Computers in Biology and Medicine 178 (2024) 108779

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Propensity Weighted federated learning for treatment effect estimation in distributed imbalanced environments

Alejandro Almodóvar *, Juan Parras, Santiago Zazo

Information Processing and Telecommunication Center, ETSI de Telecomunicación, Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Keywords:
Causal inference
Counterfactual prediction
Treatment effects
Federated learning
Propensity score

ABSTRACT

Estimating treatment effects from observational data in medicine using causal inference is a very relevant task due to the abundance of observational data and the ethical and cost implications of conducting randomized experiments or experimental interventions. However, how could we estimate the effect of a treatment in a hospital that has very restricted access to treatment? In this paper, we want to address the problem of distributed causal inference, where hospitals not only have different distributions of patients, but also different treatment assignment criteria. Furthermore, it is necessary to take into account that due to privacy restrictions, personal patient data cannot be shared between hospitals. To address this problem, we propose an adaptation of the federated learning algorithm `FederatedAveraging` to one of the most advanced models for the prediction of treatment effects based on neural networks, TEDVAE. Our algorithm adaptation takes into account the shift in the treatment distribution between hospitals and is therefore called *Propensity Weighted* `FederatedAveraging` (PW `FedAvg`). As the distributions of the assignment of treatments become more unbalanced between the nodes, the estimation of causal effects becomes more challenging. The experiments show that PW `FedAvg` manages to reduce errors in the estimation of individual causal effects when imbalances are large, compared to *Vanilla* `FedAvg` and other federated learning-based causal inference algorithms based on the application of federated learning to linear parametric models, Gaussian Processes and Random Fourier Features.

1. Introduction

Causal inference is the task of estimating the effect of a variable (treatment) on a target variable with observational data, which is challenging due to the misspecification of the causal graph in real data [1, 2]. In particular, the variables that causally determine the treatment variable can bias the estimation if not considered properly [3].

This definition of causal inference is particularly useful to understand why it is so important in healthcare: it allows one to evaluate a drug without conducting random control trials (RCTs), which are usually too expensive and unethical; that is, observational data do not modify the treatment assignment criteria of physicians. These criteria depend on the characteristics of the patient (covariates and other unobserved features).

We are interested in a particular case in which treatment assignment criteria can also depend on external factors in distributed environments. That is, settings in which there are several hospitals which act as data processing information nodes, and the selection treatment criteria is different in each node. More specifically, we are interested in the problem of prediction of causal effects of a binary treatment, without having a causal graph available when the data are tabular, with numerical covariates and outcome.

Suppose that we are trying to evaluate the efficiency of a new medication in several hospitals, but some of these hospitals have restrictions in that treatment supply due to the scarcity of that drug (for example, imagine undeveloped countries with very limited resources). In that case, doctors will prescribe this drug to a very small number of patients. However, in hospitals that do not have restrictions on drug availability (in developed countries), the criterion for treatment assignment is different, and the number of treated patients will be higher. This imbalance between treated and control patients is a distribution shift [4] that we call *propensity score shift*.

This problem could be solved by combining the data of all hospitals, taking into account the hospital to each patient, and using causal inference methods. However, for medical purposes, this process is usually not possible, since privacy restrictions do not allow us to share explicit data about patients.

After studying current local causal inference techniques and approaches for distributed privacy-constrained learning, in this paper, we

* Corresponding author.
E-mail addresses: alejandro.almodovar@upm.es (A. Almodóvar), j.parras@upm.es (J. Parras), santiago.zazo@upm.es (S. Zazo).

https://doi.org/10.1016/j.compbiomed.2024.108779
Received 17 January 2024; Received in revised form 16 May 2024; Accepted 15 June 2024
Available online 28 June 2024

propose a framework that combines federated learning with variational disentangled models to address this challenge. Specifically, we adopt Federated Averaging (FedAvg) [5], and use TEDVAE (Treatment Effect with Disentangled Variational Autoencoder) [6] a model of causal inference that achieves good performance in the estimation of treatment effects due to the partial discovery of the causal graph due to the disentanglement of the latent space of a variational autoencoder, for the estimation of the treatment effect. However, as we explain in Section 3.1, some biases can arise in the estimation of causal effects due to the heterogeneity of the treatment assignment in different nodes. To mitigate the biases of the imbalanced treatment assignment criteria, an adaptation of FedAvg has been developed, which is the focus of this article and is called *Propensity Weighted* FedAvg. This adaptation achieves better performance than the standard application of FedAvg and other federated learning-based causal inference algorithms based on the application of federated learning on linear parametric models, Gaussian Processes and Random Fourier Features, as we show in Section 4 by conducting experiments in semi-synthetic causal inference datasets in which we modify the imbalance in treatment assignment distribution between nodes.

Associated with privacy, FedAvg aligns with privacy regulations in the sense that it maintains user data at their local nodes, without sharing individual-level data [7]. Therefore, the privacy restrictions taken into account in this paper imply that individual patient feature data are not shared. However, its preservation of privacy is not absolute; there is a potential for indirect data leakage through shared model updates. This limitation requires additional actions such as differential privacy [8] or encryption [9] to strengthen its compliance with stringent privacy laws and standards, which is part of the future work of this paper.

1.1. Related work

In the realm of classical causal inference, one of the most common strategies to address the prediction of binary treatment effects is the use of propensity score to mitigate observational biases. In [10], the reader can find a collection of propensity score methods that have formed one of the bases of the causal inference field. Among these, we can find: propensity score weighting, which gives more importance to the less represented group (between treated and control) to compute causal effects; propensity score matching [11], which finds similarities between individuals in the covariate space and computes the average causal effect by subtracting the outcome of a treated patient and the outcome of the more similar control patient, and propensity score stratification [12], which makes subgroups of patients with similar propensity scores and computes causal effects within each group. However, all these methods are designed to compute Average Treatment Effects (ATEs); [13] proposes an extension to estimate heterogeneous treatment effects. More recent advances include techniques such as causal forests [14–16], which are an extension of the random forest algorithm, designed specifically for estimating heterogeneous treatment effects, utilizing a large number of decision trees to model the variance in treatment effects across different subpopulations; doubly-robust estimation [17,18] which ensures consistency in ATE estimation when the potential outcome regression or propensity score estimation is consistent; targeted learning and regularization [19–21], which performs an iterative process in which potential outcomes are predicted in one step and a regression is performed on clever variables calculated from an estimation of the propensity score in other step and when the contributions of the clever covariates tends to zero, the model is asymptotically doubly-robust.; Bayesian Additive Regression Trees (BART) [22,23], using Bayesian non-parametrics for flexible modeling of outcomes, particularly effective in estimating complex causal effects; model agnostic algorithms known as meta-learners [24], which are algorithms to compute individual causal effects and can be adapted to any regression model; balancing representations [25], which leverages nonlinear representation learning, and neural network-based approaches such as Treatment-Agnostic Representation Network (TARNet) (used as module of TEDVAE) [26,27] SITE [28], which uses neural network representation learning to balance distributions and [29], which merges the ideas of multi-head learning of TARNet with targeted learning.

Generative models have also gained attention. CEVAE [30] proposes the use of a variational autoencoder (VAE) to infer substitute confounders and latent information from the covariates; TEDVAE [6] aims to disentangle the covariates into several latent factors that affect only treatment, only outcome, or both (confounders) and Intact-VAE [31] proposes an adaptation of identifiable VAE [32] for causal inference in the presence of hidden confounders. Generative Adversarial Networks (GANs) have also been used to infer the distribution of counterfactual outcomes [33,34].

On the other hand, federated learning [5] has been one of the most recognized approaches to address decentralized privacy-constrained training, which is especially interesting in healthcare [35]. In line with our purposes, [36] study the convergence behavior of federated learning in non-IID settings, providing insight into how the algorithm performs when data are unevenly distributed across nodes, and [37] address robustness, exploring techniques to enhance the reliability of federated learning under non-IID conditions.

In the federated causal inference domain, various techniques have been explored, such as parametric linear models [38], which have the disadvantage of not being able to model complex relationships between variables and individual heterogeneous effects; Collaborative Linked Analysis [39], which shares summary statistics to compute average effects in a communication efficient manner, but does not compute individual effects, or aggregation techniques [40] which consist of weighting the estimands by a density ratio to address heterogeneity across nodes, but only achieves the estimation of average treatment effects. In [41], an adaptation of Gaussian processes (GP) is studied to estimate individual causal effects and the application of federated learning; however, this approach has scalability problems and does not account for heterogeneity between nodes. Lastly, [42] proposes the application of Random Fourier Features (RFF), which employs adaptive kernel functions to estimate individual effects under unobserved confounding and the presence of proxy variables, although the hidden confounding is outside of the scope of this paper. To our knowledge, only in [42], dissimilarities in the data distribution of the nodes are considered for estimating individual treatment effects.

During our experiments, we will compare with FedCI and CausalRFF because they are methods designed to estimate individual effects, as is our adaptation of TEDVAE. In addition, we will include the work of [38] because it includes in his article a detailed and clear analytical work and is indisputably one of the starting points of federated causal inference.

