Breaking Data Silos: Multi-Source Average Treatment Effect Estimation beyond Meta-Analysis

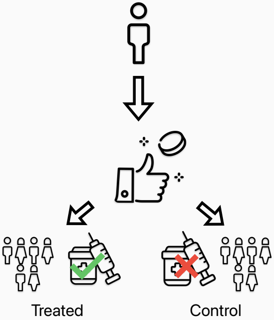
Rémi Khellaf, Aurélien Bellet and Julie Josse (INRIA, Montpellier)

July 6, 2025

Federated causal inference

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

Randomized Controlled Trials (RCTs):



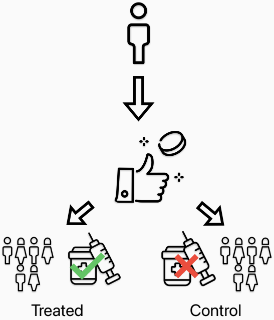
+ : direct causal association

– : limited scope (eligibility criteria), small sample sizes, not always feasible

Federated causal inference

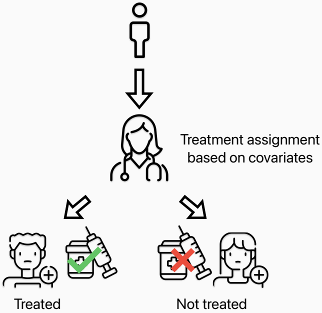
Goal of causal inference: measure the effect of a treatment on an outcome

Randomized Controlled Trials (RCTs):



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

Observational Data:

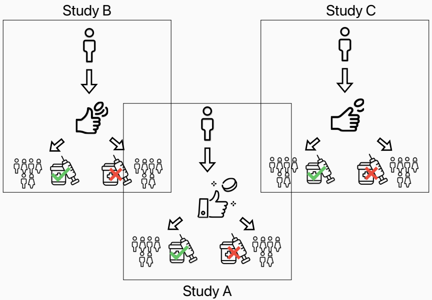


+ : abundant, large scope, always available
− : naturally scattered across sites (e.g., hospitals), confounding factors

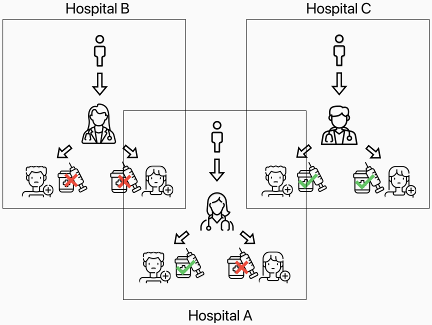
Federated causal inference

Multi-source causal inference: higher validity and generalization

Randomized Controlled Trials (RCTs):



Observational Data:



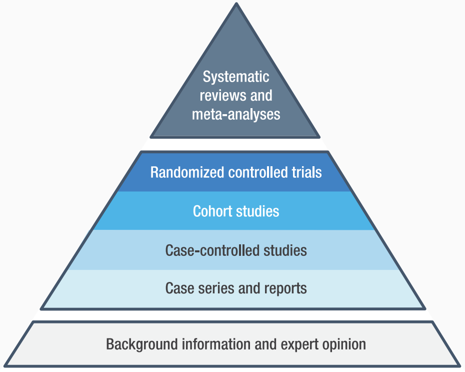
+ : direct causal association
– : limited scope (eligibility criteria), small sample sizes, not always feasible

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

Classic approach: Meta-analysis

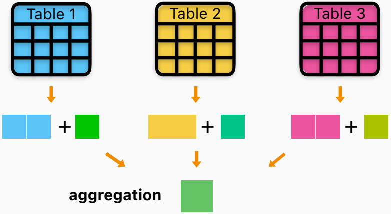
Meta-analysis (MA) combines effects from multiple studies

It is at the top of the evidence hierarchy



Classic approach: Meta-analysis

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

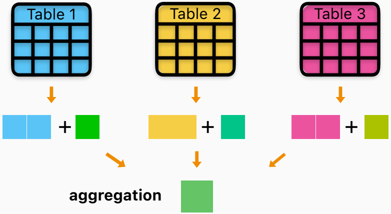


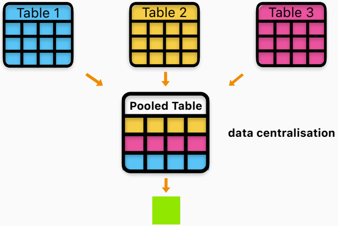
Aggregated Data (AD):

- 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):

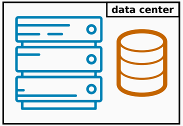
- Studies report summary statistics + effect sizes which are aggregated into a single one.
- **Limitation**: Prone to **ecological bias**.

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

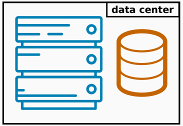
IPD cannot always be pooled altogether



- 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)

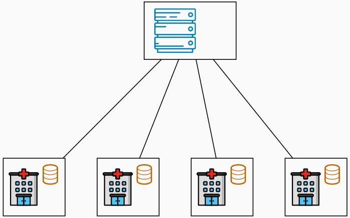
Enabling individual patient data analysis with federated learning

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)

Multiple RCTs

Reminder: classic RCT framework

- Estimate effect of treatment W on outcome Y given covariates X, with $W_i \sim \mathcal{B}(p)$
- Average Treatment Effect (ATE) measured as a risk difference $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$

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	??
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	3.7	2.0	F	0	2.8	??	2.8
n	2.5	1.7	M	1	3.2	3.2	??

Our setting: decentralized heterogeneous RCTs

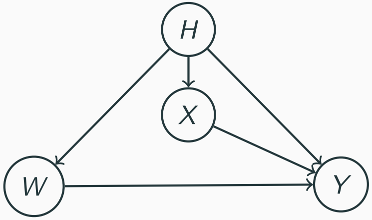
- We consider K decentralized and potentially heterogeneous RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

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: decentralized heterogeneous RCTs

- We consider K decentralized and potentially heterogeneous RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

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

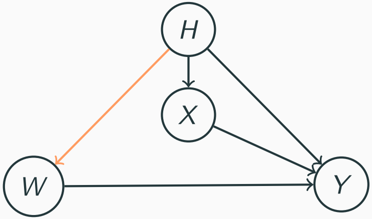


Our setting: decentralized heterogeneous RCTs

- We consider K decentralized and potentially heterogeneous RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

