Appendix:
Bayesian Federated Estimation of Causal Effects from Observational Data

Thanh Vinh Vo[1] **Young Lee**[2] **Trong Nghia Hoang**[3] **Tze-Yun Leong**[1]

[1]School of Computing, National University of Singapore
[2]Harvard University
[3]School of Electrical Engineering and Computer Science, Washington State University

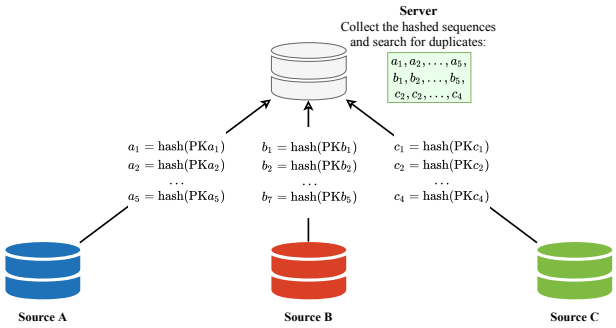

Figure 1: The secure preprocessing procedures to identify duplicated individuals among multiple sources. PKa_i ($i = 1,...,5$), PKb_i ($i = 1,...,7$), PKc_i ($i = 1,...,4$) are the primary keys of each individual in each source. a_i ($i = 1,...,5$), b_i ($i = 1,...,7$), c_i ($i = 1,...,4$) are the hashed sequences of these individuals.

A THE PREPROCESSING PROCEDURE

The assumptions were described briefly in Section 3.2 of the main text. Here we present the preprocessing procedures to remove duplicated individuals.

The preprocessing procedure are summarized as follows. Firstly, each source would use a one-way hash function (such as MD4, MD5, SHA or SHA256) to encrypt each individuals' primary key and then send the hashed sequences to a server. By doing this, the individuals' data are secured. Note that the one-way hash function is agreed among the sources so that they would use the same function. Then, the server collects all hashed sequences from all sources and perform a matching algorithm to see if there exists repeated individuals among different sources. For each repeated individual, the server randomly choose to keep it on a small number (predefined) of sources and inform the other sources to exclude this individual from the training process. The whole procedure is to ensure that an individual does not exists in a huge number of sources, thus prevent learning a biased model. We summarize the procedure in Figure 1.

Assumption 4 and the preprocessing procedure are required for data that are highly repeated in different sources only. For data that are not likely to have a high number of repetitions such as patients from different hospitals of different countries, the above assumption and the preprocessing procedure are not required. Note that the existing methods also need Assumption 4 since they need to combine data and remove repeated individuals.

In this work, we assume that all of the assumptions described in this section are satisfied, and the preprocessing procedure was performed if it is necessary.

B THE FEDERATED EVIDENCE LOWER BOUND

Naively applying variational inference would lead to a non-decomposable ELBO. The proposed ELBO can be decomposed into multiple components, thus enabling federated optimization. We give a full derivation as follows:

$$\begin{aligned}
&\log p(\mathbf{y}_{\text{obs}} \mid \mathbf{X}, \mathbf{w}) \\
&= \log \int p(\mathbf{y}_{\text{obs}}, \mathbf{g}, \Psi, \Sigma \mid \mathbf{X}, \mathbf{w}) d\mathbf{g} d\Psi d\Sigma \\
&= \log \int p(\mathbf{y}_{\text{obs}} \mid \mathbf{g}, \Psi, \Sigma, \mathbf{X}, \mathbf{w}) p(\mathbf{g}, \Psi, \Sigma | \mathbf{X}, \mathbf{w}) d\mathbf{g} d\Psi d\Sigma.
\end{aligned}$$

From Figure 2, we see that $\mathbf{g}, \Psi, \Sigma \perp\!\!\!\perp \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}}$ (for all $\mathsf{s} = 1, 2, \ldots, m$), i.e, $\mathbf{g}, \Psi, \Sigma$ are independent with $\mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}}$ when $\mathbf{y}^{\mathsf{s}}_{\text{obs}}, \mathbf{y}^{\mathsf{s}}_{\text{mis}}$ are not given. Thus, $p(\mathbf{g}, \Psi, \Sigma | \mathbf{X}, \mathbf{w}) = p(\mathbf{g}, \Psi, \Sigma)$.