1.2. Contributions

In our work, we build on these advances and tailor our approach to the specific challenges of estimating treatment effects in privacy-constrained distributed settings, where treatment assignment criteria are different in distinct nodes.

Specifically, the contribution of this paper is threefold:

- We expose the idea of applying federated averaging to an advanced causal inference models (TEDVAE) and analyze its limitations in cases where the distribution of treatments is unbalanced. Note that this case is the study of *Vanilla* FedAvg.
- We develop a modification of FedAvg, called *Propensity Weighted* FedAvg, which accounts for the propensity score shift and weights in a specific manner the outcome regressors. We show that this method outperforms the standard implementation of FedAvg and other state-of-the-art methods.

- We propose a method to evaluate the performance of the different distributed methods in imbalanced environments. The method consists in gradually increasing the imbalance in treated/control patients in each node and evaluating the causal estimation (using PEHE, introduced in §4) in each node separately, accounting for the performance degradation at each level of imbalance and comparing this performance with the centralized model in which one model is trained with all available data and with isolated models, where the nodes do not share any information and only train their models with their local data, as shown in Fig. 7.

Finally, it should be noted that a conference version of our algorithm can be found in [43], where the `FedAvg` adaptation is proposed and preliminary results are given. In this publication, an extended mathematical justification of the algorithm and more detailed results are given, more scenarios where the imbalances are highlighted, and the algorithm is tested on more benchmarking datasets.

2. Problem definition

2.1. Local causal inference

Causal inference is a very general concept, generally referring to the task of solving any type of causal query, from the perspective of intervention, observation, or control [44]. In our case, we focus on the task of estimating the effect of treatment on observational data, since making interventions is usually very costly and even unethical in medicine.

Consider the data set $D = \{\mathbf{X}_i, T_i, Y_i\}_{i=1}^{N}$, where the subindex $i \in \mathcal{M} = \{1, \dots, N\}$ represents the indices of individuals, and $N = |\mathcal{M}|$ is the total number of data points in the data set. Assume that the samples are i.i.d. observations: $D \overset{iid}{\sim} \mathbb{P}$. In this notation, $\mathbf{X}_i \in \mathbb{R}^{D_x}$ is a vector of covariates (D_x is the number 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 Treatment Effect of T_i on Y_i ($ITE \equiv \tau_i$), following the Neyman–Rubin potential outcome framework [45], as: $\tau_i \equiv Y_i(T_i = 1) - Y_i(T_i = 0)$.

Conventional causal inference methods [11–15,17,19,22,23,25–28,45–47] estimate individual conditioning (ITE) and/or average treatment effect (ATE) on covariates, assuming that the data generation process meets the standard causal inference assumptions: (1) unconfoundedness, (2) positivity, (3) consistency, and (4) *no interference*; following the backdoor criterion [48]:

$$\begin{aligned} \hat{\tau}(\mathbf{x}_i) &= \mathbb{E}[Y|T = 1, \mathbf{X} = \mathbf{x}_i] - \mathbb{E}[Y|T = 0, \mathbf{X} = \mathbf{x}_i] \\ \hat{ATE} &= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y|T = 1, \mathbf{X} = \mathbf{x}_i] - \mathbb{E}[Y|T = 0, \mathbf{X} = \mathbf{x}_i]] \end{aligned} \tag{1}$$

The backdoor criterion consists in conditioning on the variables that block all backdoor paths between the treatment and the target variable in the causal graph. Since the causal graph is usually unknown in real data, conditioning on all covariates can lead to an error by opening backdoor paths or increasing the variance of the estimator.

The model selected in this paper for local causal inference, TEDVAE, is shown in Fig. 1. The first stage of this model consists of three encoders (E_T, E_C, E_Y) with the objective of making a partial discovery of the causal graph by disentangling the latent space, isolating the risk variables (variables that affect only the outcome: $\mathbf{z}_y$), the confounders (variables that affect both the assignment of treatment and the outcome: $\mathbf{z}_c$) and the instrumental variables (variables that only affect the assignment of treatment: $\mathbf{z}_t$) [49].

Then, from the latent space, three tasks have to be solved: (1) prediction of treatment assignment through a classificator C_T, (2) reconstruction of covariates through a decoder D, and (3) prediction of potential outcomes through a well-known causal inference model called TARNet (Treatment-Agnostic Representation Network) [26]. TARNet consists of several shared layers (FC from fully connected), which are updated for all the datapoints in the training process, and two heads, one for each potential outcome, which are only updated for control/treated individuals, respectively: Reg_{Y0} is a regressor used to predict the outcome without treatment, and Reg_{Y1} is used only to predict the outcome in the treated.

The three components of the latent space must be informative to reconstruct the covariates, but only instrumental variables and confounders are used to predict the assignment of treatment in the classifier C_T, and only confounders and risk variables are used to predict the outcome in TARNet. This is how disentanglement is encouraged.

By conditioning only on confounders and risk variables (Fig. 2), instead of conditioning on all covariates, both the bias and the variance of the estimation of treatment effects are reduced [50,51].

To achieve this, in the first place, TEDVAE proposes a modification of the standard Evidence Lower Bound (ELBO) [52], which is composed of an error of covariate reconstruction and three Kullback–Leibler (D_{KL}) divergence terms, where the priors selected are isotropic Gaussians ($p_\theta(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for $\mathbf{z}_t, \mathbf{z}_c, \mathbf{z}_y$), since D_{KL} has an analytic expression.

$$\begin{aligned} l_{\text{ELBO}}(\mathbf{x}, y, t; \Theta) = \mathbb{E}_{q_{\phi_C} q_{\phi_T} q_{\phi_Y}} &\left[\log p_\theta\left(\mathbf{x} \mid \mathbf{z}_t, \mathbf{z}_c, \mathbf{z}_y\right)\right] \\ &- D_{KL}\left(q_{\phi_T}\left(\mathbf{z}_t \mid \mathbf{x}\right) \parallel p_{\theta_t}\left(\mathbf{z}_t\right)\right) \\ &- D_{KL}\left(q_{\phi_C}\left(\mathbf{z}_c \mid \mathbf{x}\right) \parallel p_{\theta_c}\left(\mathbf{z}_c\right)\right) \\ &- D_{KL}\left(q_{\phi_Y}\left(\mathbf{z}_y \mid \mathbf{x}\right) \parallel p_{\theta_y}\left(\mathbf{z}_y\right)\right) \end{aligned} \tag{2}$$

In addition, the objective function of TEDVAE (L_{TEDVAE}) includes the loss of prediction of treatment and the potential outcomes, respectively.

$$\begin{aligned} L_{\text{TEDVAE}}(\Omega; D) = & l_{\text{ELBO}}(\mathbf{x}_i, y_i, t_i; \Theta) \\ &+ \alpha_t \mathbb{E}_{q_{\phi_T} q_{\phi_C}}\left[\log p_{\varphi_t}\left(t_i \mid \mathbf{z}_{t,i}, \mathbf{z}_{c,i}\right)\right] \\ &+ \alpha_y \mathbb{E}_{q_{\phi_Y} q_{\phi_C}}\left[\log p_{\varphi_y}\left(y_i \mid t_i, \mathbf{z}_{c,i}, \mathbf{z}_{y,i}\right)\right] \end{aligned} \tag{3}$$

The terms $\alpha_t, \alpha_y \in \mathbb{R}^+$ are hyperparameters, that have to be adapted to encourage disentanglement. Note that the regressors of the potential outcomes are only updated for treated and control patients, respectively, since we assume that the potential outcome follows the next Gaussian distribution:

$$p_{\varphi_y}(y_i | t_i, \mathbf{z}_{c,i}, \mathbf{z}_{y,i}) = \frac{1}{\hat{\sigma}_i \sqrt{2\pi}} e^{-(x-\hat{\mu}_i)^2 / 2\hat{\sigma}_i^2} \tag{4}$$

where $(\hat{\mu}_i, \hat{\sigma}_i) = t_i \cdot f_{\varphi_{Y1}}(f_{\varphi_y}(\mathbf{z}_{c,i}, \mathbf{z}_{y,i})) + (1 - t_i) \cdot f_{\varphi_{Y0}}(f_{\varphi_y}(\mathbf{z}_{c,i}, \mathbf{z}_{y,i}))$.

2.2. Distributed causal inference

Suppose that we have K information processing nodes, each $k \in \{1, \dots, K\}$ node with a data set $D^{(k)} = \{\mathbf{X}_i^{(k)}, T_i^{(k)}, Y_i^{(k)}\}_{i \in \mathcal{M}^{(k)} \subset \mathcal{M}}$, where $\mathcal{M}^{(k)}$ is the set of patient indices, $N^{(k)} = |\mathcal{M}^{(k)}|$ is the number of samples from node k and $N = \sum_{k=1}^{K} N^{(k)}$, where N is the total number of patients. There are no repeated patients: $\mathcal{M}^{(k)} \cap \mathcal{M}^{(j)} = \varnothing$. Furthermore, $\mathbf{X}_i^{(k)} \in \mathbb{R}^{D_{xk}}$, 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. The union of treated and control patients represents all patients in that node: $\mathcal{C}^{(k)} \cup \mathcal{T}^{(k)} = \mathcal{M}^{(k)}$.