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 allocation

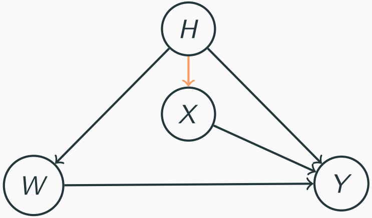


Our setting: decentralized heterogeneous RCTs

- We consider K decentralized and potentially heterogeneous RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

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

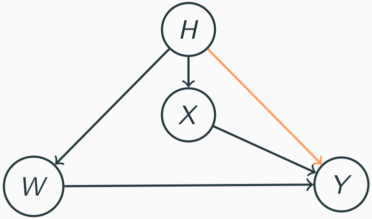


Our setting: decentralized heterogeneous RCTs

- We consider K decentralized and potentially heterogeneous RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

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

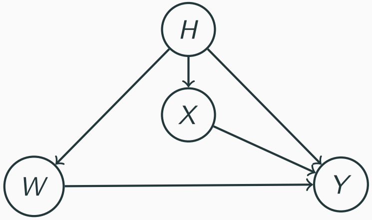
Presence of **center effects**



Our setting: decentralized heterogeneous RCTs

- We consider K decentralized and potentially heterogeneous RCTs (studies) from different sources and want to estimate the ATE given by $\tau = \mathbb{E}\left(\mathbb{E}(Y^{(1)} - Y^{(0)} \mid H)\right)$

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



How to estimate τ without pooling together individual-level data?

Model and assumptions

- For now, same linear outcome model for all studies:

$$\forall k : \quad Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}, \quad \text{with } \mathbb{E}\left[X_k^\top \varepsilon_{k,i}^{(w)}\right] = 0, \mathbb{V}\left(\varepsilon_{k,i}^{(w)} \mid X_k\right) = \sigma^2$$

Model and assumptions

- For now, same linear outcome model for all studies:

$$\forall k: \quad Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}, \quad \text{with } \mathbb{E}\left[X_k^\top \varepsilon_{k,i}^{(w)}\right] = 0, \mathbb{V}\left(\varepsilon_{k,i}^{(w)} \mid X_k\right) = \sigma^2$$

- Standard assumptions (consistency, positivity, unconfoundedness)

Model and assumptions

- For now, same linear outcome model for all studies:

$$\forall k: \quad Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}, \quad \text{with } \mathbb{E}\left[X_k^\top \varepsilon_{k,i}^{(w)}\right] = 0, \mathbb{V}\left(\varepsilon_{k,i}^{(w)} \mid X_k\right) = \sigma^2$$

- Standard assumptions (consistency, positivity, unconfoundedness)
- We aim to estimate the ATE $\tau = \mathbb{E}[Y^{(1)} - Y^{(1)}] = \mathbb{E}[X'] \left(\theta^{(1)} - \theta^{(0)}\right)$.

Model and assumptions

- For now, same linear outcome model for all studies:

$$\forall k: \quad Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}, \quad \text{with } \mathbb{E}\left[X_k^\top \varepsilon_{k,i}^{(w)}\right] = 0, \mathbb{V}\left(\varepsilon_{k,i}^{(w)} \mid X_k\right) = \sigma^2$$

- Standard assumptions (consistency, positivity, unconfoundedness)
- We aim to estimate the ATE $\tau = \mathbb{E}[Y^{(1)} - Y^{(1)}] = \mathbb{E}[X'] \left(\theta^{(1)} - \theta^{(0)}\right)$.
- Ideal baseline: estimator $\hat{\tau}_{\text{pool}} = \frac{1}{n}\sum_{i=1}^{n} X_i'(\hat{\theta}_{\text{pool}}^{(1)} - \hat{\theta}_{\text{pool}}^{(0)})$ on pooled data, where

$\hat{\theta}_{\text{pool}}^{(w)} = (\hat{c}_{\text{pool}}^{(w)}, \hat{\beta}_{\text{pool}}^{(w)}) = \left({X'^{(w)}}^\top X'^{(w)}\right)^{-1} {X'^{(w)}}^\top Y^{(w)}$ is the OLS estimator and $X'^{(w)} = [1, X^{(w)}]$

Model and assumptions

- For now, same linear outcome model for all studies:

$$\forall k : \quad Y_{k,i}^{(w)} = c^{(w)} + X_{k,i}\beta^{(w)} + \varepsilon_{k,i}^{(w)}, \quad \text{with } \mathbb{E}\left[X_k^\top \varepsilon_{k,i}^{(w)}\right] = 0, \mathbb{V}\big(\varepsilon_{k,i}^{(w)} \mid X_k\big) = \sigma^2$$

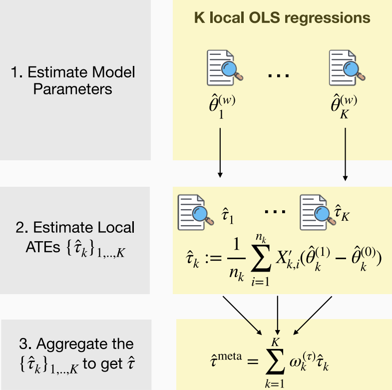
- Standard assumptions (consistency, positivity, unconfoundedness)
- We aim to estimate the ATE $\tau = \mathbb{E}[Y^{(1)} - Y^{(1)}] = \mathbb{E}[X'] \left(\theta^{(1)} - \theta^{(0)}\right)$.
- Ideal baseline: estimator $\hat{\tau}_{\mathrm{pool}} = \frac{1}{n}\sum_{i=1}^{n} X_i'(\hat{\theta}_{\mathrm{pool}}^{(1)} - \hat{\theta}_{\mathrm{pool}}^{(0)})$ on pooled data, where

  $\hat{\theta}_{\mathrm{pool}}^{(w)} = (\hat{c}_{\mathrm{pool}}^{(w)}, \hat{\beta}_{\mathrm{pool}}^{(w)}) = \left({X'^{(w)}}^\top X'^{(w)}\right)^{-1} {X'^{(w)}}^\top Y^{(w)}$ is the OLS estimator and $X'^{(w)} = [1, X^{(w)}]$
- $\hat{\tau}_{\mathrm{pool}}$ always has lower variance than the simple difference-in-means estimator [Benkeser et al., 2021, Lei and Ding, 2021]

