Federated learning for causal inference using deep generative disentangled models

Alejandro Almodóvar, Juan Parras, Santiago Zazo
Information Processing and Telecommunication Center, ETSI de Telecomunicación
Universidad Politécnica de Madrid
alejandro.almodovar@upm.es

Abstract

In the context of decentralized and privacy-constrained healthcare data settings, we introduce an innovative approach to estimate individual treatment effects (ITE) via federated learning. Emphasizing the critical importance of data privacy in healthcare, especially when drawing on data from various global hospitals, we address challenges arising from data scarcity and specific treatment assignment criteria influenced by the availability of the medication of interest. Our methodology uses federated learning applied to neural network-based generative causal inference models to bridge the gap between decentralized and centralized ITE estimation on a benchmark dataset.

1 Introduction

Causal inference is vital to understand the effects of treatment in diverse fields, but in decentralized, privacy-constrained settings, it presents significant challenges. We aim to estimate individual treatment effects (ITEs) while respecting stringent privacy constraints.

In distributed environments such as healthcare facilities, changes in propensity scores and in the distribution of covariates between nodes pose obstacles. Limited data availability further complicates unbiased causal effect estimation. To address these challenges, we adopt federated learning, specifically `FederatedAveraging` (`FedAvg`) [1], and utilize TEDVAE (Treatment Effect with Disentangled Variational Autoencoder) [2] for the estimation of the treatment effect. In the domain of federated causal inference, researchers have explored various techniques: parametric models [3] and Gaussian Processes (GPs) [4] and Random Fourier Features (RFF) [5] have been adapted for federated settings. These algorithms often rely on asymptotic properties and may not be suitable for scenarios with limited sample sizes and changes in treatment assingment distributions, which is often the case in the medical setting. We extend the principles of federated learning to the domain of causal inference, acknowledging a limited number of samples and changes in the distribution of treatment assignment.

2 Problem definition

2.1 Local causal inference

Consider the data set $\mathcal{D} = \{\mathbf{X}_i, T_i, Y_i\}_{i=1}^{N}$, where the subindex $i \in \mathcal{M} = \{1, ..., N\}$ represents the index an individual datapoint and $N = |\mathcal{M}|$ is the total number of data points in the data set. Assume that the samples are i.i.d. observations: $\mathcal{D} \overset{iid}{\sim} \mathbb{P}$. In this notation, $\mathbf{X}_i \in \mathcal{X} \subseteq \mathbb{R}^{D_x}$ is a vector of covariates, $T_i \in \{0, 1\}$ is the treatment , and $Y_i \in \mathbb{R}$ represents the *outcome*. Let us also define the individual causal effect of T_i on Y_i ($ITE \equiv \tau_i$), following the Neyman-Rubin potential outcome framework [6], as: $\tau_i \equiv Y_i(T_i = 1) - Y_i(T_i = 0)$

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

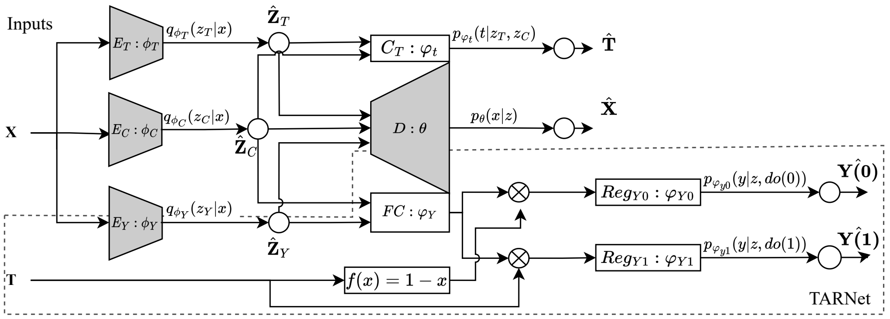

Figure 1: TEDVAE model. $\circ$ represents sampling and $\otimes$ the product.