The objective of federated learning applied to causal inference is to improve the performance of the prediction of causal effects of a treatment, having the data distributed in several information processing nodes, where particular information about individuals cannot be shared due to privacy constraints. Therefore, the model of node k does not have access to any datapoint of node j when $k \neq j$.

Data collected at the different nodes can be distributed non-identically ($D^{(j)} \overset{iid}{\sim} \mathbb{P}^{(j)}$, $D^{(k)} \overset{iid}{\sim} \mathbb{P}^{(k)}$, $\mathbb{P}^{(j)} \neq \mathbb{P}^{(k)}$, with $j, k \in \{1, \dots, K\}$, $j \neq k$). Let us define three conditions related to the distribution of the variables, to study in distributed causal inference [38]:

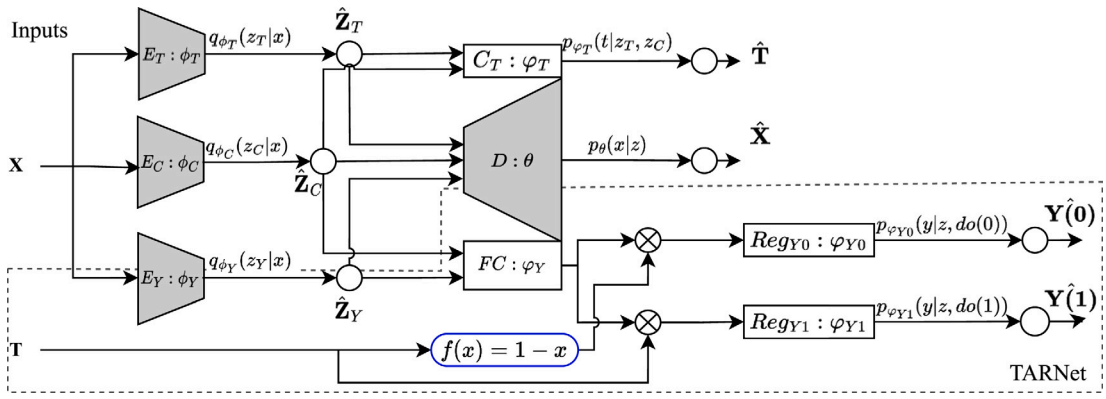


Fig. 1. TEDVAE local model. ∘ represents sampling and ⊗ the product. Gray boxes belong to VAE (with parameters Θ). The rounded blue box with $f(x)$ represents the factor to train each head of TARNet separately: Reg_{Y0} only is updated for individuals with $T = 0$ and Reg_{Y1} only is updated for individuals with $T = 1$. Since the whole model can be considered as a neural network as well, all the boxes will be referenced in this text as modules of the whole network. $M : \vartheta$ express the name of the module (M) and its parameters (ϑ). The output of any module M with parameters ϑ with input $\mathbf{x}_i$ is expressed as $f_\vartheta(\mathbf{x}_i)$.

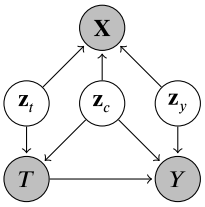


Fig. 2. TEDAVE causal graph.

Condition 1. The set of covariates is the same in all nodes: $D_{xj} = D_{xk}$, $\forall j \neq k$.

Condition 2. The covariate distribution is stable between nodes: $p^{(j)}(\mathbf{X}) = p^{(k)}(\mathbf{X})$, $\forall j \neq k$.

Condition 3. The propensity score is stable between nodes: $p^{(j)}(T|\mathbf{X}) = p^{(k)}(T|\mathbf{X})$, $\forall j \neq k$.

2.3. Conditions of our problem

In this text, we assume that Condition 1 holds, but not necessarily Conditions 2 and 3. We want to focus on a scenario in which some underdeveloped countries do not have access to some drugs. This strategy can help 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.

However, 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 (see Fig. 3).

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 assumption of positivity in each isolated node even when the assumption is met for the entire dataset D.

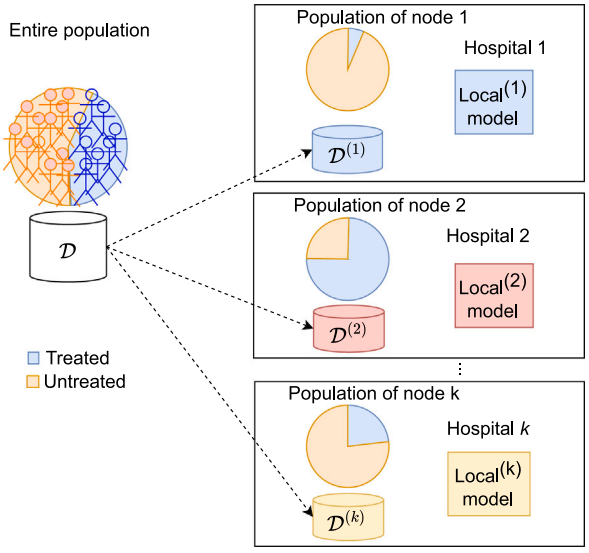


Fig. 3. Schema of treatment imbalance in nodes. Our method is designed to address imbalances in the assignment of treatment between nodes. We assume that the entire population meets the standard assumptions of causal inference and that the proportion of treated and control patients is balanced. However, by dividing the population into different nodes, these conditions are violated.

3. Method of federated learning for causal inference

Definitions. Let us define some terms for the explanation of this section: the parameters of neural networks are expressed in Greek letters, as Fig. 1 shows. Each box in Fig. 1 is a neural network. The letter $\Omega^{(k)}$ refers to the set of all the parameters of the neural network model (TEDVAE) of node k and $\Omega = \{\Omega^{(1)}, \ldots, \Omega^{(k)}\}$ refers to the set of parameters of the models of all nodes. Let us define the parameters of the VAE as $\Theta^{(k)} = \{\phi_T^{(k)}, \phi_C^{(k)}, \phi_Y^{(k)}, \theta^{(k)}\}$, and the set of all parameters in node k is $\Omega^{(k)} = \{\Theta^{(k)}, \varphi_T^{(k)}, \varphi_Y^{(k)}, \varphi_{Y1}^{(k)}, \varphi_{Y0}^{(k)}\}$.

We also include the concept of central *server*, which is a node that coordinates the averaging process. The superscript 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: $N_T^S = \sum_k N_T^{(k)}$ and $N_C^S = \sum_k N_C^{(k)}$.

Methodology. We propose applying `FedAvg` with a star topology as presented in Fig. 4, to our local causal inference model, TEDVAE, to improve performance in the prediction of treatment effects.

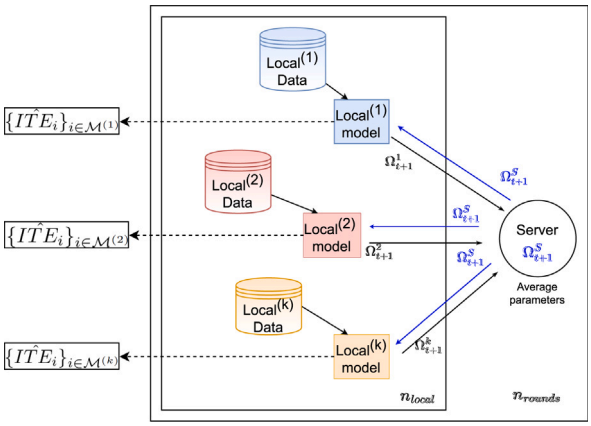


Fig. 4. FedAvg schema.

The prediction problem in federated learning has a global objective function to minimize:

$$\mathcal{L}(\Omega; \mathcal{D}) \equiv \sum_{k=1}^{K} \frac{N^{(k)}}{N} L^{(k)}_{\text{TEDVAE}}(\Omega^{(k)}; \mathcal{D}^{(k)}) \tag{5}$$

where $L^{(k)}_{\text{TEDVAE}}$ is the objective function of the local nodes.

To minimize the global function, FedAvg computes several iterations (n_{rounds} iterations) composed of two stages: (1) local models train several epochs locally (n_{local} epochs), (2) local nodes send their parameters to a server, which makes an average and re-send them to the nodes.

$$\Omega^{S}_{t+1} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \Omega^{(k)}_{t+1} \tag{6}$$

3.1. Limitation of FedAvg

FedAvg has demonstrated robustness and convergence guarantees in convex problems [5]. When data is IID in the nodes, the weights that reduce the variance and the bias of the training process are proportional to the number of samples of each node, as shown in Eq. (6) [53]. However, the violations of Conditions 2 and 3 make the data non-IID due to *propensity score shift*, so the divergence between centralized weights and averaged weights is greater than in IID setting [54].