In addition, from Figure 2, we also have

$$p(\mathbf{y}_{\text{obs}} \mid \mathbf{g}, \Psi, \Sigma, \mathbf{X}, \mathbf{w}) = \prod_{\mathsf{s}=1}^{m} p(\mathbf{y}^{\mathsf{s}}_{\text{obs}} \mid \mathbf{g}, \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}}).$$

Thus,

$$\log p(\mathbf{y}_{\text{obs}} \mid \mathbf{X}, \mathbf{w})$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

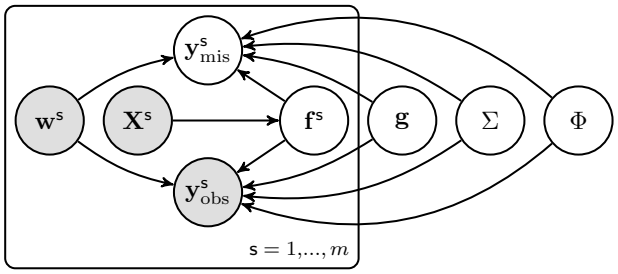

Figure 2: Graphical model that summarizes the proposed framework with treatment $\mathbf{w}^{\mathsf{s}}$, covariate $\mathbf{X}^{\mathsf{s}}$, and the two potential outcomes $\mathbf{y}^{\mathsf{s}}_{\text{mis}}$ and $\mathbf{y}^{\mathsf{s}}_{\text{obs}}$. The quantity $\mathbf{f}^{\mathsf{s}}$ is idiosyncratic to the sources and $\mathbf{g}$ contains shared characteristics across all the sources. Σ and Ψ are shared parameters. Note that this is not a causal graph.

$$
\begin{aligned}
&= \log \int q(\mathbf{g}, \Psi, \Sigma) \prod_{\mathsf{s}=1}^{m} p(\mathbf{y}^{\mathsf{s}}_{\text{obs}} \,|\, \mathbf{g}, \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}}) \\
&\qquad\qquad \times \frac{p(\mathbf{g}, \Psi, \Sigma)}{q(\mathbf{g}, \Psi, \Sigma)} d\mathbf{g} d\Psi d\Sigma \\
&\geq \int q(\mathbf{g}, \Psi, \Sigma) \log \left(\prod_{\mathsf{s}=1}^{m} p(\mathbf{y}^{\mathsf{s}}_{\text{obs}} \,|\, \mathbf{g}, \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}}) \right. \\
&\qquad\qquad \left. \times \frac{p(\mathbf{g}, \Psi, \Sigma)}{q(\mathbf{g}, \Psi, \Sigma)} \right) d\mathbf{g} d\Psi d\Sigma \\
&= \sum_{\mathsf{s}=1}^{m} \mathbb{E}_q[\log p(\mathbf{y}^{\mathsf{s}}_{\text{obs}} \,|\, \mathbf{g}, \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}})] \\
&\qquad - \mathbb{D}_{\text{KL}}[q(\mathbf{g}, \Psi, \Sigma) \| p(\mathbf{g}, \Psi, \Sigma)] \\
&= \sum_{\mathsf{s}=1}^{m} \left(\mathbb{E}_q[\log p(\mathbf{y}^{\mathsf{s}}_{\text{obs}} \,|\, \mathbf{g}, \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}})] \right. \\
&\qquad \left. - \frac{1}{m} \mathbb{D}_{\text{KL}}[q(\mathbf{g}, \Psi, \Sigma) \| p(\mathbf{g}, \Psi, \Sigma)] \right) \\
&= \sum_{\mathsf{s}=1}^{m} \mathbf{L}^{\mathsf{s}},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{L}^{\mathsf{s}} := \; & \mathbb{E}_q[\log p(\mathbf{y}^{\mathsf{s}}_{\text{obs}} \,|\, \mathbf{g}, \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}})] \\
& - \frac{1}{m} \mathbb{D}_{\text{KL}}[q(\mathbf{g}, \Psi, \Sigma) \| p(\mathbf{g}, \Psi, \Sigma)].
\end{aligned}
$$

Hence, we can divide the ELBO into multiple components, which leads to federated training of the model. Without the proposed model, the ELBO cannot be decomposed into multiple components and hence cannot be trained in a federated setting.