Conventional causal inference methods [6–18] estimate individual and/or average treatment effect conditioning on covariates, assuming that the data meets the standard unconfoundedness, positivity, consistency, and *no interference* (these last two are SUTVA) assumptions, following backdoor criterion [19]: $\hat{\tau}(\mathbf{x}_i) = \mathbb{E}[Y|T = 1, \mathbf{X} = \mathbf{x}_i] - \mathbb{E}[Y|T = 0, \mathbf{X} = \mathbf{x}_i]$.

2.2 Distributed causal inference

Suppose that we have K information processing nodes, each $k \in \{1, ..., K\}$ node with a data set $\mathcal{D}^k = \{\mathbf{X}_i^k, T_i^k, Y_i^k\}_{i \in \mathcal{M}^k \subset \mathcal{M}}$, where $\mathcal{M}^k$ is the set of patient indices, $N^k = |\mathcal{M}^k|$ is the number of samples from node k and $N = \sum_{i=1}^K N^k$ (there are not repeated patients: $\mathcal{M}^k \cap \mathcal{M}^j = \varnothing$). Also, $\mathbf{X}_i^k \in \mathcal{X} \subseteq \mathbb{R}^{D_{xk}}$, where D_{xk} is the number of covariates of each node, $T_i^k \in \{0, 1\}$ and $Y_i^k \in \mathbb{R}$. The sets of patient indices treated and control (untreated) patients of node k are $\mathcal{T}^k$ and $\mathcal{C}^k$. The number of treated and control patients in each node is $N_T^k = |\mathcal{T}^k|$ and $N_C^k = |\mathcal{C}^k|$, respectively.

Data collected in the different nodes can be distributed non-identically ($\mathcal{D}^j \overset{iid}{\sim} \mathbb{P}^j$, $\mathcal{D}^k \overset{iid}{\sim} \mathbb{P}^k$, $\mathbb{P}^j \neq \mathbb{P}^k$, with $j, k \in \{1, ..., K\}$, $j \neq k$). Let us define three conditions to study in distributed causal inference [3]: (**Condition 1**) The set of covariates is the same in all nodes: $D_{xj} = D_{xk}$, $\mathcal{X}^j = \mathcal{X}^k$, (**Condition 2**) the covariate distribution is stable across nodes: $p^j(\mathbf{X}) = p^k(\mathbf{X})$ and (**Condition 3**) the propensity score is stable between nodes: $p^j(T|\mathbf{X}) = p^k(T|\mathbf{X})$.

2.3 Conditions of our problem

In this text, we will assume that Condition 1 is valid, but not Conditions 2 and 3. We want to focus on a scenario where some underdeveloped countries do not have access to some drugs. This strategy can help to estimate the effect of a treatment from data in developed countries. This imbalance causes very important changes in the propensity score and in the distribution of covariates.

In addition, we assume that the classical assumptions of causal inference are satisfied for the joint dataset $\mathcal{D} = \{\mathbf{X}_i, T_i, Y_i\}_{i=1}^N \sim \mathbb{P}_{data}$. Ideally, the causal effects of T on the outcome could be estimated from the union of all datasets, as if all data were located in the same centralized node.

On the other hand, we consider that individual-level data cannot be shared between nodes, so this joint distribution is not available in each node. Due to the distribution shift across nodes, the estimated causal effect will be different in each node and none of them will have to coincide with the estimate in the centralized case. Furthermore, the decentralization of information implies that the number of samples in each node is less than the total number of samples, which increases the variance of the estimators in datasets with a limited number of samples. We must take into account that due to propensity score shift, it is more difficult to meet the positivity assumption in each isolated node even when the assumption is met for the entire dataset $\mathcal{D}$.

3 Federated learning method for causal inference

Definitions. Let us define some terms for the explanation of this section: parameters of neural networks are expressed with Greek letters, as shows Figure 1. The letter Ω^k refers to the set of all parameters of the neural network model (TEDVAE) of node k and $\Omega = \{\Omega^1, ..., \Omega^k\}$ refers to the set of parameters of the models of all nodes. Let us define $\Theta^k = \{\phi_T^k, \phi_C^k, \phi_Y^k, \theta^k, \varphi_t^k, \varphi_Y^k\}$; then the set of all parameters is $\Omega^k = \{\Theta^k, \phi_{Y1}^k, \phi_{Y0}^k\}$. The output of a module (ϑ) with input $\mathbf{x}_i$ is expressed in this text as $f_\vartheta(\mathbf{x}_i)$. The subscript S refers to the server parameters. Furthermore, consider that N_T^S and N_C^S are the sum of treated and control patients in all nodes, respectively.