Although the performance of standard FedAvg has been satisfactory compared to other Federated methods for Causal Inference, we have found that we can improve training by introducing a specific adaptation to deal with unbalanced treatment assignment criteria.

The intuition is as follows. The averaging process in FedAvg is computed by weighting the parameters of each node by the number of samples in that node. This averaging gives more weight to the parameters of the model with less variance, in order to reduce the variance of the averaged estimator. However, our causal inference model includes two regressors to predict both potential outcomes: $Reg0$ to predict $Y(T=0)$ and $Reg1$ to predict $Y(T=1)$. These regressors are only updated for control and treated patients, respectively. Therefore, its variance is inversely proportional to the number of control and treated patients, rather than the total number of samples at each node. In our setting, the data are non-IID, specifically, there is *propensity score shift*. As the imbalance in the propensity score increases, the difference between training samples in those regressors becomes larger, though the number of samples in each node was the same.

To clarify this premise, let us develop the global function, particularizing the objective function of TEDVAE.

$$\begin{aligned}
\mathcal{L}(\Omega; \mathcal{D}) &\equiv \sum_{k=1}^{K} \frac{N^{(k)}}{N} L^{(k)}_{\text{TEDVAE}}(\Omega^{(k)}; \mathcal{D}^{(k)}) \\
&= \sum_{k=1}^{K} \frac{N^{(k)}}{N} \Bigg\{ l_{\text{ELBO}}\left(\mathbf{x}^{(k)}_i, y^{(k)}_i, t^{(k)}_i; \Theta^{(k)}\right) \\
&\quad + \frac{1}{N^{(k)}} \sum_{i \in \mathcal{M}^{(k)}} \underbrace{\alpha_t \left[\log p^{(k)}_{\varphi_T}\left(t^{(k)}_i \mid \mathbf{z}^{(k)}_{t,i}, \mathbf{z}^{(k)}_{c,i}\right)\right]}_{l_T(\varphi^{(k)}_T, \phi^{(k)}_T, \phi^{(k)}_C; \mathcal{D}^{(k)})} \\
&\quad + \frac{1}{N^{(k)}_T} \sum_{i \in \mathcal{T}^{(k)}} \underbrace{\alpha_y \log p^{(k)}_{\varphi_{Y1}}(y^{(k)}_i | t^{(k)}_i = 1, \mathbf{z}^{(k)}_{c,i}, \mathbf{z}^{(k)}_{y,i})}_{l_{Y1}(\varphi^{(k)}_Y, \varphi^{(k)}_{Y1}, \phi^{(k)}_Y, \phi^{(k)}_C; \mathcal{D}^{(k)})} \\
&\quad + \frac{1}{N^{(k)}_C} \sum_{i \in \mathcal{C}^{(k)}} \underbrace{\alpha_y \log p^{(k)}_{\varphi_{Y0}}(y^{(k)}_i | t^{(k)}_i = 0, \mathbf{z}^{(k)}_{c,i}, \mathbf{z}^{(k)}_{y,i})}_{l_{Y0}(\varphi^{(k)}_Y, \varphi^{(k)}_{Y0}, \phi^{(k)}_Y, \phi^{(k)}_C; \mathcal{D}^{(k)})} \Bigg\}
\end{aligned} \tag{7}$$

Note that we have differentiated four terms:

- l_{ELBO}: loss function of the VAE. Updates all VAE modules: Θ. All individuals in the dataset contribute to this loss.
- l_T: loss function of treatment prediction. Encourage latent spaces $\mathbf{z}_t, \mathbf{z}_c$ to be informative in predicting treatment. This term updates the encoder parameters E_T and E_C and the treatment classifier C_T. All individuals in the dataset contribute to this loss.
- l_{Y0}: loss function of the prediction of the potential outcome for $T=0$. Updates the encoders E_C and E_Y, the shared layers of TARNet (FC) and the regressor $Reg0$. Only **control** individuals contribute to this loss.
- l_{Y1}: loss function of the prediction of the potential outcome for $T=1$. Updates the encoders E_C and E_Y, the shared layers of TARNet (FC) and the regressor $Reg1$. Only **treated** individuals contribute to this loss.

The modules of predictors of potential outcomes $Reg0$ and $Reg1$ are the only modules that are not updated for all samples in the node datasets (see Appendix B for local updates of all parameters). As a consequence, the parameters on the server after the averaging process for these modules are:

$$\begin{aligned}
\varphi^{S}_{Y0_{t+1}} &= \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{Y0_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \left(\varphi^{(k)}_{Y0_t} - \eta \nabla_{\varphi_{Y0}} \frac{1}{N^{(k)}_C} \sum_{i \in \mathcal{C}^{(k)}} l_{Y0}(\cdot)\right) \\
&= \varphi^{S}_{Y0_t} - \eta \sum_{k=1}^{K} \nabla_{\varphi_{Y0}} \sum_{i \in \mathcal{C}^{(k)}} \frac{N^{(k)}}{N \cdot N^{(k)}_C} l_{Y0}(\cdot) \\
\varphi^{S}_{Y1_{t+1}} &= \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{Y1_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \left(\varphi^{(k)}_{Y1_t} - \eta \nabla_{\varphi_{Y1}} \frac{1}{N^{(k)}_T} \sum_{i \in \mathcal{T}^{(k)}} l_{Y1}(\cdot)\right) \\
&= \varphi^{S}_{Y1_t} - \eta \sum_{k=1}^{K} \nabla_{\varphi_{Y1}} \sum_{i \in \mathcal{T}^{(k)}} \frac{N^{(k)}}{N \cdot N^{(k)}_T} l_{Y1}(\cdot)
\end{aligned} \tag{8}$$

where η is the learning rate and $(\cdot)$ represents the arguments of the functions that can be consulted in Eq. (7).

These factors are particularly problematic in our scenario, where the propensity score is not constant at different nodes (Condition 3). Consider the two-node scenario of Fig. 5, where $N^{(1)} = N^{(2)}$. This scenario is intended to represent a real case where node 1 is a hospital in an underdeveloped country, where there is no access to treatment (there is only one patient treated) and node 2 is a hospital where treatment is provided according to unbiased medical criteria.

Due to the imbalance in treatment assignment, we can observe that $N^{(1)}/N^{(1)}_T \gg N^{(2)}/N^{(2)}_T$. This fact implies that the contributions of patients treated at node 1 to the averaged gradients are much higher than the contributions of patients treated at node 2. This may lead to an increase in the variance of the averaged estimator and bias if the

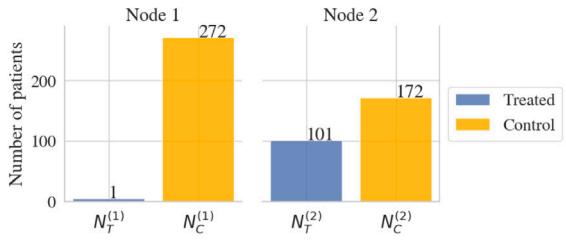


Fig. 5. Two nodes imbalanced scenario. Number of patients in each node. Node 1 is very unbalanced: 1 treated patient, 272 untreated; node 2 is balanced: 101 treated patients, 172 untreated patients. There is the same number of patients in each node, 273 patients in total in each node.

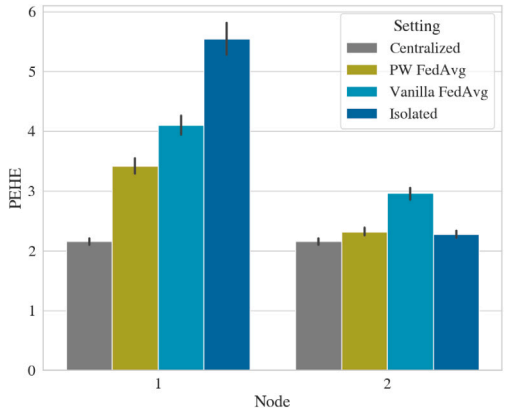


Fig. 6. Error (Mean and 95% CI) in treatment effect estimation (PEHE) in an very imbalanced scenario. **Lower is better**. *Centralized* trains with all the data in a single model; *Vanilla FedAvg* is standard `FedAvg` without address imbalances; *Isolated* trains separated models in each node with their local data, and *PW FedAvg* (Ours) is the best distributed strategy because it considers the imbalance.

samples from node 1 do not represent the entire distribution of treated patients.

In Fig. 6 we can observe the error committed in the estimation of treatment effects (PEHE, see Section 4), comparing the centralized case with the standard FedAvg implementation, using the IHDP database (see Section 4.1) and the scenario of Fig. 5. It can be seen that the error in the estimation, both in node 1 (very unbalanced) and in node 2 (balanced), increases. In contrast, the error committed when training in isolation is much higher in the unbalanced node, whereas it remains low in the balanced node. In view of these observations, we have developed an adaptation of the `FedAvg` algorithm, which weights differently the contributions of the nodes differently in the regressors of the potential outcomes: *Propensity weighted* `FedAvg`. We detail our proposal in the next section (Section 3.2).

