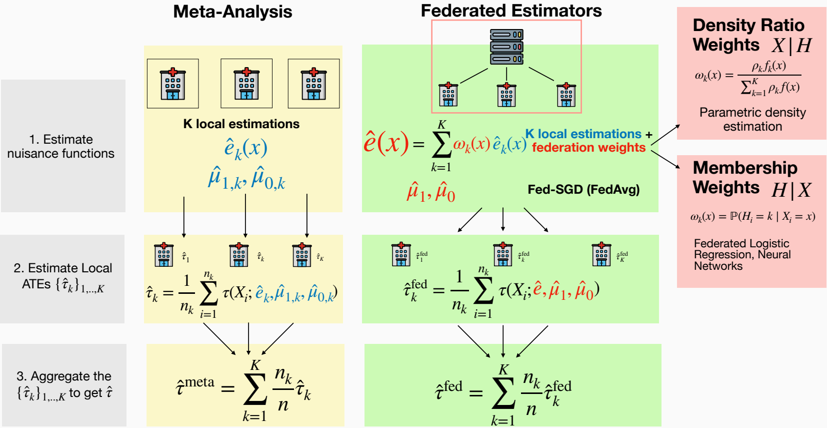


Multi-site estimators: summary



- Meta requires local overlap, federated estimators just global overlap.

Multi-site estimators: summary



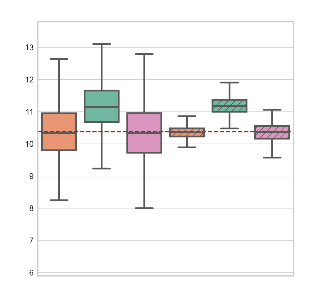
- Meta requires local overlap, federated estimators just global overlap.

Numerical illustration

- $K = 3$ sites, $d = 10$, $n_k = 500$
- Non-linear μ_1, μ_0 estimated with misspecified federated linear regression $\to$ double robustness of Fed-AIPW
- $e_k(x) = \text{Logistic}(\gamma_k, x)$
- MW: Federated logistics, do not work well with $\neq \Sigma_k$'s
- No local overlap: $e_2(x) = 0$ site 2 is an external control arm $\to$ no meta-analysis

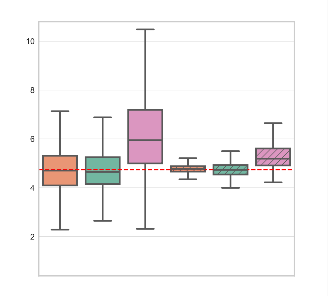
DGP $X|H$

$$X \mid H = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$



DGP $H|X$

$$\mathbb{P}(H = k \mid X) = \text{Logistic}(\theta_k, X)$$



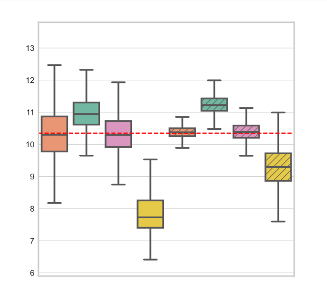
Centralized Oracle   Fed-MW   Fed-DW   Meta-SW   IPW   AIPW   True ATE

Numerical illustration

- $K = 3$ sites, $d = 10$, $n_k = 500$
- Non-linear μ_1, μ_0 estimated with misspecified federated linear regression $\rightarrow$ double robustness of Fed-AIPW
- $e_k(x) = \text{Logistic}(\gamma_k, x)$
- MW: Federated logistics, do not work well with $\neq \Sigma_k$'s
- Poor local overlap: $||\gamma_2||_1$ is large $\rightarrow e_2(x)$ close to 0 for some x's