The prediction problem in Federated Learning has a global objective function to minimize:

$$\mathcal{L}(\Omega; \mathcal{D}) \equiv \sum_{k=1}^{K} \frac{N^k}{N} L^k(\Omega^k; \mathcal{D}^k) \text{ where } L^k(\Omega^k; \mathcal{D}^k) = \sum_{i \in \mathcal{M}^k} \frac{1}{N^k} l[f^k(\mathbf{x}_i^k, t_i^k; \Omega^k), y_i^k] \tag{1}$$

If the data were IID and the number of samples was large enough, $\mathbb{E}_{\mathcal{M}^k}[L^k(\Omega^k; \mathcal{D}^k)] = \mathcal{L}(\Omega; \mathcal{D})$ over any node k. However, our problem considers the non-IID setting and the limited number of samples. In the `FedAvg` algorithm, the training process in each node minimizes the local objective function $L^k(\Omega^k)$: $\Omega_{t+1}^k = \Omega_t^k - \eta \nabla_{\Omega_t^k} \frac{1}{N^k} \sum_{i \in \mathcal{M}^k} l(f^k(\mathbf{x}_i^k, t_i^k; \Omega_t^k), y_i)$

Vanilla (standard) `FedAvg` computes in a central server a weighted average of the parameters of each node, weighting them with the number of samples in each node:

$$\Omega_{t+1}^S = \sum_{k=1}^{K} \frac{N^k}{N} \Omega_{t+1}^k = \Omega_t^S - \eta \sum_{k=1}^{K} \frac{1}{N} \nabla_{\Omega_t^k} \sum_{i \in \mathcal{M}^k} l(f^k(\mathbf{x}_i^k, t_i^k; \Omega_t^k), y_i^k) \tag{2}$$

The objective function $L(\Omega^k; \mathcal{D}^k)$ of the TEDVAE [2] model parameterized in Figure 1 is in Eqs. 3, 4. The objective of this model is to disentangle the covariates into instrumental variables ($\mathbf{z}_t$), confounders ($\mathbf{z}_c$), and adjustment variables ($\mathbf{z}_y$), achieving a partial discovery of the causal graph.

$$\begin{aligned} L_{\text{TEDVAE}}(\Omega; \mathcal{D}) = \frac{1}{N} \sum_{i \in \mathcal{M}} & l_{\text{ELBO}}(\mathbf{x}_i, y_i, t_i; \Theta) \\ + \alpha_t \mathbb{E}_{q_{\phi_t} q_{\phi_c}} & \left[\log p_{\varphi_t}\left(t_i \mid \mathbf{z}_{t,i}, \mathbf{z}_{c,i}\right)\right] \\ + \alpha_y \mathbb{E}_{q_{\phi_y} q_{\phi_c}} & \left[\log p_{\varphi_y}\left(y_i \mid t_i, \mathbf{z}_{c,i}, \mathbf{z}_{y,i}\right)\right] \end{aligned} \tag{3}$$

$$\begin{aligned} l_{\text{ELBO}}(\mathbf{x}, y, t; \Theta) = {} & \mathbb{E}_{q_{\phi_c} q_{\phi_t} q_{\phi_y}}\left[\log p_\theta\left(\mathbf{x} \mid \mathbf{z}_t, \mathbf{z}_c, \mathbf{z}_y\right)\right] \\ & - D_{KL}\left(q_{\phi_t}\left(\mathbf{z}_t \mid \mathbf{x}\right) \| p_{\theta_t}\left(\mathbf{z}_t\right)\right) \\ & - D_{KL}\left(q_{\phi_c}\left(\mathbf{z}_c \mid \mathbf{x}\right) \| p_{\theta_c}\left(\mathbf{z}_c\right)\right) \\ & - D_{KL}\left(q_{\phi_y}\left(\mathbf{z}_y \mid \mathbf{x}\right) \| p_{\theta_y}\left(\mathbf{z}_y\right)\right). \end{aligned} \tag{4}$$