3.2. *Propensity weighted* `FederatedAveraging`.

We propose an approach to improve the performance of the standard (from now *Vanilla*) `FedAvg`, which leverages *Propensity score weighting* [10,12,55] and is based on the idea of averaging with a higher weight the modules that have trained with a greater number of samples.

Although theoretical guarantees on the convergence of this algorithm are not available [36] due to the complex nonconvex objective function of TEDVAE, we experimentally prove, on benchmarking datasets for Causal Inference, that our approach achieves better metrics than both *Vanilla* `FedAvg` and other distributed approaches for Causal Inference.

The proposed algorithm, called *Propensity weighted* `FedAvg`, is an adaptation of the averaging process in the central server.

This adaptation consists of two steps on the central server: (1) isolate the parameters of the $Reg0$ and $Reg1$ regressors for each node and (2) perform the averaging process separately in the regressors following Eq. (9).

$$\begin{aligned}\varphi^S_{Y0_{t+1}} &= \sum_{k=1}^{K} \frac{N_C^{(k)}}{N_C^S} \varphi^{(k)}_{Y0_{t+1}} = \varphi^S_{Y0_{t+1}} - \eta \sum_{k=1}^{K} \frac{1}{N_C^S} \varphi^{(k)}_{Y0_t} \\ \varphi^S_{Y1_{t+1}} &= \sum_{k=1}^{K} \frac{N_T^{(k)}}{N_T^S} \varphi^{(k)}_{Y1_{t+1}} = \varphi^S_{Y1_{t+1}} - \eta \sum_{k=1}^{K} \frac{1}{N_T^S} \varphi^{(k)}_{Y1_t}\end{aligned} \tag{9}$$

Note that now the contribution of all control patients is the same for $Reg0$ and the contribution of all treated patients is the same for $Reg1$.

The rest of the modules of the networks are averaged weighted by the number of samples of each node, since all of them are updated for all individuals in a dataset. That is, being ϑ any module of TEDVAE, except the regressors $Reg0$ and $Reg1$:

$$\vartheta^S_{t+1} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \vartheta^{(k)}_{t+1} = \vartheta^S_{t+1} - \eta \sum_{k=1}^{K} \frac{1}{N} \vartheta^{(k)}_t \tag{10}$$

We experimentally demonstrate in Section 4 that our method outperforms the *Vanilla* `FedAvg` when nodes are imbalanced. The implementation of the algorithm and the experiments conducted in Section 4 can be found in https://github.com/aalmodovares/federated_tedvae.

4. Experiments on benchmark datasets

Several experiments have been conducted on benchmark datasets in order to give a complete review of the performance of this algorithm.

The objective, when using a federated learning technique, is to improve the performance of both nodes by training in isolation only with their own data (Fig. 7(a)). On the other hand, the upper bound of performance is found in centralized training (Fig. 7(b)), where all patients are considered to be in a single node and a single model is trained. Therefore, the results of our algorithm must be between these two limits, better results mean being closer to the centralized training.

The comparison will be carried out comparing our implementation of *Propensity Weighted* `FedAvg` on TEDVAE (PW FedAvg) with centralized TEDVAE (Centralized), which trains with all dataset (D); the node-wise isolated training (Isolated), in which each node trains with their data separately, without sharing any information; the *Vanilla* `FedAvg` implementation (Vanilla FedAvg), which does not consider propensity imbalances; the Federated Causal Inference method of [41], based on Gaussian Processes (FedCI); the CausalRFF method of [42] based on Random Fourier Features (CausalRFF) and the federated approach of [38] (Fed MLE).

Since the data used are semi-synthetic, the true values of both potential outcomes are known, and the performance of the model is evaluated using *precision of estimating heterogeneous effects* (PEHE) [22]. PEHE is the mean squared error between predicted and true individual causal effect causal effects, which is expressed as:

$$\text{PEHE} = \mathbb{E}_X[(\hat{\tau}(x) - \tau(x))^2], \tag{11}$$

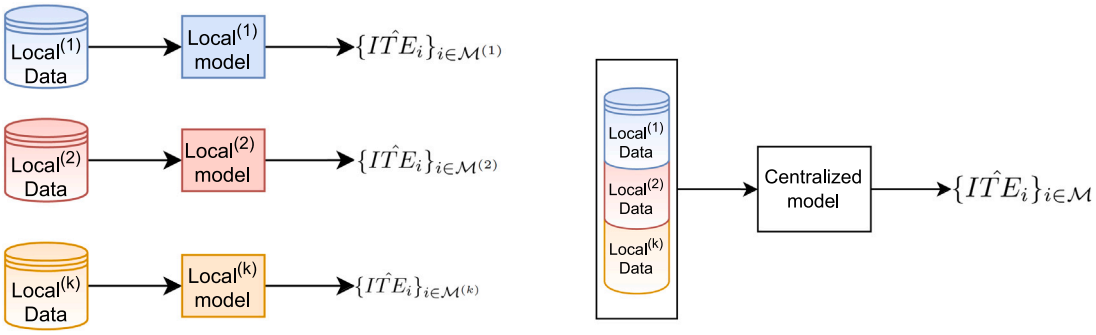
where $\hat{\tau}(x)$ is the estimated treatment effect for subgroup x, and $\tau(x)$ is the true treatment effect for that subgroup. The true causal effect of treatment for patient i is the difference between the two true potential outcomes of that patient:

$$\tau(x_i) = y_{1,i} - y_{0,i} \tag{12}$$

where $y_{t,i}$ is true the potential outcome individual i for each treatment t. Note that we only have access to both potential outcomes because the data are semi-synthetic.

On the other hand, the predicted causal effect of individual i is computed by:

$$\hat{\tau}(x_i) = t_i(y_{1,i} - \hat{y}(0, x_i)) + (1 - t_i)(\hat{y}(1, x_i) - y_{0,i}) \tag{13}$$



(a) Node-wise isolated training/inference (b) Centralized training/inference

Fig. 7. Lower and upper bound of performance. In isolated training (worst case), there is no communication between nodes. In contrast, in the centralized case (better case), a single model trains with all samples.

where t_i is the treatment given to the individual i, and $\hat{y}(t, x_i)$ is the predicted outcome for individual i, treated with treatment t. Note that the predicted Individual Treatment Effect (ITE) is computed using the factual outcome and the counterfactual outcome predicted by our model.

The given results are out-of-sample measurements of PEHE and absolute ATE errors after the training process. The test set used for the inference of potential outcomes is the same for all nodes.

4.1. Datasets

4.1.1. IHDP

Infant Health and Development Program (IHDP) represents a well-established benchmark dataset introduced by [22]. In reality, IHDP is a set of semi-synthetic datasets, where the outcomes are generated as functions of the covariates, described in [56], which can be found in [57]. There are two settings to generate the outcomes: **(1)** setting A, where both potential outcomes are linear combinations of the features and the causal effect is homogeneous and **(2)** setting B, where one of the potential outcomes is an exponential combination of covariates and treatment, so the causal effects are heterogeneous. We have used 100 replications of the dataset. Since the parameters of the functions that generate the outcome are random variables, each replication generates the potential outcomes of different functions, resulting from the sampling process of these parameters. This dataset has 25 covariates ($D_x = 25$), a binary treatment ($T \in \{0, 1\}$), and a continuous simulated outcome ($Y \in \mathbb{R}$). The dataset includes 747 samples: 139 treated individuals and 608 control individuals.

The true causal effect is computed from *noiseless* versions of potential outcomes. Experiments have been performed for both settings of the datasets.

4.1.2. ACIC16

Atlantic Causal Inference Conference 2016 (ACIC16) [58]: The unit contains 4802 samples, 58 covariates ($D_x = 58$), a binary treatment ($T \in \{0, 1\}$) that depends on the covariates ($p(T|\mathbf{X}) \neq p(T)$) and a continuous outcome ($Y \in \mathbb{R}$) simulated with an exponential function. The datasets have been downloaded from [59]. For this study only one of the 77 realizations (each realization has different levels of variability of the effect of treatment, connections between outcomes and assignments, and overlap levels), setting 2, which has a polynomial model for treatment assignment and an exponential model for outcome generation [58]. Ten replications of the data set were used. We remove categorical covariates following [60,61]. The experiment carried out with this dataset is similar to the first one conducted on IHDP datasets: there are two nodes with the same number of patients, and the imbalance increases. The test set has 200 samples with the original distribution of treated patients and is the same for all nodes.

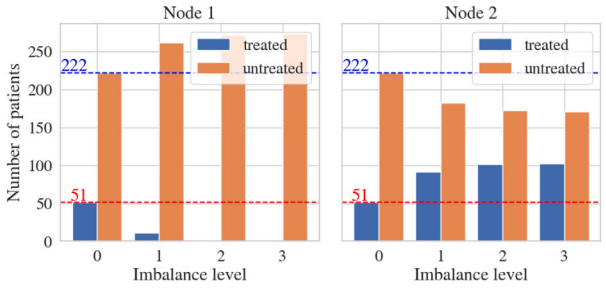


Fig. 8. Imbalances levels IHDP 1. From balanced to completely unbalanced. The same number of patients in each node. **Level 0** (balanced): 51/222 treated/control patients at each node. **Level 1**: 11/262 treated/control patients in node 1, 91/182 treated/control patients in node 2. **Level 2**: 1/272 treated/control patients in node 1, 101/192 treated/control patients in node 2. **Level 3**: 0/273 treated/control patients in node 1, 102/191 treated/control patients in node 2.