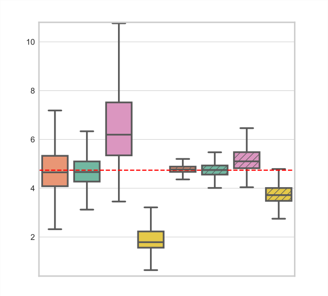
DGP $X|H$

$$X \mid H = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$



DGP $H|X$

$$\mathbb{P}(H = k \mid X) = \text{Logistic}(\theta_k, X)$$



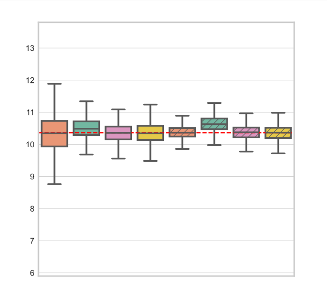
Centralized Oracle | Fed-MW | Fed-DW | Meta-SW | IPW | AIPW | True ATE

Numerical illustration

- $K = 3$ sites, $d = 10$, $n_k = 500$
- Non-linear μ_1, μ_0 estimated with misspecified federated linear regression $\rightarrow$ double robustness of Fed-AIPW
- $e_k(x) = \text{Logistic}(\gamma_k, x)$
- MW: Federated logistics, do not work well with $\neq \Sigma_k$'s
- Good local overlaps: all $\mathcal{O}_k$'s are small and close to $\mathcal{O}_{\text{global}}$.

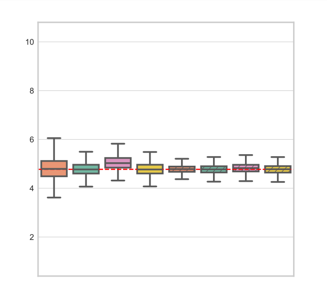
DGP $X|H$

$$X \mid H = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$



DGP $H|X$

$$\mathbb{P}(H = k \mid X) = \text{Logistic}(\theta_k, X)$$



Centralized Oracle   Fed-MW   Fed-DW   Meta-SW   IPW   AIPW   True ATE

Conclusion

Limits of our approach:

- MW vs. DW: MW with Neural Networks always works but requires more local data
- Cross-silos setting:
  - small K since number of federation weights' parameters grows with K
  - large n_k to estimate e_k's, outcome models, membership probabilities/density parameters

Perspectives:

- Handle centre effects beyond parametric modelling of $e(X, H)$
- Handle covariate mismatch across sources
- Consider non-collapsible measures (e.g., odd-ratios)
- Provide robust privacy guarantees (differential privacy)

Thank you for your attention!

Questions?

References i

[Athey et al., 2019] Athey, S., Tibshirani, J., and Wager, S. (2019).
Generalized random forests.

[Guo et al., 2024] Guo, T., Karimireddy, S. P., and Jordan, M. I. (2024).
Collaborative heterogeneous causal inference beyond meta-analysis.
arXiv preprint arXiv:2404.15746.

[Han et al., 2021] Han, L., Hou, J., Cho, K., Duan, R., and Cai, T. (2021).
Federated adaptive causal estimation (face) of target treatment effects.
arXiv preprint arXiv:2112.09313.

[Han et al., 2023] Han, L., Shen, Z., and Zubizarreta, J. R. (2023).
Multiply robust federated estimation of targeted average treatment effects.
In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., 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.*

[Kairouz et al., 2021] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021).
Advances and open problems in federated learning.
Foundations and trends® in machine learning, 14(1–2):1–210.

References ii

[Makhija et al., 2024] Makhija, D., Ghosh, J., and Kim, Y. (2024).
Federated learning for estimating heterogeneous treatment effects.
CoRR, abs/2402.17705.

[Vo et al., 2022a] Vo, T. V., Bhattacharyya, A., Lee, Y., and Leong, T.-Y. (2022a).
An adaptive kernel approach to federated learning of heterogeneous causal effects.
Advances in Neural Information Processing Systems, 35:24459–24473.

[Vo et al., 2022b] Vo, T. V., Lee, Y., Hoang, T. N., and Leong, T.-Y. (2022b).
Bayesian federated estimation of causal effects from observational data.
In *UAI*.

[Xiong et al., 2023] Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. (2023).
Federated causal inference in heterogeneous observational data.
Statistics in Medicine, 42(24):4418–4439.