C PROOF OF LEMMA 1

Proof. We denote $\xi^{\mathsf{s}}_0 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n_{\mathsf{s}}})$ and $\xi^{\mathsf{s}}_1 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n_{\mathsf{s}}})$. Then, from the model definition (Eq. (5) in the main text), we have

$$
\begin{aligned}
&\begin{bmatrix} y^{\mathsf{s}}_1(0) & \ldots & y^{\mathsf{s}}_{n_{\mathsf{s}}}(0) \\ y^{\mathsf{s}}_1(1) & \ldots & y^{\mathsf{s}}_{n_{\mathsf{s}}}(1) \end{bmatrix} \\
&\quad = \Psi^{\frac{1}{2}} \begin{bmatrix} f^{\mathsf{s}}_1(0) + g^{\mathsf{s}}(0) & \ldots & f^{\mathsf{s}}_{n_{\mathsf{s}}}(0) + g^{\mathsf{s}}(0) \\ f^{\mathsf{s}}_1(1) + g^{\mathsf{s}}(1) & \ldots & f^{\mathsf{s}}_{n_{\mathsf{s}}}(1) + g^{\mathsf{s}}(1) \end{bmatrix} \\
&\qquad + \Sigma^{\frac{1}{2}} \begin{bmatrix} \varepsilon^{\mathsf{s}}_1(0) & \ldots & \varepsilon^{\mathsf{s}}_{n_{\mathsf{s}}}(0) \\ \varepsilon^{\mathsf{s}}_1(1) & \ldots & \varepsilon^{\mathsf{s}}_{n_{\mathsf{s}}}(1) \end{bmatrix}.
\end{aligned}
$$

The above equation is equivalent to the following

$$
\mathbf{Y}^{\mathsf{s}} = \begin{bmatrix} \boldsymbol{\mu}_0 & \boldsymbol{\mu}_1 \end{bmatrix} (\Psi^{\frac{1}{2}})^{\top} + \begin{bmatrix} \varepsilon^{\mathsf{s}}_0 & \varepsilon^{\mathsf{s}}_1 \end{bmatrix} (\Sigma^{\frac{1}{2}})^{\top},
$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_0 &= \mu_0(\mathbf{X}^{\mathsf{s}}) + \mathbf{g}^{\mathsf{s}}_0 + (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \xi^{\mathsf{s}}_0, \\
\boldsymbol{\mu}_1 &= \mu_1(\mathbf{X}^{\mathsf{s}}) + \mathbf{g}^{\mathsf{s}}_1 + (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \xi^{\mathsf{s}}_1.
\end{aligned}
$$

Further expanding the right hand side, we have

$$
\begin{aligned}
\mathbf{Y}^{\mathsf{s}} &= \begin{bmatrix} \mu_0(\mathbf{X}^{\mathsf{s}}) + \mathbf{g}^{\mathsf{s}}_0 & \mu_1(\mathbf{X}^{\mathsf{s}}) + \mathbf{g}^{\mathsf{s}}_1 \end{bmatrix} (\Psi^{\frac{1}{2}})^{\top} \\
&\quad + (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \begin{bmatrix} \xi^{\mathsf{s}}_0 & \xi^{\mathsf{s}}_1 \end{bmatrix} (\Psi^{\frac{1}{2}})^{\top} + \begin{bmatrix} \varepsilon^{\mathsf{s}}_0 & \varepsilon^{\mathsf{s}}_1 \end{bmatrix} (\Sigma^{\frac{1}{2}})^{\top} \\
\operatorname{vec}(\mathbf{Y}^{\mathsf{s}}) &= \left(\Psi^{\frac{1}{2}} \otimes \mathbf{I}_{n_{\mathsf{s}}} \right) \begin{bmatrix} \mu_0(\mathbf{X}^{\mathsf{s}}) + \mathbf{g}^{\mathsf{s}}_0 \\ \mu_1(\mathbf{X}^{\mathsf{s}}) + \mathbf{g}^{\mathsf{s}}_1 \end{bmatrix} \\
&\quad + \left(\Psi^{\frac{1}{2}} \otimes (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \right) \begin{bmatrix} \xi^{\mathsf{s}}_0 \\ \xi^{\mathsf{s}}_1 \end{bmatrix} + (\Sigma^{\frac{1}{2}} \otimes \mathbf{I}_{n_{\mathsf{s}}}) \begin{bmatrix} \varepsilon^{\mathsf{s}}_0 \\ \varepsilon^{\mathsf{s}}_1 \end{bmatrix},
\end{aligned}
$$