We have omitted the superscript k in Figure 1 and Eqs. 3, 4, for clarity. The terms $\alpha_t, \alpha_y \in \mathbb{R}^+$ are hyperparameters. The Evidence Lower Bound (ELBO) [20, 21], is composed by Gaussian priors ($\mathcal{N}(\mathbf{0}, \mathbf{I})$) and Gaussian posteriors, with the particularity of the inclusion of three Kullback-Leibler divergence terms due to the decomposition of latent space. Note that $p_{\varphi_y}(y_i | t_i, \mathbf{z}_{c,i}, \mathbf{z}_{y,i}) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i) \quad \text{where} \quad (\hat{\mu}_i, \hat{\sigma}_i) = t_i \cdot f_{\varphi_{Y1}}(\mathbf{z}_{c,i}, \mathbf{z}_{y,i}) + (1 - t_i) \cdot f_{\varphi_{Y0}}(\mathbf{z}_{c,i}, \mathbf{z}_{y,i})$.

The key point of our *propensity adaptation*, is in the expectation included in Eq. 3 can be rewritten due to the previous equation as: $\frac{1}{N} \sum_{i \in \mathcal{M}} \mathbb{E}_{q_{\phi_y} q_{\phi_c}}\left[\log p_{\varphi_y}\left(y_i \mid t_i, \cdot\right)\right] = \frac{1}{N_T} \sum_{i \in \mathcal{T}} \mathbb{E}_{q_{\phi_y} q_{\phi_c}} \log p_{\varphi_{Y1}}(y_i | t_i = 1, \cdot) + \frac{1}{N_C} \sum_{i \in \mathcal{C}} \mathbb{E}_{q_{\phi_y} q_{\phi_c}} \log p_{\varphi_{Y0}}(y_i | t_i = 0, \cdot)$

Eq. 2 shows that, for a fully connected model, sharing the parameters of the nodes after computing the gradient descent in each node separately is equivalent to sharing the gradients of the nodes and computing the gradient descent averaging the gradients in the server. However, due to the particular optimization of TARNet modules Reg_{Y0} and Reg_{Y1}, this equivalence does not hold: if we compute the averaging as in Eq. 2 (*Vanilla* `FedAvg`), the averaged parameters are in Eq. 5

$$\begin{aligned} \varphi_{Y1_{t+1}}^S &= \varphi_{Y1_t}^S - \eta \textstyle\sum_{k=1}^{K} \frac{N^k}{N \cdot N_T^k} \nabla_{\varphi_{Y1}} \sum_{i \in \mathcal{T}^k} l_T(\varphi_{Y1}^k, \phi_y^k, \phi_c^k; \mathcal{D}^k), \\ \varphi_{Y0_{t+1}}^S &= \varphi_{Y0_t}^S - \eta \textstyle\sum_{k=1}^{K} \frac{N^k}{N \cdot N_C^k} \nabla_{\varphi_{Y0}} \sum_{i \in \mathcal{C}^k} l_C(\varphi_{Y0}^k, \phi_y^k, \phi_c^k; \mathcal{D}^k) \end{aligned} \tag{5}$$

Proposed algorithm [1]. Our approach to mitigate propensity score imbalances across nodes is an adaptation of `FedAvg` [1] over a neural network-based model for causal inference called TEDVAE [2], which we call *propensity adaptation*.