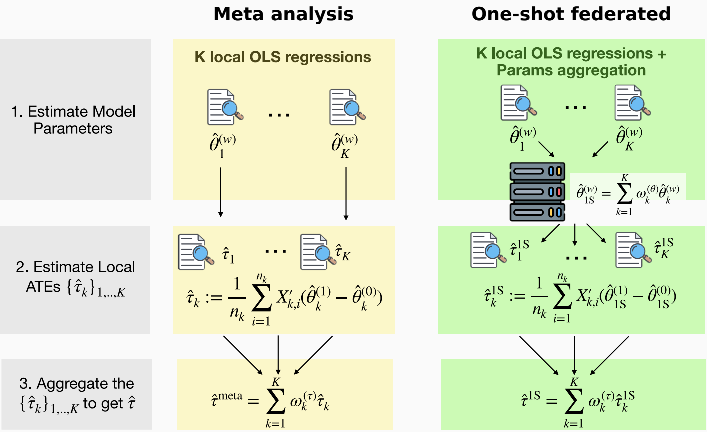
Federated Estimators

Three types of federated estimators

Meta analysis

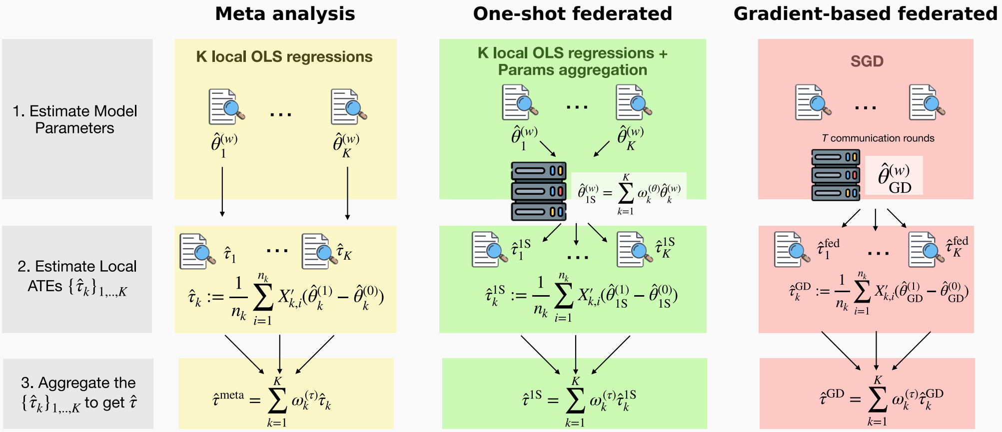


Three types of federated estimators



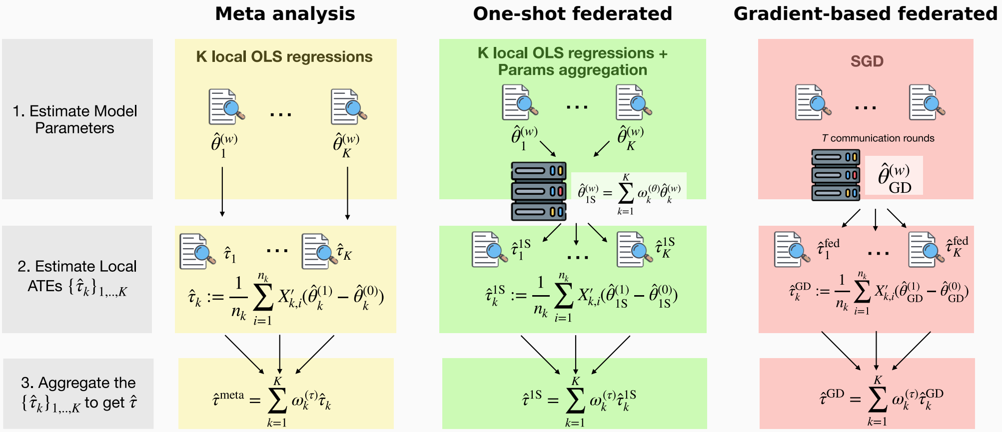
- Meta and one-shot require local sample size $n_k^{(w)} \geq d$ for k, w

Three types of federated estimators



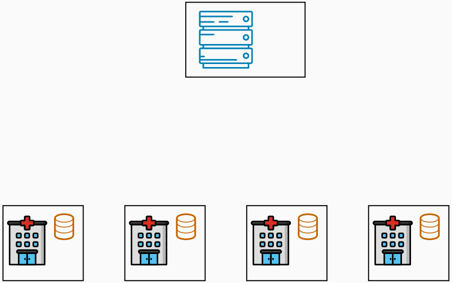
- Meta and one-shot require local sample size $n_k^{(w)} \geq d$ for k, w

Three types of federated estimators



- Meta and one-shot require local sample size $n_k^{(w)} \geq d$ for k, w
- Aggregation: sample size weights (SW) or inverse variance weights (IVW)

A baseline FL algorithm: FedAvg



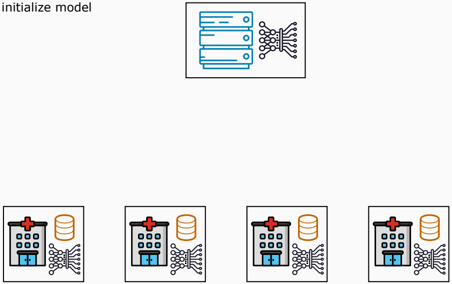
Algorithm FedAvg (server-side)

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

A baseline FL algorithm: FedAvg



Algorithm FedAvg (server-side)

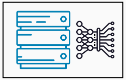
initialize global model parameters θ_0
for each round $t = 1$ to T **do**
  for each client $k \in K$ in parallel **do**
    $\theta_k \leftarrow \textsc{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \sum_{k \in K} \frac{n_k}{n} \theta_k$ `// FedAvg`

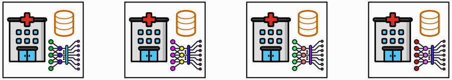
Algorithm $\textsc{ClientUpdate}(k, \theta)$

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

A baseline FL algorithm: FedAvg

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**

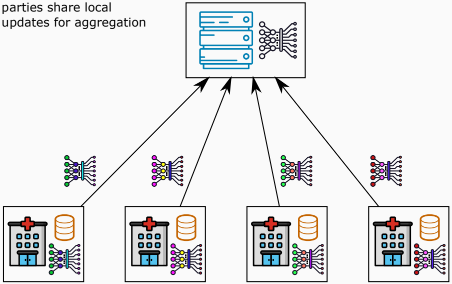
  $\mathcal{B}_k \leftarrow$ mini-batch of B samples from $\mathcal{D}_k$

  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

A baseline FL algorithm: FedAvg



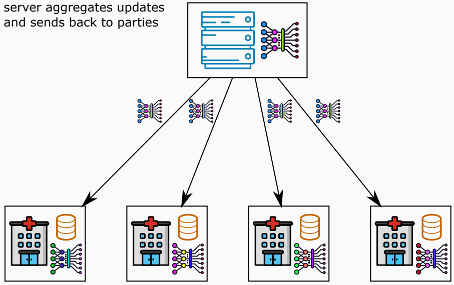
Algorithm FedAvg (server-side)

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

A baseline FL algorithm: FedAvg



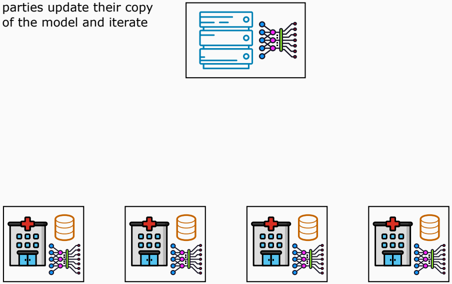
Algorithm FedAvg (server-side)

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

A baseline FL algorithm: FedAvg



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

A baseline FL algorithm: 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.