where $\operatorname{vec}(\cdot)$ denotes the vectorization of a matrix, which converts a matrix into a column vector.

For the second term on the right hand side of the above equation, note that $\xi^{\mathsf{s}}_0 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n_{\mathsf{s}}})$ and $\xi^{\mathsf{s}}_1 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n_{\mathsf{s}}})$, so we have the following

$$
\begin{aligned}
&\begin{bmatrix} \xi^{\mathsf{s}}_0 \\ \xi^{\mathsf{s}}_1 \end{bmatrix} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{2n_{\mathsf{s}}}) \\
&\left(\Psi^{\frac{1}{2}} \otimes (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \right) \begin{bmatrix} \xi^{\mathsf{s}}_0 \\ \xi^{\mathsf{s}}_1 \end{bmatrix} \\
&\qquad \sim \mathsf{N}\left(\mathbf{0}, \left(\Psi^{\frac{1}{2}} \otimes (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \right) \mathbf{I}_{2N} \left(\Psi^{\frac{1}{2}} \otimes (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \right)^{\top} \right) \\
&\left(\Psi^{\frac{1}{2}} \otimes (\mathbf{K}^{\mathsf{s}})^{\frac{1}{2}} \right) \begin{bmatrix} \xi^{\mathsf{s}}_0 \\ \xi^{\mathsf{s}}_1 \end{bmatrix} \sim \mathsf{N}\left(\mathbf{0}, \Psi \otimes \mathbf{K}^{\mathsf{s}} \right).
\end{aligned}
$$

For the last term, note that $\varepsilon^{\mathsf{s}}_0 \sim \mathsf{N}(0, \mathbf{I}_{n_{\mathsf{s}}}), \varepsilon^{\mathsf{s}}_1 \sim \mathsf{N}(0, \mathbf{I}_{n_{\mathsf{s}}})$, thus

$$
\begin{aligned}
&\begin{bmatrix} \varepsilon^{\mathsf{s}}_0 \\ \varepsilon^{\mathsf{s}}_1 \end{bmatrix} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{2n_{\mathsf{s}}}) \\
&\left(\Sigma^{\frac{1}{2}} \otimes \mathbf{I}_{n_{\mathsf{s}}} \right) \begin{bmatrix} \varepsilon^{\mathsf{s}}_0 \\ \varepsilon^{\mathsf{s}}_1 \end{bmatrix} \sim \mathsf{N}\left(\mathbf{0}, \left(\Sigma^{\frac{1}{2}} \otimes \mathbf{I}_{n_{\mathsf{s}}} \right) \mathbf{I}_{2n} \left(\Sigma^{\frac{1}{2}} \otimes \mathbf{I}_{n_{\mathsf{s}}} \right)^{\top} \right) \\
&\left(\Sigma^{\frac{1}{2}} \otimes \mathbf{I}_{n_{\mathsf{s}}} \right) \begin{bmatrix} \varepsilon^{\mathsf{s}}_0 \\ \varepsilon^{\mathsf{s}}_1 \end{bmatrix} \sim \mathsf{N}\left(\mathbf{0}, \Sigma \otimes \mathbf{I}_{n_{\mathsf{s}}} \right).
\end{aligned}
$$