To avoid the discrepancy of Eq. 5 weights $\left(\frac{N^k}{N \cdot N_T^k}\right)$ with the mean of gradients $\left(\frac{N_T^k}{N_T}\right)$,our algorithm weights the parameters of the Reg_{Y1} and Reg_{Y0} regressors by the number of treated and control patients, respectively (N_T^k, N_C^k). The weight of the regressors depends on the control/treated samples in each node. The process can stop in any epoch between the averaging moments. An implementation of `FedAvg` without *propensity adaptation* may lead, in a limiting case in which there are no

[1] Code available in `https://github.com/aalmodovares/federated_tedvae`

	Setting A		Setting B	
	node 1	node 2	node 1	node 2
TV Cen	1.16(0.26)		3.07(0.72)	
TV Fed	1.18(0.31)	1.20(0.31)	3.55(0.86)	3.41(0.69)
TV Fed V	1.15(0.37)	1.15(0.29)	3.61(0.80)	3.50(0.72)
TV Iso	1.21(0.41)	1.27(0.29)	4.83(0.81)	4.64(0.65)
CausalRFF	2.99(1.73)	2.96(1.72)	6.88(1.39)	6.80(1.37)
FedCI	2.56(0.45)	2.63(0.83)	4.88(1.95)	4.94(2.16)

Table 1: Out-of-sample PEHE results for the original distribution sampled dataset of 83 samples in each node for IHDP setting A and B respectively. **Lower is better.** With equilibrated nodes, *propensity adaptation* and *Vanilla* `FedAvg` have similar metrics.

treated patients from a node, to averaging the parameters with a module that has not been trained even once locally.

4 Experiments on IHDP

The comparison will be carried out comparing our implementation of *propensity adaptation* of `FedAvg` on TEDVAE (**TV Fed**) with centralized TEDVAE (TV cen), which trains with all dataset $\mathcal{D}$); the node-wise isolated training (TV iso), in which each node trains with their data separately, without sharing any information; the *Vanilla* `FedAvg` implementation (TV Fed V), which does not consider propensity imbalances; the Federated Causal Inference method of [4] based on GPs (FedCI), and the CausalRFF method of [5] based on RFF (CausalRFF).

The experiments have been carried out on 20 replications of IHDP [9], semi-synthetic datasets commonly used to evaluate causal inference methods, where the outcome is a known combination of the input data. Since the potential outcomes are known, the real value of the ITE can be calculated, and the Precision in Estimation of Heterogeneous Effects (PEHE) can be presented as an evaluation metric:

$PEHE = \mathbb{E}[(\hat{\tau}(x) - \tau(x))^2]$, where $\hat{\tau}(x)$ is the estimated treatment effect for subgroup x, and $\tau(x)$ is the true treatment effect for that subgroup.

Algorithm 1: *Prop. Adap.* of `FedAvg`

```
Input  : List of nodes $C_1, ..., C^K$ and their
         parameters $\Omega^1, ..., \Omega^K$
Output: List of node parameters $\Omega^1, ..., \Omega^K$
Server execution:
Initialize global model parameters $\Omega_0^S$;
$\Omega_0^k \leftarrow \Omega_0^S$ for $k$ in $\{1, ..., K\}$
for n in $n_{rounds}$ do
    for t in $n_{fedavg}$ do
        for each node $C^k$ in parallel do
            $\Omega_{t+1}^k, \leftarrow$ TrainNode($k, \Omega_t^k, \mathcal{D}^k$)
        end
    end
    send $\{\Omega_{t+1}^k\}_{k=1}^K$ to server
    $\Omega_{t+1}^S = \left\{ \begin{array}{l} \sum_{k=1}^K \frac{N^k}{N} \Theta_{t+1}^k, \\ \sum_{k=1}^K \frac{N_T^k}{N_T^S} \varphi_{Y1_{t+1}}^k, \\ \sum_{k=1}^K \frac{N_C^k}{N_C^S} \varphi_{Y1_{t+1}}^k \end{array} \right\}$ // Avg
    for each node $C^k$ in parallel do
        $\Omega_{t+1}^k, \leftarrow \{\Theta_{t+1}^S, \varphi_{Y0_{t+1}}^S, \varphi_{Y0_{t+1}}^S\}$
    end
end
TrainNode(k, $\Omega_t^k, \mathcal{D}^k$):
    $\Omega_{t+1}^k \leftarrow \Omega_t^k - \eta \nabla_{\Omega_t^k} L^k(\Omega_t^k; \mathcal{D}^k)$ // GD
  return $\Omega_{t+1}^k$
```

There are two settings of data generation in IHDP: setting A, where both potential outcomes are linear combinations of the covariates and the treatment, and setting B, where one of the potential outcomes is an exponential combination of the features. The surface of ITE is more complex in setting B. Two experiments are presented for IHDP datasets setting A and B respectively. The mean and standard deviation presented in PEHE results come from the evaluation of 20 IHDP replications.