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

Federated Averaging (FedAvg) for Linear Regression

Linear Regression

- $Y = X\beta + \varepsilon$. Estimate β by minimizing the MSE:

$$\arg\min_\beta \ell(\beta; X_i, Y_i) \text{ with } \ell(\beta; X_i, Y_i) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

Gradient Descent (GD)

1. Initialize β_0 with zeros
2. Update $\beta_{t+1} := \beta_t - \eta\nabla\ell(\beta_t)$ with $\nabla\ell(\beta_t) = -\frac{2}{n}\sum_{i=1}^{n} X_i^T(Y_i - X_i\beta)$
3. Repeat for E steps until convergence

Choices: learning rate η & E to get $\hat{\beta}_{\text{GD}} \approx \hat{\beta}_{\text{OLS}}$ with equality as $E \to \infty$.
Choose $\eta < \frac{2}{\lambda_{\text{max}}}$ where λ_{max} is the highest eigenvalue of X^TX.

Federated Averaging (FedAvg) for Linear Regression

FedAvg Objective

- $Y = X\beta + \varepsilon$. Estimate $\hat{\beta}_{\text{FedAvg}}$ by minimizing:

$$\arg\min_\beta \sum_{k=1}^{K} \frac{n_k}{n} \ell_k(\beta) \text{ with } \ell_k(\beta) = \frac{1}{n_k} \sum_{i=1}^{n_k} (Y_i^k - X_i^k \beta)^2$$

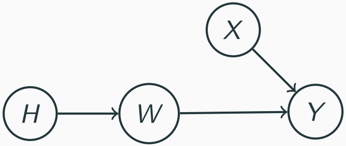
Federated Learning extends GD to a distributed setting

1. Initialize β_0 on central server with zeros (globally shared)
2. For each **communication round** $t = 1, \ldots, T$:
   - Each site $k = 1, \ldots, K$ performs E **gradient steps** on its data: $\beta_{t+1}^k = \beta_t^k - \eta \nabla \ell_k(\beta_t^k)$ where $\nabla \ell_k(\beta_t^k) = -\frac{2}{n_k} \sum_{i=1}^{n_k} X_i^{k,T} (Y_i^k - X_i^k \beta_t^k)$
   - Parameters are sent to the server for aggregation: $\beta_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \beta_{t+1}^k$

Choices: learning rate η, communication T & L.

$T = 1$ & $L \to \infty$: One-shot federated learning, meta analysis on β.

Homogeneous setting



- The source membership variable H only affects the treatment allocation scheme
- Let $W_{k,i} \sim \mathcal{B}(p_k)$

Summary of results

Estimators are unbiased but differ by their asymptotic variance and communication costs:

Estimator	Notation	$\mathbb{V}^\infty$	Com. rounds	Com. cost
Meta-SW	$\hat{\tau}_{\text{Meta-SW}}$	$\frac{\sigma^2}{n}\sum_{k=1}^{K}\frac{\rho_k}{p_k(1-p_k)}+\frac{1}{n}\lVert\beta^{(1)}-\beta^{(0)}\rVert_\Sigma^2$	1	$O(1)$
Meta-IVW	$\hat{\tau}_{\text{Meta-IVW}}$	$\left(\sum_{k=1}^{K}\left(\sigma^2\frac{n\rho_k}{p_k(1-p_k)}+\frac{1}{n_k}\lVert\beta^{(1)}-\beta^{(0)}\rVert_\Sigma^2\right)^{-1}\right)^{-1}$	1	$O(1)$
1S-SW	$\hat{\tau}_{\text{1S-SW}}$	V_{pool}	2	$O(d)$
1S-IVW	$\hat{\tau}_{\text{1S-IVW}}$	V_{pool}	2	$O(d^2)$
GD	$\hat{\tau}_{\text{GD}}$	V_{pool}	$T+1$	$O(Td)$
Pool	$\hat{\tau}_{\text{pool}}$	$V_{\text{pool}}=\frac{\sigma^2}{n}\frac{1}{p(1-p)}+\frac{1}{n}\lVert\beta^{(1)}-\beta^{(0)}\rVert_\Sigma^2$	—	—

with $\rho_k=\mathbb{P}(H=k)=\mathbb{E}\left[\frac{n_k}{n}\right]$ and $p=\sum_{k=1}^{K}\frac{n_k}{n}p_k$

Summary of results

Estimators are unbiased but differ by their asymptotic variance and communication costs:

$$
\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) &= \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{GD}}) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{1S-SW}}) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{1S-IVW}}) \\
&\qquad \leq \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{Meta-IVW}}) \left\{ \right.
\end{aligned}
$$

Summary of results

Estimators are unbiased but differ by their asymptotic variance and communication costs:

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

Summary of results

Estimators are unbiased but differ by their asymptotic variance and communication costs:

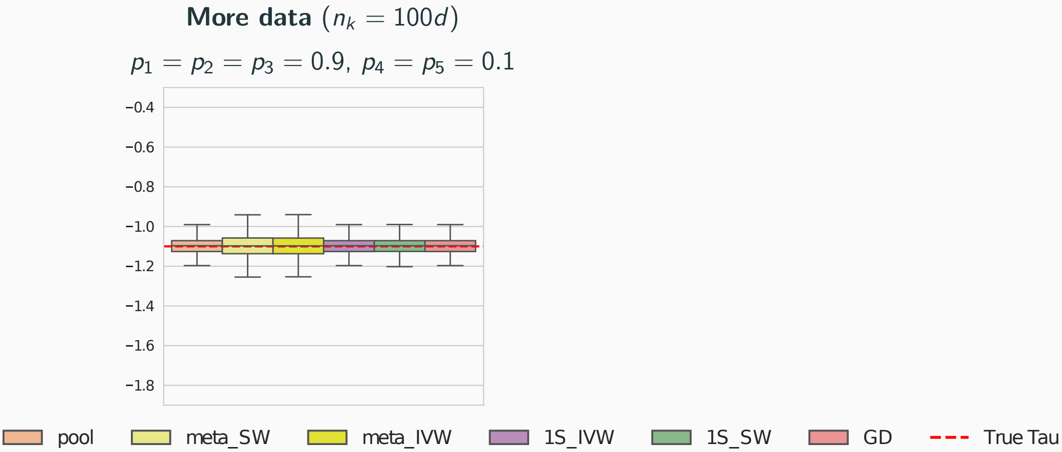
$$\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) &= \mathbb{V}^{\infty}(\hat{\tau}_{\text{GD}}) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-SW}}) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\text{1S-IVW}}) \\
&\quad \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-IVW}}) \begin{cases} = & \text{if same } \{p_k\}_k, \\ & \end{cases} \\
&\qquad \leq \mathbb{V}^{\infty}(\hat{\tau}_{\text{Meta-SW}}) \begin{cases} = & \text{if same } \{p_k(1-p_k)\}_k, \\ & \end{cases}
\end{aligned}$$

Summary of results

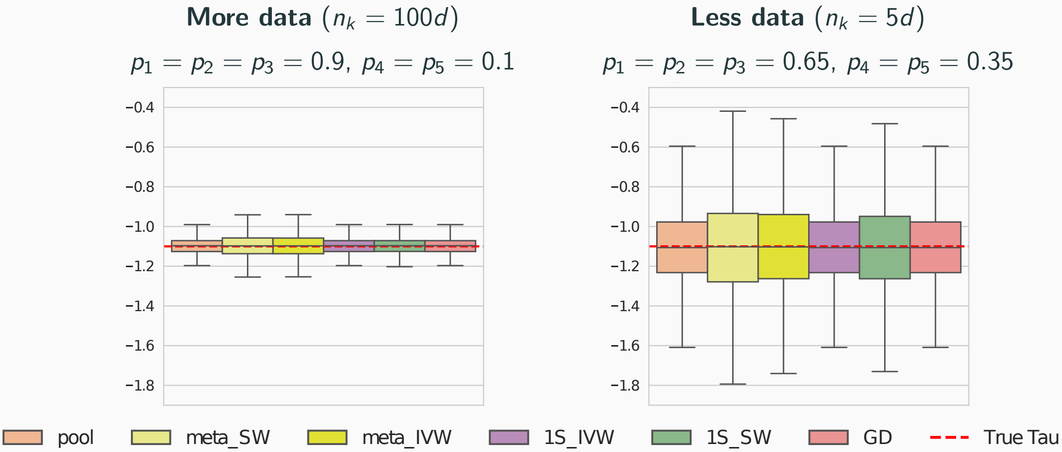
Estimators are unbiased but differ by their asymptotic variance and communication costs:

$$\begin{aligned}
\mathbb{V}^{\infty}(\hat{\tau}_{\text{pool}}) &= \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{GD}}) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{1S-SW}}) \\
&= \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{1S-IVW}}) \\
&\qquad \leq \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{Meta-IVW}}) \begin{cases} = & \text{if same } \{p_k\}_k, \\ < & \text{if different } \{p_k\}_k \end{cases} \\
&\qquad\qquad \leq \mathbb{V}^{\infty}(\hat{\tau}_{\mathrm{Meta-SW}}) \begin{cases} = & \text{if same } \{p_k(1-p_k)\}_k, \\ < & \text{if different } \{p_k(1-p_k)\}_k \end{cases}
\end{aligned}$$

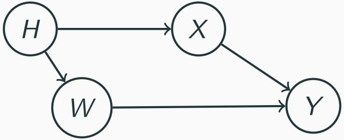
Numerical illustration ($K = 5$ and $d = 10$)



Numerical illustration ($K = 5$ and $d = 10$)



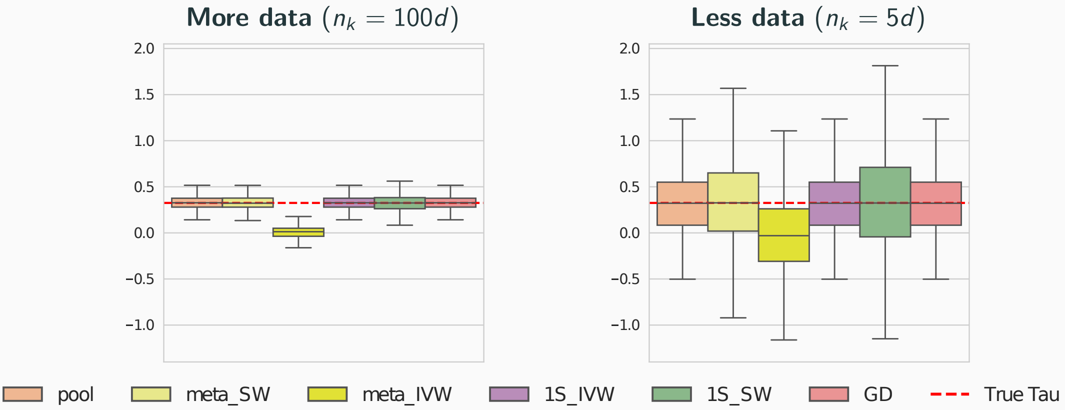
Heterogeneity in covariates distributions



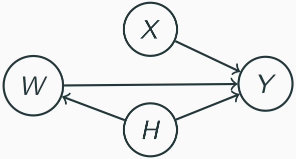
- Distributional shift across sources: $H \not\perp\!\!\!\perp X \implies \tau_k \neq \tau_{k'}$
- Global ATE is given by $\tau = \sum_{k=1}^{K} \rho_k \tau_k$ with $\rho_k = \mathbb{P}(H = k) = \mathbb{E}\left[\frac{n_k}{n}\right]$

Numerical illustration

$$X_k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

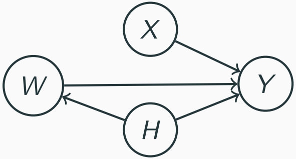


Heterogeneity from Center Effects



- Studies may have different baselines in individual outcomes due to varying practices or organizational contexts (e.g. hospital specialized in oncology)