Consequently,

$$
\operatorname{vec}(\mathbf{Y}^{\mathsf{s}}) \big| \Psi, \Sigma, \mathbf{X}^{\mathsf{s}}, \mathbf{w}^{\mathsf{s}}, \mathbf{g}^{\mathsf{s}}
$$

$$\sim \mathsf{N}\left(\left(\Psi^{\frac{1}{2}} \otimes \mathbf{I}_{n_\mathsf{s}}\right)\begin{bmatrix}\mu_0(\mathbf{X}^\mathsf{s})+\mathbf{g}_0^\mathsf{s}\\ \mu_1(\mathbf{X}^\mathsf{s})+\mathbf{g}_1^\mathsf{s}\end{bmatrix}, \Psi \otimes \mathbf{K}^\mathsf{s} + \Sigma \otimes \mathbf{I}_{n_\mathsf{s}}\right),$$

which implies that

$$\begin{bmatrix}\mathbf{y}^\mathsf{s}(0)\\ \mathbf{y}^\mathsf{s}(1)\end{bmatrix}\Big|\Psi, \Sigma, \mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \mathbf{g}^\mathsf{s}$$
$$\sim \mathsf{N}\left(\left(\Psi^{\frac{1}{2}} \otimes \mathbf{I}_{n_\mathsf{s}}\right)\begin{bmatrix}\mu_0(\mathbf{X}^\mathsf{s})+\mathbf{g}_0^\mathsf{s}\\ \mu_1(\mathbf{X}^\mathsf{s})+\mathbf{g}_1^\mathsf{s}\end{bmatrix}, \Psi \otimes \mathbf{K}^\mathsf{s} + \Sigma \otimes \mathbf{I}_{n_\mathsf{s}}\right).$$

This completes the proof. □

D PROOF OF LEMMA 2

Proof. Following the proof of Lemma 2, we note that if the observed treatment $w_i^\mathsf{s} = 0$, then the mean of $p(y_{i,\text{obs}}^\mathsf{s}|\mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \Psi, \Sigma, \mathbf{g}^\mathsf{s})$ equals to the mean of $p(y_i^\mathsf{s}(0)|\Psi, \Sigma, \mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \mathbf{g}^\mathsf{s})$ and the mean of $p(y_{i,\text{mis}}^\mathsf{s}|\mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \Psi, \Sigma, \mathbf{g}^\mathsf{s})$ equals to the mean of $p(y_i^\mathsf{s}(1)|\Psi, \Sigma, \mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \mathbf{g}^\mathsf{s})$. If the observed treatment $w_i^\mathsf{s} = 1$, then the mean of $p(y_{i,\text{obs}}^\mathsf{s}|\mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \Psi, \Sigma, \mathbf{g}^\mathsf{s})$ equals to the mean of $p(y_i^\mathsf{s}(1)|\Psi, \Sigma, \mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \mathbf{g}^\mathsf{s})$ and the mean of $p(y_{i,\text{mis}}^\mathsf{s}|\mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \Psi, \Sigma, \mathbf{g}^\mathsf{s})$ equals to the mean of $p(y_i^\mathsf{s}(0)|\Psi, \Sigma, \mathbf{X}^\mathsf{s}, \mathbf{w}^\mathsf{s}, \mathbf{g}^\mathsf{s})$. Hence, we have

$$\mu_{\text{obs}}(\mathbf{X}^\mathsf{s}) = (\mathbf{1} - \mathbf{w}^\mathsf{s}) \odot \mathbf{m}_0 + \mathbf{w}^\mathsf{s} \odot \mathbf{m}_1,$$
$$\mu_{\text{mis}}(\mathbf{X}^\mathsf{s}) = \mathbf{w}^\mathsf{s} \odot \mathbf{m}_0 + (\mathbf{1} - \mathbf{w}^\mathsf{s}) \odot \mathbf{m}_1,$$

Similarly, for the covariance matrix, each element in $\mathbf{K}_{\text{obs}}$, $\mathbf{K}_{\text{mis}}$, and $\mathbf{K}_{\text{om}}$ also depends on whether $w_i^\mathsf{s} = 0$ or $w_i^\mathsf{s} = 1$. So each element in these matrices is computed by the following kernel function