Experiment 1: Stable propensity score. In this experiment there are two nodes with a small set of randomly sampled patients (83 patients in each node), so the treatment distribution is the same in both nodes (conditions 2 and 3 holds).

Table 1 shows that, since the propensity score is the same in both nodes, the *propensity score adaptation* does not provide any improvement with respect to *Vanilla* `FedAvg`. The performance of all versions of TEDVAE outperform other methods based in GPs and RFF. Both the *Vanilla* `FedAvg` implementation and our `FedAvg` *propensity adaptation* improve prediction performance with respect to isolated TEDVAE in IHDP setting B, where the treatment effect function is more complex.

Experiment 2: Imbalanced propensity score. In this experiment, we conducted several subexperiments with a fixed a set of patients (546 patients in total), following the original distribution of the original dataset (102 treated patients and 444 untreated patients) and divided the patients into

		Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
		Node 1	Node 2	Node 1	Node 2	Node 1	Node 2	Node 1	Node 2
		51/222	51/222	11/262	91/182	1/272	101/172	0/273	102/171
Setting A	TV Cen	0.75(0.17)							
	TV Fed	0.74(0.25)	0.79(0.23)	0.68(0.15)	0.78(0.25)	0.86(0.25)	0.85(0.26)	0.90(0.30)	0.81(0.21)
	TV Fed V	0.73(0.25)	0.78(0.25)	0.75(0.15)	0.84(0.22)	1.04(0.30)	0.94(0.22)	1.21(0.41)	0.95(0.28)
	TV Iso	0.85(0.25)	0.82(0.27)	0.81(0.23)	0.83(0.18)	2.56(0.86)	0.81(0.19)	2.89(1.02)	0.85(0.25)
	CausalRFF	2.93(1.48)	2.82(1.52)	2.94(1.49)	2.88(1.79)	3.25(1.79)	2.95(1.89)	3.35(1.92)	3.25(1.79)
	FedCI	1.89(1.00)	2.36(1.32)	2.13(0.98)	1.68(0.98)	3.35(1.66)	1.75(0.89)	3.55(1.82)	2.85(2.09)
Setting B	TV Cen	2.08(0.19)							
	TV Fed	2.49(0.39)	2.47(0.41)	3.23(0.51)	2.36(0.33)	3.36(0.50)	2.38(0.26)	3.43(0.51)	2.21(0.28)
	TV Fed V	2.87(0.47)	2.86(0.42)	3.62(0.82)	2.71(0.56)	3.99(0.81)	2.86(0.42)	3.96(0.62)	2.53(0.40)
	TV Iso	2.54(0.45)	2.37(0.24)	3.60(0.66)	2.37(0.24)	5.08(1.47)	2.38(0.35)	5.45(1.13)	2.15(0.23)
	FedCI	3.75(1.95)	4.00(1.37)	4.34(1.30)	4.13(1.54)	5.61(1.22)	3.58(1.19)	5.72(1.25)	3.89(1.36)
	CausalRFF	5.50(1.42)	5.41(1.33)	5.48(1.31)	5.50(1.34)	5.48(1.45)	5.40(1.29)	5.60(1.61)	5.40(1.29)

Table 2: Out-of-sample PEHE results on IHDP settings A and B with increasing imbalances. **Lower is better.** *Propensity adaptation* achieves better metrics that *Vanilla* `FedAvg` and node-wise isolated training of TEDVAE as the imbalance increases.

the two nodes. Both nodes have the same number of patients, but we are decreasing progressively the number of treated patients in one node at the same time that they increase in the other node. The contrary is true for control patients. We start from the original distribution of treated patients in the dataset (51 treated patients and 222 untreated patients in each node), and then we unbalance the number of treated and untreated, keeping the total number of patients in each node the same. The aim of this experiment is to observe how the ITE estimation errors vary as the imbalance increases, reaching the limit experiment where one of the nodes has no treated patients, where *Vanilla* `FedAvg` and node-wise isolated training obtain the worst results.