4.2. Experiments

The experiments conducted used the previous datasets with different combinations of imbalances.

4.2.1. IHDP 1: Two imbalanced nodes

In this experiment, there are two nodes. Experiments are conducted following the imbalance levels in Fig. 8. "Imbalance 0" is the situation in which the two nodes have the same distribution of treated and untreated patients. The remaining levels increase the imbalance, decreasing the number of treated patients in node 1. In all imbalances, both nodes have the same number of samples. "Imbalance 3" is the extreme case in which there are no treated patients in node 1. A hundred samples have been reserved to test the models after training. The same test set has been used for both nodes. Therefore, the training set has, in total, 546 patients (102 treated, 444 control).

Table 1 and Fig. 9 show that the PEHE achieved by both implementations of `FedAvg` are always between the centralized performance and the isolated performance for large imbalances.

First, observing PEHE in node 1, we note that the isolated errors become very high for larger imbalances. Although *Vanilla* `FedAvg` achieves lower errors than the SOTA methods, the gap with respect to centralized training also increases with greater imbalances. Our method (PW FedAvg) achieves significantly better results than *Vanilla FedAvg*.

On the other hand, in node 2, isolated training performs well, since node 2 is balanced in terms of treatment assignment. The performance of *Vanilla* `FedAvg` worsens when the imbalance grows, due to the action of node 1. Our algorithm manages not to decrease the performance in the balanced node with respect to isolated training. Namely, PW `FedAvg` achieves better results than the other methods in both nodes:

Table 1
$\sqrt{\text{PEHE}}$ out-of-sample results in IHDP setting B: mean(std). **Lower is better**. The mean and standard deviation are from the 100 replications of IHDP. - denotes that the specific method does not work without treated patients at a node. Imbalance 0 is the scenario in which both nodes have the same number of treated patients. In the other *imbalances*, the number of treated and control groups is becoming more and more different at different nodes. a/b represents treated/control patients in each node. Results in bold represent the best results from a one-way t-test with $p < 0.05$. Our method, PW FedAvg, is marked with *. Observe that our method consistently obtains the best results among all distributed methods.

			Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
			Node 1 51/222	Node 2 51/222	Node 1 11/262	Node 2 91/182	Node 1 1/272	Node 2 101/192	Node 1 0/273	Node 2 102/191
Setting B		Centralized				2.10(0.18)				
	TEDVAE	PW FedAvg*	**2.58(0.44)**	**2.63(0.44)**	**3.09(0.52)**	**2.35(0.38)**	**3.42(0.63)**	**2.32(0.33)**	**3.45(0.62)**	**2.29(0.37)**
		Vanilla FedAvg	**2.72(0.49)**	**2.66(0.56)**	3.58(0.80)	2.78(0.55)	4.09(0.81)	2.96(0.50)	3.95(0.89)	2.59(0.41)
		Isolated	**2.69(0.43)**	**2.71(0.46)**	3.75(0.71)	**2.36(0.37)**	5.54(1.36)	**2.28(0.32)**	5.74(0.93)	**2.25(0.35)**
	SOTA	CausalRFF	4.51(1.70)	4.49(1.70)	5.75(1.67)	5.66(1.65)	5.84(1.16)	5.30(1.20)	5.86(1.27)	5.39(1.27)
		FedCI	3.78(1.64)	3.67(1.66)	3.82(1.81)	4.05(1.85)	4.81(2.00)	3.89(1.80)	5.72(1.25)	3.80(1.90)
		Fed MLE	5.64(1.08)	5.64(1.08)	5.66(1.10)	5.66(1.10)	5.55(0.96)	5.55(0.96)	–	-
Setting A		Centralized				0.85(0.35)				
	TEDVAE	PW FedAvg*	**0.84(0.60)**	**0.86(0.60)**	**0.80(0.58)**	**0.86(0.67)**	**0.82(0.58)**	**0.88(0.64)**	**0.84(0.63)**	**0.88(0.65)**
		Vanilla FedAvg	**0.88(0.79)**	**0.86(0.75)**	**0.82(0.66)**	**0.89(0.72)**	1.06(0.70)	1.16(0.74)	1.20(0.67)	1.30(0.73)
		Isolated	**0.84(0.61)**	**0.86(0.57)**	0.97(0.55)	**0.84(0.64)**	2.15(0.88)	**0.87(0.65)**	3.80(1.23)	**0.86(0.65)**
	SOTA	CausalRFF	1.31(2.70)	1.30(2.68)	1.29(2.79)	1.22(2.78)	1.50(2.77)	1.27(2.82)	1.83(2.57)	1.24(2.72)
		FedCI	1.59(2.18)	1.67(2.16)	1.74(1.81)	1.33(1.53)	2.67(2.88)	1.21(1.50)	2.74(1.28)	1.32(1.05)
		Fed MLE	3.42(1.93)	3.42(1.93)	3.36(2.01)	3.36(2.01)	3.30(2.05)	3.30(2.05)	–	–

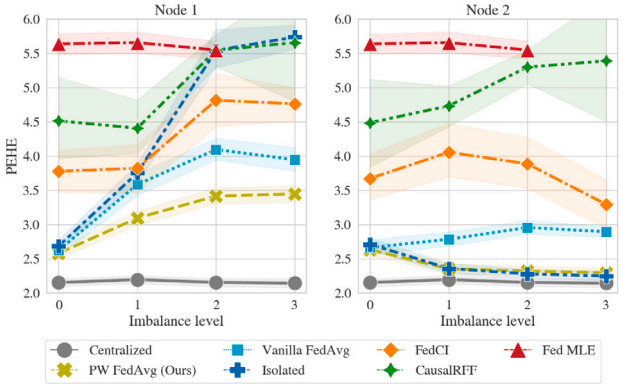


Fig. 9. Mean and 95% confidence interval of PEHE estimation across 100 realizations of **IHDP setting B**. *Propensity weighted* FedAvg (Ours) reduces the error for larger imbalances in both nodes, comparing with *Vanilla* FedAvg.

in node 1 achieves a reduction of the errors due to the imbalance and in node 2 achieves similar performance than the isolated training, which is close to the centralized node, since the patients are balanced in that node.

Both versions of FedAvg (*Vanilla* and our *Propensity Weighted*) show a decrease in errors with respect to the rest of the state-of-the-art (SOTA) methods. This fact is due to the flexibility of the local causal inference method.

Note that there is a slight reduction in PEHE in *Vanilla* FedAvg at the most extreme imbalance (when there are no patients treated at node 1). This is because the $Reg1$ at node 1 is not updated locally at any time. Therefore, it is not overfitted. When averaging between the regressors of node 1 and node 2, the convergence of $Reg1$ is slower, but it only trains on the samples of node 2, which is the only node that has treated patients.

4.2.2. Ablation study on IHDP 1

We perform an ablation study to ensure that sharing the parameters of all network modules is the best option, to reduce the amount of data transmitted in case that sharing fewer parameters offers the same results. Table 2 compares the performance of sharing all TEDVAE submodules (PW FedAvg), with sharing only some of them. For example, "Regressors & Encoder" is the implementation of FedAvg sharing and averaging only the encoder and regressor parameters. In cases where the regressors are shared, *Propensity Weighted* FedAvg has been used.

Analyzing the results, we can see that the share of all modules (All) is between the best performers in both nodes in all combinations of imbalances. When the treatment assignment is balanced (Imbalance 0), all combinations offer similar performance. However, when the imbalance increases, sharing all the modules is the best option, especially in node 1. Sharing only the regressors and the encoder (Regressors & Decoder) is the combination that is closest to the performance of sharing all the modules in node 1. However, Regressors & Decoder sharing produces the worst results in node 2, which is the balanced nodes. For that reason, sharing all the modules is the best option to achieve the lowest errors in both nodes.

4.2.3. IHDP 2: Two small balanced nodes

In this experiment, there are two balanced nodes with a small number of patients (83 patients in each node). This experiment proves the performance of our algorithm where Conditions 2 and 3 of our problem are fulfilled. Note that if these conditions are met, the distribution of patients in both nodes is the same and *Vanilla* FedAvg should work similar to PW FedAvg.

Table 3 shows the PEHE distributions. Both settings of IHDP have been included to observe the difference estimation errors depending on the data generation process. The errors committed in setting B are considerably larger than those committed in setting A, since setting A follows a linear data generation process and the treatment effect is homogeneous. In contrast, the effects of the treatment on setting B are highly non-linear and heterogeneous, making them especially difficult to predict when there is a small amount of data.