$$\begin{aligned}
k_{\text{obs}}(\mathbf{x}_i, \mathbf{x}_j) = &\big[(1-w_i)(1-w_j)\psi_{11} + w_i w_j \psi_{22} \\
&+ (1-w_i)w_j\psi_{12} + w_i(1-w_j)\psi_{21}\big]\mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) \\
&+ \big[(1-w_i)\sigma_{11} + w_i\sigma_{22}\big]\mathbb{1}_{i=j}, \\
k_{\text{mis}}(\mathbf{x}_i, \mathbf{x}_j) = &\big[w_i w_j \psi_{11} + (1-w_i)(1-w_j)\psi_{22} \\
&+ (1-w_i)w_j\psi_{21} + w_i(1-w_j)\psi_{12}\big]\mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) \\
&+ \big[w_i\sigma_{11} + (1-w_i)\sigma_{22}\big]\mathbb{1}_{i=j}, \\
k_{\text{om}}(\mathbf{x}_i, \mathbf{x}_j) = &\big[(1-w_i)(1-w_j)\psi_{21} + w_i w_j \psi_{12} \\
&+ (1-w_i)w_j\psi_{22} + w_i(1-w_j)\psi_{11}\big]\mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) \\
&+ \big[(1-w_i)\sigma_{21} + w_i\sigma_{12}\big]\mathbb{1}_{i=j},
\end{aligned}$$

where ψ_{ab} and σ_{ab} are the (a, b)–th elements of Ψ and Σ, respectively.

This completes the proof. □

E EVALUATION METRICS

The two evaluation metrics reported in our experiments are defined as follows: (i) precision in estimation of heterogeneous effects (PEHE):

$$\epsilon_{\text{PEHE}} := \sum_{\mathsf{s}=1}^{m}\sum_{i=1}^{n_\mathsf{s}}(\tau_i^\mathsf{s} - \hat{\tau}_i^\mathsf{s})^2/(mn_\mathsf{s})$$

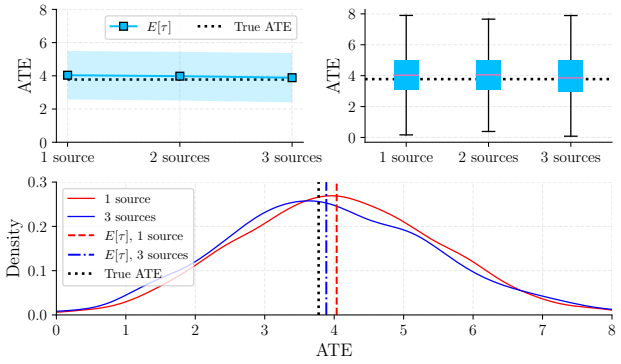

Figure 3: The estimated ATE distribution on source #1 of IHDP dataset. The dotted black lines represent the true ATE.

for evaluating ITE, and (ii) absolute error:

$$\epsilon_{\text{ATE}} := |\tau - \hat{\tau}|$$

for evaluating ATE, where τ_i^s and τ are the *true* ITE and *true* ATE, respectively, and $\hat{\tau}_i^\mathsf{s}$, $\hat{\tau}$ are their estimates.

F ADDITIONAL EXPERIMENTAL RESULTS

In this section, we present some additional results which was skipped in the main text due to limited space.

F.1 SYNTHETIC DATA: DATA-2

In this section, we present additional experimental results on DATA-2. Again, those results were skipped in the main text due to limited space. In Table 1, we present additional results of the baselines trained locally (loc) and the baselines trained with bootstrap aggregating (agg). Similar to the experiments on DATA-1 presented in the main text, the results on DATA-2 also show that FedCI achieves much lower errors, especially the error in predicting ITE.

F.2 IHDP DATASET

In this section, we present additional experimental results on the IHDP dataset. The results here were not presented in the main text due to limited space. In Table 2, we present additional results of the baselines trained locally (loc) and the baselines trained with bootstrap aggregating (agg). Similar to the experiments on synthetic data, the results presented here show that FedCI achieves much smaller errors. The reason is because FedCI has access to all the data sources in a federated fashion while the 'baselines trained locally' (loc) and the 'baselines trained with bootstrap aggregating' (agg) only have access to a local data source.

Similar to the experiment on synthetic data, the estimated distribution of ATE in the first source ($\mathsf{s} = 1$) is presented

Table 1: Out-of-sample errors on DATA-2 where top-3 performances are highlighted in bold (lower is better). The dashes (—) in 'loc' and 'agg' indicate that the numbers are the same as those of 'com'.