Heterogeneity from Center Effects

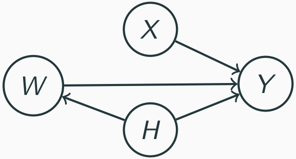


- Studies may have different baselines in individual outcomes due to varying practices or organizational contexts (e.g. hospital specialized in oncology)
- We model this by a fixed effect of the source H onto the outcome Y:

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

(Note: the CATEs $\mathbb{E}[Y(1) - Y(0)|X, H]$ remain the same across sources)

Heterogeneity from Center Effects



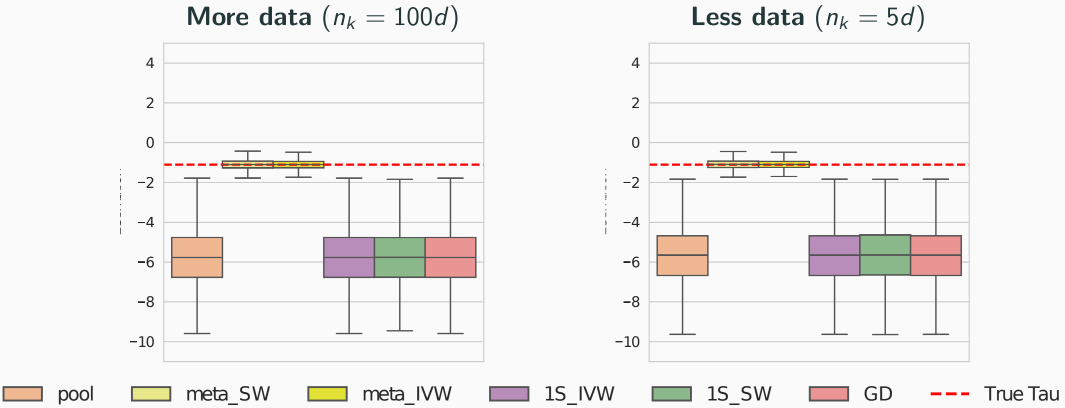
- Studies may have different baselines in individual outcomes due to varying practices or organizational contexts (e.g. hospital specialized in oncology)
- We model this by a fixed effect of the source H onto the outcome Y:

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

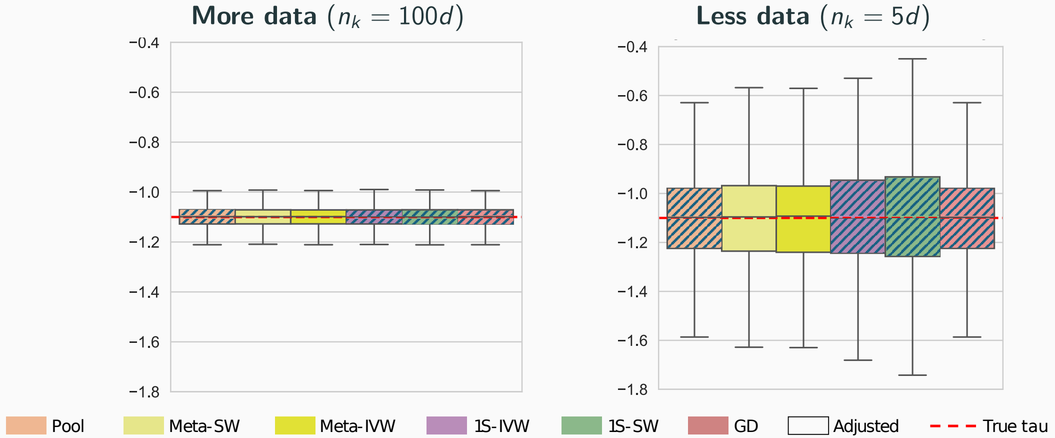
(Note: the CATEs $\mathbb{E}[Y(1) - Y(0)|X, H]$ remain the same across sources)

- Caution: H is now a confounder!

Numerical illustration



Numerical illustration



Summary: decision diagram for practitioners

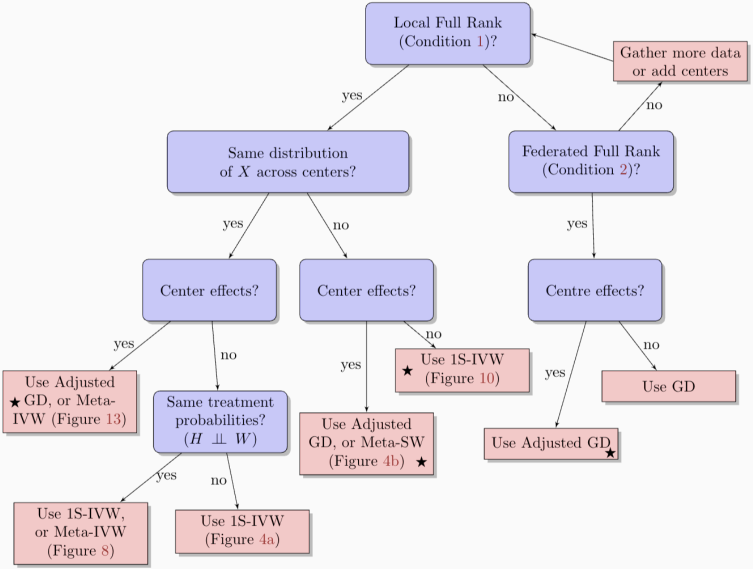


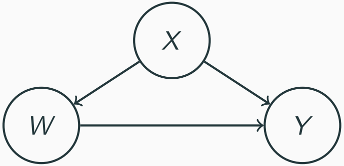
Figure 6: Decision Diagram for Practitionners. The sign ★ denotes scenarios where the DM estimator is biased.

Multiple Observational Studies

Classic framework with observational data

- Goal: estimate effect of treatment W on outcome Y given covariates X
- Observational setting: $W \not\perp\!\!\!\perp X$, treatment allocation based on patient covariates
- X is a confounder: need to account for either $\mathbb{P}(W_i = 1 \mid X_i)$ or $\mathbb{E}(Y_i \mid W_i, X_i)$

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	??