On the other hand, comparing the performance of distributed algorithms, in Table 3 it can be seen that the results are similar in *Vanilla* FedAvg as in *Propensity Weighted* FedAvg, since there is no imbalance in the nodes. In Setting A, both FedAvg implementations and isolated training perform very similar to centralized training, since the prediction task is very simple and the data in each node produce good estimates of the causal effect. However, all TEDVAE approaches perform better than SOTA methods.

In Setting B, it can be observed that there is more difference between the centralized training and both *FedAvg* approaches, since the surface of the treatment effects is more complex. Both *Vanilla* and PW FedAvg achieve similar results and perform significantly better than isolated training and SOTA methods.

Then, it is sensible to use PW FedAvg even when the data are balanced between nodes.

Table 2
Ablation study. $\sqrt{\text{PEHE}}$ (Mean(std)) out-of-sample results in IHDP setting B. **Lower is better**. Results in bold represent the best results from a one-way t-test with $p < 0.05$. Sharing all modules is particularly necessary when the imbalances are high between nodes, which is a realistic condition of our problem. Each row represents a different combination of shared parameters. For example, "All" means that all modules have been shared, while "Regressors & Encoder" means that only the parameters of the regressors and the encoders of each node have been shared and averaged.

	Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
Shared modules	Node 1 51/222	Node 2 51/222	Node 1 11/262	Node 2 91/182	Node 1 1/272	Node 2 101/192	Node 1 0/273	Node 2 102/191
All	**2.58(0.44)**	**2.63(0.44)**	**3.09(0.52)**	**2.35(0.38)**	**3.42(0.63)**	**2.32(0.33)**	**3.45(0.62)**	**2.29(0.37)**
Regressors & Encoder	**2.68(0.45)**	**2.64(0.45)**	3.51(0.63)	**2.39(0.37)**	4.43(1.13)	**2.34(0.34)**	4.29(1.00)	**2.30(0.35)**
Regressors & Decoder	2.78(0.46)	2.78(0.43)	3.39(0.57)	2.55(0.36)	**3.64(0.60)**	2.54(0.36)	3.71(0.67)	2.52(0.33)
Regressors	**2.63(0.45)**	**2.65(0.46)**	3.51(0.69)	**2.35(0.32)**	4.30(0.97)	**2.35(0.31)**	4.36(1.17)	**2.29(0.35)**
Encoder & Decoder	**2.74(0.43)**	**2.73(0.46)**	3.78(0.77)	**2.39(0.38)**	5.68(1.57)	**2.33(0.33)**	5.73(0.98)	**2.29(0.37)**
Encoder	**2.75(0.45)**	**2.73(0.48)**	3.82(0.78)	**2.38(0.36)**	5.59(1.39)	**2.33(0.31)**	5.79(1.10)	**2.30(0.33)**
Decoder	**2.71(0.47)**	**2.71(0.45)**	3.77(0.75)	**2.37(0.38)**	5.65(1.43)	**2.31(0.32)**	5.83(0.97)	**2.29(0.34)**

Table 3
Out-of-sample $\sqrt{\text{PEHE}}$ (Mean(std)) results for the original distribution sampled dataset of 83 samples in each node for IHDP setting A and B respectively. **Lower is better.** Imbalances are not specified, since in both nodes there are the same number of treated and control patients, respectively. Therefore, the distribution of patients is balanced between the nodes. With equilibrated nodes, *Propensity Weighted* and *Vanilla* `FedAvg` have similar metrics. Results in bold represent the best results from a one-way t-test with $p < 0.05$. Our method, PW FedAvg, is marked with *.

		Setting A		Setting B	
		Node 1	Node 2	Node 1	Node 2
TEDVAE	Centralized	1.16(0.26)		3.07(0.72)	
	PW FedAvg*	**1.18(0.31)**	**1.20(0.31)**	**3.55(0.86)**	**3.41(0.69)**
	Vanilla FedAvg	**1.15(0.37)**	**1.15(0.29)**	**3.61(0.80)**	**3.50(0.72)**
	Isolated	**1.21(0.41)**	**1.27(0.29)**	4.83(0.81)	4.64(0.65)
SOTA	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)
	Fed MLE	3.56(1.25)	3.53(1.23)	5.58(1.68)	5.64(1.81)

Table 4
Three imbalanced nodes. $\sqrt{\text{PEHE}}$ (Mean(std)) out-of-sample results in IHDP setting B. **Lower is better**. In both node 1 and node 2, the PW FedAvg is the distributed algorithm with the best performance (excluding the centralized method, which acts as a lower bound). A representation of the imbalances can be consulted in Appendix A. Results in bold represent the best results from a one-way t-test with $p < 0.05$. Our method, PW FedAvg, is marked with *. Observe that our method consistently obtains the best results among all distributed methods when there are imbalances between nodes.

		Imbalance 0			Imbalance 1			Imbalance 2			Imbalance 3		
		Node 1 33/148	Node 2 33/148	Node 3 33/148	Node 1 9/172	Node 2 9/172	Node 3 81/108	Node 1 1/180	Node 2 1/180	Node 3 97/92	Node 1 0/181	Node 2 0/181	Node 3 99/90
TEDVAE	Centralized	2.10(0.18)											
	PW FedAvg *	**2.57(0.44)**	**2.63(0.42)**	**2.60(0.43)**	**3.07(0.58)**	**3.12(0.57)**	**2.31(0.40)**	**3.54(0.65)**	**3.51(0.62)**	**2.28(0.38)**	**3.68(0.85)**	**3.68(0.92)**	**2.30(0.40)**
	Vanilla FedAvg	**2.58(0.41)**	**2.62(0.44)**	**2.63(0.46)**	3.61(0.57)	3.58(0.52)	2.85(0.41)	4.26(0.83)	4.26(0.84)	3.13(0.52)	4.26(0.79)	4.32(0.88)	2.83(0.42)
	Isolated	**2.52(0.46)**	**2.57(0.46)**	**2.54(0.49)**	3.99(0.76)	4.04(0.75)	**2.57(0.36)**	6.01(1.74)	5.86(1.94)	**2.53(0.39)**	5.74(0.98)	6.03(0.93)	**2.55(0.37)**
SOTA	CausalRFF	3.80(1.38)	3.98(1.12)	4.27(1.81)	4.31(1.78)	4.26(1.73)	4.19(1.59)	4.40(1.85)	4.48(1.90)	4.28(1.94)	4.51(2.01)	4.74(2.07)	4.37(2.11)
	FedCI	3.60(2.26)	3.29(1.56)	3.57(1.94)	4.12(1.81)	3.96(1.59)	3.73(1.53)	4.49(2.70)	4.55(2.65)	3.68(1.59)	4.54(1.27)	4.57(1.22)	3.68(0.99)
	Fed MLE	5.66(0.98)	5.66(0.98)	5.66(0.98)	5.56(0.91)	5.56(0.91)	5.56(0.91)	5.60(1.03)	5.60(1.03)	5.60(1.03)	–	–	–

4.2.4. IHDP 3: Three imbalanced nodes

This experiment counts with more hospitals (nodes), to expand the multinode evaluation. In the proposed scenario, we have two hospitals where the treated patients decrease progressively, while in the other hospital (good node) the numbers increase. The total number of samples is the same in all nodes. Table 4 shows conclusions similar to those of previous experiments. When the nodes are balanced, all approaches of `FedAvg` have similar results to isolated training. Note that in this case, the distance with centralized training is greater than in the two-node experiment. This is because there, although the nodes are balanced, there are fewer patients in each node. When the imbalance is greater, we can observe that the gap between isolated and centralized training increases in nodes 1 and 2, while it remains almost unchanged in node 3. PW `FedAvg` achieves lower errors than *Vanilla* `FedAvg` and the isolated training in both nodes when imbalances increase.

A similar experiment has been conducted with 5 different hospitals, where there are two good nodes and 3 bad nodes. The results can be consulted in Appendix A.

4.2.5. ACIC16: Two imbalanced nodes

To diversify the study datasets and verify that the results can be extrapolated to other data, the last experiment was conducted with the ACIC16 dataset ACIC16. In this case, there is no good node, since the original distribution of patients has been selected completely balanced: in total, there are 800 treated and 800 control patients. Therefore, in Imbalance 0, both nodes have the same number of treated and control patients. When the imbalance increases, node 1 loses treated patients and gets control patients, and in node 2 the contrary is true. A representation of these imbalances can be found in Appendix A. The number of patients is always the same. The test set is the same for both nodes and contains 200 samples, 100 treated and 100 control patients.

Observing Table 5, we can see that in Imbalance 0, both `FedAvg` approaches and isolated training have similar performance and are close to centralized training. However, in Imbalance 1, the PEHE is greatly increased because both nodes have very few samples of a type. Although the increase in PEHE is also noticeable in PW `FedAvg` (in Imbalance 3 is three times the PEHE in the centralized case), the difference with respect to centralized training is always smaller than in *Vanilla* `FedAvg` and isolated training, and the reduction in PEHE achieved with *Propensity Weighting* increases as the imbalance grows.