Method	The error of ITE ($\sqrt{\epsilon_{\text{PEHE}}}$)			The error of ATE (ϵ_{ATE})		
	1 source	3 sources	5 sources	1 source	3 sources	5 sources
BART_{loc}	—	18.4±0.3	18.3±0.2	—	3.37±0.7	2.90±0.6
$\text{X-Learner}_{\text{loc}}$	—	22.7±0.5	22.8±0.5	—	3.55±1.3	3.09±0.8
$\text{R-Learner}_{\text{loc}}$	—	26.3±0.2	26.1±0.2	—	19.7±0.3	19.5±0.3
$\text{OthoRF}_{\text{loc}}$	—	38.3±1.4	40.0±0.9	—	4.09±0.9	4.40±1.2
$\text{TARNet}_{\text{loc}}$	—	37.6±0.6	37.1±0.4	—	7.31±0.4	7.25±0.3
$\text{CFR Wass}_{\text{loc}}$	—	37.2±0.7	37.0±0.5	—	7.24±0.3	7.12±0.2
$\text{CFR MMD}_{\text{loc}}$	—	37.2±0.6	36.8±0.4	—	7.21±0.4	7.11±0.3
$\text{CEVAE}_{\text{loc}}$	—	21.4±0.7	19.8±0.6	—	2.11±0.4	1.97±0.2
BART_{agg}	—	17.9±0.2	17.7±0.2	—	3.91±0.8	3.15±0.7
$\text{X-Learner}_{\text{agg}}$	—	18.2±0.4	17.1±0.2	—	3.43±1.3	3.07±0.8
$\text{R-Learner}_{\text{agg}}$	—	26.2±0.3	26.1±0.2	—	19.7±0.4	19.6±0.3
$\text{OthoRF}_{\text{agg}}$	—	25.0±1.3	17.3±0.6	—	4.56±1.1	**1.30±0.4**
$\text{TARNet}_{\text{agg}}$	—	36.5±0.3	36.1±0.3	—	7.26±0.3	7.18±0.3
$\text{CFR Wass}_{\text{agg}}$	—	35.2±0.5	35.0±0.3	—	7.13±0.3	6.97±0.2
$\text{CFR MMD}_{\text{agg}}$	—	35.2±0.5	35.1±0.4	—	7.10±0.4	7.05±0.2
$\text{CEVAE}_{\text{agg}}$	—	19.2±0.8	18.3±0.7	—	2.02±0.3	1.91±0.4
BART_{com}	**18.0±0.4**	**17.7±0.2**	17.4±0.1	**3.54±1.3**	2.94±0.8	**1.84±0.5**
$\text{X-Learner}_{\text{com}}$	21.1±0.9	17.9±0.4	**16.2±0.2**	4.55±1.4	3.29±1.0	2.37±0.8
$\text{R-Learner}_{\text{com}}$	25.9±0.6	23.5±0.5	21.3±0.4	19.0±0.8	15.6±0.7	12.3±0.6
$\text{OthoRF}_{\text{com}}$	37.8±2.7	**10.7±0.5**	**9.83±0.5**	7.88±2.2	**1.99±0.4**	2.36±0.6
$\text{TARNet}_{\text{com}}$	36.1±0.4	35.5±0.2	35.0±0.2	7.11±0.4	7.10±0.3	7.08±0.2
$\text{CFR Wass}_{\text{com}}$	35.1±0.4	34.5±0.2	34.1±0.2	7.10±0.4	7.01±0.3	6.90±0.2
$\text{CFR MMD}_{\text{com}}$	35.1±0.4	35.0±0.2	34.9±0.2	7.12±0.4	7.02±0.3	7.01±0.2
$\text{CEVAE}_{\text{com}}$	**20.1±0.5**	18.4±0.6	16.6±0.6	**1.50±0.3**	**1.38±0.4**	1.89±0.2
FedCI	**9.28±0.4**	**6.34±0.2**	**5.53±0.1**	**2.37±0.5**	**1.47±0.4**	**0.74±.2**

Table 2: Out-of-sample errors on IHDP dataset where top-3 performances are highlighted in bold (lower is better). The dashes (—) in 'loc' and 'agg' indicate that the numbers are the same as those of 'com'.