Classic (oracle) centralized ATE estimators

- Denote $e(x) = \mathbb{P}(W = 1 \mid X = x)$ (propensity score) and $\mu_w(x) = \mathbb{E}(Y \mid W = w, X = x)$

Classic (oracle) centralized ATE estimators

- Denote $e(x) = \mathbb{P}(W = 1 \mid X = x)$ (propensity score) and $\mu_w(x) = \mathbb{E}(Y \mid W = w, X = x)$

Inverse Propensity Weighting (IPW):

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

Classic (oracle) centralized ATE estimators

- Denote $e(x) = \mathbb{P}(W = 1 \mid X = x)$ (propensity score) and $\mu_w(x) = \mathbb{E}(Y \mid W = w, X = x)$

Inverse Propensity Weighting (IPW):

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

Classic (oracle) centralized ATE estimators

- Denote $e(x) = \mathbb{P}(W = 1 \mid X = x)$ (propensity score) and $\mu_w(x) = \mathbb{E}(Y \mid W = w, X = x)$

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):

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

which is doubly robust

Classic (oracle) centralized ATE estimators

- Denote $e(x) = \mathbb{P}(W = 1 \mid X = x)$ (propensity score) and $\mu_w(x) = \mathbb{E}(Y \mid W = w, X = x)$

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):

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

which is doubly robust

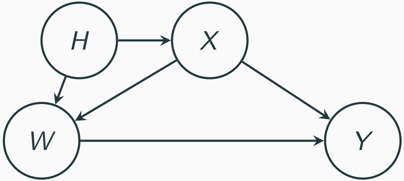
Assumptions for consistency:

- **SUTVA**: $Y = WY(1) + (1 - W)Y(0)$
- **Unconfoundedness**: $Y(0), Y(1) \perp\!\!\!\perp W \mid X$
- **Bounded outcomes**
- **Overlap**: $\exists \eta > 0$, $\forall x \in \mathcal{X}$, $\eta < e(x) < 1 - \eta$

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



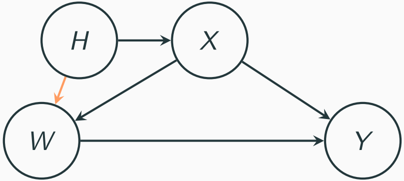
Our setting: multi-site decentralized observational data

- 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**

$$e_k(x) = \mathbb{P}(W \mid X = x, H = k)$$



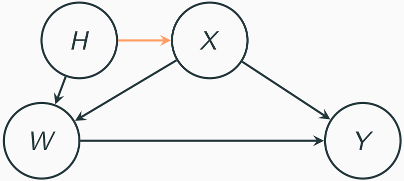
Our setting: multi-site decentralized observational data

- 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

Heterogeneity in **covariates distribution**

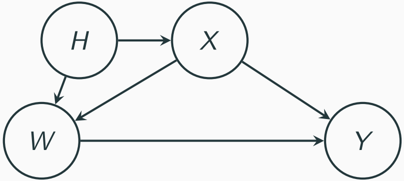
$$X \mid H = k \nsim X \mid H = k'$$



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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	1	4.5	5.0	F	1	4.1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	1	3.7	2.0	F	0	2.8
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	n_K	2.5	1.7	M	0	3.2



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

Federated Estimators

How to design a Federated IPW estimator?

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

- **FL folks have thought of:**
  - learning a global propensity score model $e(x) = \mathbb{P}(W_i = 1 \mid X = x)$ [Guo et al., 2025] but this is very restrictive (note: we would also like e to be non-parametric for consistency)

How to design a Federated IPW estimator?

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

- **FL folks have thought of:**
  - learning a global propensity score model $e(x) = \mathbb{P}(W_i = 1 \mid X = x)$ [Guo et al., 2025] but this is very restrictive (note: we would also like e to be non-parametric for consistency)
- **Causal folks have thought of:**
  - One-shot averaging of local propensity models $e_k(x) = \mathbb{P}(W_i = 1 \mid X = x, H = k)$, restricting to parameters assumed to be shared across sites [Xiong et al., 2023]

How to design a Federated IPW estimator?

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

- **FL folks have thought of:**
  - learning a global propensity score model $e(x) = \mathbb{P}(W_i = 1 \mid X = x)$ [Guo et al., 2025] but this is very restrictive (note: we would also like e to be non-parametric for consistency)
- **Causal folks have thought of:**
  - One-shot averaging of local propensity models $e_k(x) = \mathbb{P}(W_i = 1 \mid X = x, H = k)$, restricting to parameters assumed to be shared across sites [Xiong et al., 2023]
  - Reweighting with density ratios $f_k(X)/f(X)$, either parametrically under strong assumptions [Han et al., 2023] or non-parametrically without FL algorithm [Guo et al., 2024]

Our approach: decompose the global propensity score

- Using simple manipulations we can rewrite:

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

Our approach: decompose the global propensity score

- Using simple manipulations we can rewrite:

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

- Therefore, each site can learn its local propensity score independently with the (non-parametric) model of their choice $\rightarrow$ maximum flexibility

Our approach: decompose the global propensity score

- Using simple manipulations we can rewrite:

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

- Therefore, each site can learn its local propensity score independently with the (non-parametric) model of their choice $\rightarrow$ maximum flexibility
- Learning the membership weights is a federated multi-class classification problem that can be solved using standard FL methods

Our approach: decompose the global propensity score

- Using simple manipulations we can rewrite:

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

- Therefore, each site can learn its local propensity score independently with the (non-parametric) model of their choice $\rightarrow$ maximum flexibility
- Learning the membership weights is a federated multi-class classification problem that can be solved using standard FL methods
- Membership weights can be rewritten as density ratios: $\mathbb{P}(H = k \mid X) = \frac{f_k(X)}{\sum_{k'=1}^{K} f_{k'}(X)}$, where $f_k(X)$ is the density of X at site k $\rightarrow$ enables one-shot estimation procedure under parametric assumptions of the local distributions.

Our approach: decompose the global propensity score

- Using simple manipulations we can rewrite:

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

- Therefore, each site can learn its local propensity score independently with the (non-parametric) model of their choice $\rightarrow$ maximum flexibility
- Learning the membership weights is a federated multi-class classification problem that can be solved using standard FL methods
- Membership weights can be rewritten as density ratios: $\mathbb{P}(H = k \mid X) = \frac{f_k(X)}{\sum_{k'=1}^{K} f_{k'}(X)}$, where $f_k(X)$ is the density of X at site k $\rightarrow$ enables one-shot estimation procedure under parametric assumptions of the local distributions.
- For AIPW, need to also learn $\mu_w(x) = \mathbb{E}(Y \mid W = w, X = x)$ for $w \in \{0, 1\}$ $\rightarrow$ again a standard federated regression problem, as in the case of RCTs [Khellaf et al., 2025b]

Theoretical results

- We need the additional assumption of site ignorability: : $Y(0), Y(1) \perp\!\!\!\perp H \mid X$
  - $\Rightarrow$ Common conditional outcome models $\{\mu_1, \mu_0\}$ across sites
  - $\Rightarrow$ H is not a confounder (no site-specific effect): learning $e(X)$ suffices to deconfound