Table 2 shows that the *propensity adaptation* of `FedAvg` remains close to the centralized case, while isolated structures and *Vanilla* `FedAvg` offer a performance that worsens as the imbalance in the propensity score increases. Note that *Vanilla* `FedAvg` not only obtains worse results in the most unbalanced node (which does not have treated patients), but also worsens the prediction in node 2, where a sufficient number of treated and untreated patients is present; while the *propensity adaptation* manages to stay closer to the centralized case in both nodes. In the same way, we can observe that the performance of our algorithm is superior to that of FedCI and CausalRFF for both cases.

5 Conclusion

In conclusion, this study has demonstrated the remarkable potential of federated learning as an effective and privacy-preserving approach in the context of causal inference in sensitive domains such as healthcare, where centralized data processing is impractical due to privacy restrictions, and propensity score and covariate distributions vary between nodes. By comparing the performance of federated learning with node-wise isolated training and centralized training, we have consistently observed that federated learning achieves better results in terms of Predicted Error in Heterogeneous Effect (PEHE), especially when the treatment assignment criteria is very different between nodes. The federated implementation of TEDVAE outperforms the methods of Federated Causal Inference and CausalRFF used for comparison, since it allows modeling complex non-linear relationships between variables and complex surfaces of treatment effects, in addition to partially discovering the causal graph through disentanglement.

Acknowledgements

This research received the support of *Genomed4all* and *Synthema* projects, funded by a European Union´s Horizon Europe research and innovation programme, under grant agreement IDs 101017549 and 101095530 respectively. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[2] W. Zhang, L. Liu, and J. Li, "Treatment effect estimation with disentangled latent factors," in *AAAI Conference on Artificial Intelligence*, 2020.

[3] R. Xiong, A. Koenecke, M. Powell, Z. Shen, J. T. Vogelstein, and S. Athey, "Federated causal inference in heterogeneous observational data," *Statistics in Medicine*, vol. 42, no. 24, pp. 4418–4439, 2023.

[4] T. V. Vo, Y. Lee, T. N. Hoang, and T.-Y. Leong, "Bayesian federated estimation of causal effects from observational data," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, vol. 180 of *Proceedings of Machine Learning Research*, pp. 2024–2034, PMLR, 01–05 Aug 2022.

[5] T. V. Vo, A. Bhattacharyya, Y. Lee, and T.-Y. Leong, "An adaptive kernel approach to federated learning of heterogeneous causal effects," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24459–24473, 2022.

[6] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[7] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[8] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[9] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, pp. 217–240, 3 2011.

[10] Y. Xie, J. E. Brand, and B. Jann, "Estimating heterogeneous treatment effects with observational data," *Sociological Methodology*, vol. 42, no. 1, pp. 314–347, 2012. PMID: 23482633.

[11] E. H. Kennedy, "Towards optimal doubly robust estimation of heterogeneous causal effects," *arXiv preprint arXiv:2004.14497*, 2020.

[12] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[13] P. R. Hahn, J. S. Murray, and C. M. Carvalho, "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)," *Bayesian Analysis*, vol. 15, pp. 965–1056, 2020.

[14] F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag, "Generalization bounds and representation learning for estimation of potential outcomes and causal effects," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 7489–7538, 2022.

[15] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International conference on machine learning*, pp. 3076–3085, PMLR, 2017.

[16] S. Li and Y. Fu, "Matching on balanced nonlinear representations for treatment effects estimation," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[18] M. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics, Springer New York, 2011.

[19] M. A. Hernán and J. M. Robins, *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[20] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[21] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American Statistical Association*, vol. 112, pp. 859–877, apr 2017.