Method	The error of ITE ($\sqrt{\epsilon_{\text{PEHE}}}$)			The error of ATE (ϵ_{ATE})		
	1 source	2 sources	3 sources	1 source	2 sources	3 sources
BART_{loc}	—	5.83±2.6	6.56±3.3	—	2.09±0.9	1.38±0.5
$\text{X-Learner}_{\text{loc}}$	—	4.14±1.5	4.54±1.9	—	1.51±0.7	0.77±0.5
$\text{R-Learner}_{\text{loc}}$	—	6.35±1.9	6.16±2.0	—	2.13±0.7	1.44±0.3
$\text{OthoRF}_{\text{loc}}$	—	4.33±1.6	4.59±1.9	—	1.10±0.6	0.75±0.3
$\text{TARNet}_{\text{loc}}$	—	3.71±1.0	3.83±1.1	—	1.31±0.5	0.98±0.4
$\text{CFR Wass}_{\text{loc}}$	—	3.35±0.8	3.12±0.7	—	0.87±0.5	0.82±0.4
$\text{CFR MMD}_{\text{loc}}$	—	3.40±0.9	3.15±1.2	—	1.17±0.5	0.63±0.3
$\text{CEVAE}_{\text{loc}}$	—	3.78±0.7	3.93±0.8	—	1.91±0.3	2.37±0.2
BART_{agg}	—	4.05±1.9	3.69±1.8	—	2.09±1.0	1.30±0.5
$\text{X-Learner}_{\text{agg}}$	—	3.98±1.5	4.28±1.9	—	1.51±0.7	0.83±0.5
$\text{R-Learner}_{\text{agg}}$	—	4.76±1.3	4.46±1.6	—	1.92±0.5	1.41±0.2
$\text{OthoRF}_{\text{agg}}$	—	3.40±1.1	4.26±1.9	—	0.87±0.3	1.20±0.6
$\text{TARNet}_{\text{agg}}$	—	3.52±0.9	3.81±1.2	—	1.23±0.4	0.95±0.4
$\text{CFR Wass}_{\text{agg}}$	—	3.21±0.7	2.93±0.9	—	0.80±0.3	0.71±0.2
$\text{CFR MMD}_{\text{agg}}$	—	3.17±0.8	2.91±1.3	—	1.12±0.5	0.57±0.3
$\text{CEVAE}_{\text{agg}}$	—	3.63±0.7	3.73±0.5	—	0.92±0.2	0.84±0.5
BART_{com}	5.98±2.7	4.32±2.1	4.04±2.0	1.80±1.1	2.09±1.1	1.21±0.6
$\text{X-Learner}_{\text{com}}$	4.22±1.6	4.15±1.5	4.06±1.8	1.64±0.7	1.93±0.8	0.84±0.4
$\text{R-Learner}_{\text{com}}$	6.97±2.1	4.43±1.4	4.47±1.7	3.15±0.5	1.34±0.5	1.10±0.3
$\text{OthoRF}_{\text{com}}$	4.49±1.9	3.81±1.3	3.75±1.5	1.86±0.8	1.61±0.6	1.56±0.8
$\text{TARNet}_{\text{com}}$	4.50±1.4	3.15±0.8	3.79±1.1	**1.52±0.5**	1.18±0.4	0.91±0.3
$\text{CFR Wass}_{\text{com}}$	**4.37±1.2**	2.93±0.6	2.85±0.9	**1.18±0.7**	**0.72±0.2**	0.67±0.1
$\text{CFR MMD}_{\text{com}}$	4.43±1.3	**2.85±0.6**	**2.83±1.1**	2.32±0.8	**0.63±0.2**	**0.54±0.2**
$\text{CEVAE}_{\text{com}}$	**3.16±0.6**	**2.34±0.6**	**2.31±0.7**	2.02±0.4	**0.53±0.1**	**0.48±0.2**
FedCI	**2.88±0.8**	**2.36±0.5**	**2.35±0.6**	**1.43±0.7**	1.03±0.4	**0.51±0.2**

in Figure 3. Again, the figures show that the true ATE is inside the estimated interval and the estimated mean ATE shifts towards its true value (dotted lines) when more data sources are used.