Theorem (Variance comparison of oracle estimators — informal)

We have

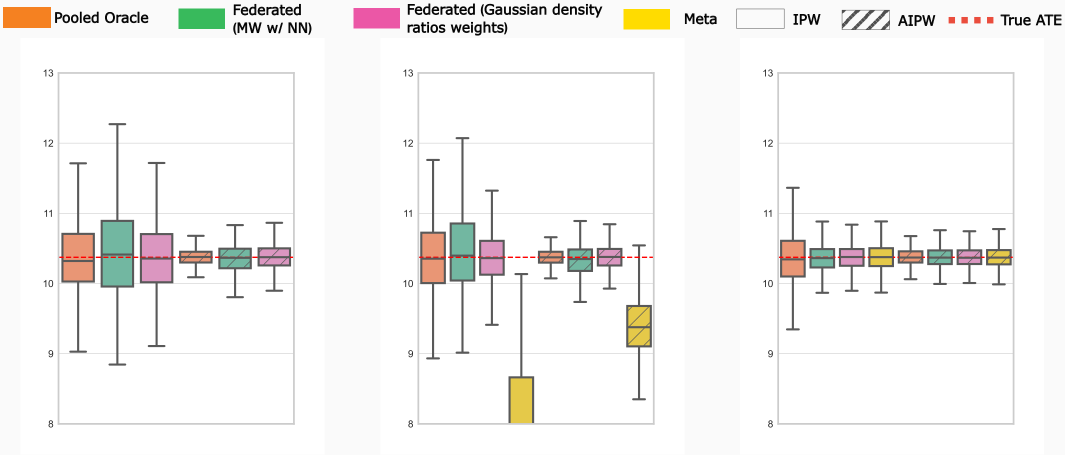
$$\mathbb{V}[\hat{\tau}^*_{\mathrm{IPW}}] = \mathbb{V}[\hat{\tau}^{\mathrm{fed}^*}_{\mathrm{IPW}}] \leq \mathbb{V}\left[\hat{\tau}^{\mathrm{meta}^*}_{\mathrm{IPW}}\right],$$

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

with equality when the local propensity scores are equal.

- Our approach is superior to meta-analysis when local overlap is low: we leverage heterogeneity in treatment assignment to improve overlap

Empirical results



(a) No local overlap

external control arm in site 2

(b) Poor local overlap

$\min(e_2(x)) = 10^{-3}$

(c) Good local overlap

$\min(e_2(x)) = 0.1$

Conclusion & Perspectives

Conclusion & Perspectives

- **Key takeaway:** Federated learning can address data-sharing challenges in causal inference, but dedicated methods are needed to ensure causal validity

Conclusion & Perspectives

- **Key takeaway:** Federated learning can address data-sharing challenges in causal inference, but dedicated methods are needed to ensure causal validity
- **Many open problems:**
  - Non-collapsible causal measures (e.g., odds ratio)
  - Differential privacy guarantees (see [Lebeda et al., 2025] for the centralized case)
  - Move beyond ATE towards more personalization
  - Transfer treatment effects to different target populations

Thank you for your attention!

Questions?

[Benkeser et al., 2021] Benkeser, D., Díaz, I., Luedtke, A., Segal, J., Scharfstein, D., and Rosenblum, M. (2021).
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.

[Berenfeld et al., 2025] Berenfeld, C., Boughdiri, A., Colnet, B., van Amsterdam, W. A. C., Bellet, A., Khellaf, R., Scornet, E., and Josse, J. (2025).
Causal Meta-Analysis: Rethinking the Foundations of Evidence-Based Medicine.
Technical report, arXiv:2505.20168.

[Berlin et al., 2002] Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., and Feldman, H. I. (2002).
Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head.
Statistics in medicine, 21(3):371–387.

[Duflo et al., 2007] Duflo, E., Glennerster, R., and Kremer, M. (2007).
Using randomization in development economics research: A toolkit.
Handbook of development economics, 4:3895–3962.

[French Health Authority, 2024] French Health Authority (2024).
Pricing & reimbursement of drugs and hta policies in france.

References ii

[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.

[Guo et al., 2025] Guo, Z., Li, X., Han, L., and Cai, T. (2025).
Robust inference for federated meta-learning.
Journal of the American Statistical Association, pages 1–16.

[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 iii

[Kaul et al., 2025] Kaul, G., Makowski, M. R., Rückert, D., and Braren, R. F. (2025).
Real world federated learning with a knowledge distilled transformer for cardiac ct imaging.
NPJ Digital Medicine, 8(1):1–9.

[Khellaf et al., 2025a] Khellaf, R., Bellet, A., and Josse, J. (2025a).
Federated causal inference from multi-site observational data via propensity score aggregation.

[Khellaf et al., 2025b] Khellaf, R., Bellet, A., and Josse, J. (2025b).
Federated causal inference: Multi-study ate estimation beyond meta-analysis.
AISTATS, 10.

[Lebeda et al., 2025] Lebeda, C., Even, M., Bellet, A., and Josse, J. (2025).
Model Agnostic Differentially Private Causal Inference.
Technical report, arXiv:2505.19589.

[Lei and Ding, 2021] Lei, L. and Ding, P. (2021).
Regression adjustment in completely randomized experiments with a diverging number of covariates.
Biometrika, 108(4):815–828.

References iv

[Moghadas et al., 2021] Moghadas, S. M., Vilches, T. N., Zhang, K., Wells, C. R., Shoukat, A., Singer, B. H., Meyers, L. A., Neuzil, K. M., Langley, J. M., Fitzpatrick, M. C., et al. (2021). **The impact of vaccination on coronavirus disease 2019 (covid-19) outbreaks in the united states.** *Clinical Infectious Diseases*, 73(12):2257–2264.

[Ogier du Terrail et al., 2023] Ogier du Terrail, M. et al. (2023). **Federated learning for predicting neoadjuvant chemotherapy response in triple-negative breast cancer.** *Nature Medicine*, 29(3):456–464.

[Sarthak Pati et al., 2022] Sarthak Pati, Ujjwal Baid, B. E. et al. (2022). **Federated learning enables big data for rare cancer boundary detection.** *Nature Medicine*, 28(5):1035–1043.

[Tudur Smith and Williamson, 2016] Tudur Smith, C, M. M. N. S. I. A. S. M. R. R. R. M. and Williamson, P. (2016). **Individual participant data meta-analyses compared with meta-analyses based on aggregate data.** *Cochrane Database of Systematic Reviews*, (9).

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

References v

[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.