Table 5
$\sqrt{\text{PEHE}}$ out of sample results in ACIC2016: mean(std). **Lower is better**. The mean and standard deviation are from the 10 replications of the dataset. The metrics worsen enormously as the imbalances grow, since in this experiment there is no balanced node. Results in bold represent the best results from a one-way t-test with $p < 0.05$. Our method, PW FedAvg, is marked with *. PW FedAvg is the best distributed algorithm in imbalanced scenarios.

		Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
		Node 1 400/400	Node 2 400/400	Node 1 25/775	Node 2 775/25	Node 1 1/799	Node 2 799/1	Node 1 0/800	Node 2 800/0
	Centralized	0.97(0.24)							
TEDVAE	PW FedAvg*	**0.99(0.33)**	**0.98(0.30)**	**1.96(0.94)**	**1.93(0.95)**	**2.91(1.23)**	**2.86(1.18)**	**2.95(1.38)**	**2.89(1.36)**
	Vanilla FedAvg	**1.04(0.29)**	**1.05(0.29)**	2.17(0.49)	2.15(0.55)	3.42(0.89)	3.44(0.99)	4.20(1.27)	4.08(1.24)
	Isolated	**1.01(0.35)**	**0.98(0.37)**	2.25(0.79)	**2.12(0.68)**	4.48(1.35)	4.65(1.40)	5.19(1.30)	5.39(1.38)
SOTA	CausalRFF	3.76(1.23)	3.79(1.26)	4.14(1.92)	4.54(2.45)	3.62(2.44)	4.23(2.21)	5.01(2.02)	4.91(1.79)
	FedCI	3.56(5.92)	2.78(4.73)	3.00(5.07)	4.47(6.49)	5.13(7.30)	4.49(7.47)	6.37(3.34)	5.19(2.78)
	Fed MLE	4.10(0.94)	4.10(0.94)	4.08(1.04)	4.08(1.04)	3.85(0.93)	3.85(0.93)	–	–

More details of the results can be consulted in Appendix A, where measurements of ATE error and more representations of these experiments can be found as well.

5. Conclusion

This study emphasizes the importance of distributed causal inference for estimating the effect of treatment in environments where the distributions of treated/control patients are imbalanced across nodes due to external factors. This problem is particularly interesting in the healthcare domain, which is a privacy-constrained setting in which individual patient data cannot be shared. We addressed the challenge by implementing an algorithm for federated learning in conjunction with a deep model of causal inference based on variational autoencoders. The main contribution of this paper is to propose a framework for Federated Learning applied to one of the most advanced Causal Inference methods and develop an adaptation of `FedAvg` to the propensity score shift problem, called *Propensity Weighted FedAvg*.

We showed empirically, by benchmarking data sets for causal inference, that our adaptation bridges the gap between centralized training and isolated training and the standard implementation of `FedAvg` in the task of predicting causal effects. It also outperforms other state-of-the-art methods for distributed causal inference, since the complexity of the causal inference model allows to model complex non-linear surfaces of treatment effects. The most important drawback of this method is that the amount of data shared is much larger than in the other methods, and the training process must be synchronous.

In future work, there are three research lines. First, advanced privacy preservation techniques must be studied and tested in combination with `FedAvg` must been studied and tested. Second, this algorithm can be tested with other powerful causal inference methods in which the prediction of both potential outcomes is performed in different modules, such as [62,63]. However, the next challenge is to solve the problem of covariate set mismatching. In medical data there is usually heterogeneity of data, and it is common to have different multidomains in each node (images, analysis, demographic and social data...); so it is also common that not all hospitals have access to the same kind of data. We can relate this difficult challenge to other latent models such as [64,65], which take into account this characteristic of the data.

CRediT authorship contribution statement

Alejandro Almodóvar: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Juan Parras:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology. **Santiago Zazo:** Writing – review & editing, Supervision.

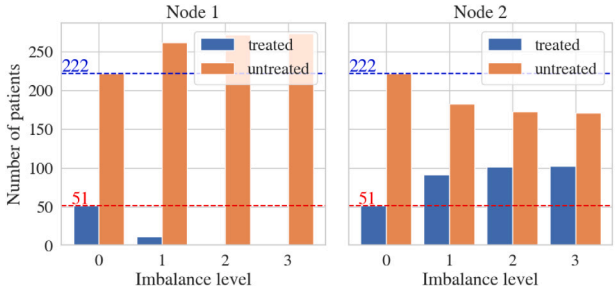


Fig. 10. Increasing imbalances in experiments IHDP 1. From balanced distribution to completely unbalanced. The same number of patients in each node. **Level 0** (balanced): 51/222 treated/control patients in each node. **Level 1**: 11/262 treated/control patients in node 1, 91/182 treated/control patients in node 2. **Level 2**: 1/272 treated/control patients in node 1, 101/192 treated/control patients in node 2. **Level 3**: 0/273 treated/control patients in node 1, 102/191 treated/control patients in node 2.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alejandro Almodovar, Juan Parras, Santago Zazo reports financial support was provided by European Commission. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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.

Appendix A. Imbalances and results representations

In this section, we give a visual representation of the imbalances proposed for the experiments, in order to see more clearly how the imbalances between treated/control patients are created in each node. In addition, visual representations of the results tables are included.

Although in tables the mean and standard deviation of the $\sqrt{\text{PEHE}}$ are presented, in the following figures the mean and the 95% confidence interval computed using bootstrap across the 100 realizations are plotted.

Table 6
ATE error out of sample results in IHDP setting B. **Lower is better**. The results in bold represent the best results from a one-way t-test with $p < 0.1$. Our method, PW FedAvg, is marked with *. For larger imbalances, PW FedAvg is better than Vanilla FedAvg in the balanced node and better than Isolated training in the unbalanced node.

	Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
	Node 1 51/222	Node 2 51/222	Node 1 11/262	Node 2 91/182	Node 1 1/272	Node 2 101/192	Node 1 0/273	Node 2 102/191
Centralized				0.36(0.25)				
PW FedAvg*	**0.48(0.34)**	**0.49(0.36)**	**0.66(0.49)**	**0.38(0.32)**	**0.82(0.65)**	**0.35(0.23)**	**0.84(0.66)**	**0.39(0.31)**
Vanilla FedAvg	**0.56(0.44)**	**0.51(0.39)**	**0.85(0.67)**	0.57(0.47)	1.16(0.92)	0.78(0.55)	1.23(0.69)	0.71(0.31)
TV Iso	**0.59(0.39)**	**0.53(0.41)**	1.02(0.81)	**0.39(0.30)**	2.51(1.73)	**0.40(0.27)**	3.19(1.21)	**0.38(0.32)**
CausalRFF	0.99(0.62)	0.99(0.84)	1.13(0.30)	1.14(0.31)	1.72(1.28)	0.97(0.90)	1.40(1.22)	0.97(0.60)
FedCI	1.02(0.33)	0.99(0.33)	0.96(0.36)	1.01(0.37)	1.16(0.40)	0.98(0.36)	1.17(0.24)	0.85(0.30)
Fed MLE	1.69(0.93)	1.69(0.93)	1.69(0.89)	1.69(0.89)	1.63(0.81)	1.63(0.81)	–	–

Table 7
|ATE error| (Mean(std)) Out-of-sample results in IHDP setting A. **Lower is better**. The results in bold represent the best results from a one-way t-test with $p < 0.1$. Our method, PW FedAvg, is marked with *. For larger imbalances, PW FedAvg is better than Vanilla FedAvg in the balanced node and better than Isolated training in the unbalanced node.

	Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
	Node 1 51/222	Node 2 51/222	Node 1 11/262	Node 2 91/182	Node 1 1/272	Node 2 101/192	Node 1 0/273	Node 2 102/191
Centralized				0.15(0.10)				
PW FedAvg*	**0.13(0.11)**	**0.15(0.14)**	**0.19(0.15)**	**0.16(0.16)**	**0.23(0.16)**	**0.16(0.13)**	**0.25(0.22)**	**0.19(0.14)**
Vanilla FedAvg	**0.18(0.23)**	**0.19(0.20)**	**0.20(0.16)**	**0.23(0.17)**	0.49(0.30)	0.51(0.29)	0.65(0.35)	0.62(0.31)
Isolated	**0.18(0.15)**	**0.20(0.14)**	0.37(0.29)	**0.17(0.14)**	1.54(0.79)	**0.15(0.11)**	3.47(1.06)	**0.18(0.15)**
CausalRFF	0.72(0.86)	0.74(0.82)	0.84(0.97)	0.77(0.91)	1.26(1.01)	0.75(0.97)	1.98(1.00)	0.69(0.88)
FedCI	0.52(0.44)	0.53(0.43)	0.45(0.36)	0.47(0.31)	0.73(0.58)	0.44(0.30)	0.77(0.22)	0.43(0.22)
Fed MLE	1.07(0.80)	1.07(0.80)	1.05(0.89)	1.05(0.89)	0.94(0.86)	0.94(0.86)	–	–

Table 8
Three imbalanced nodes.|ATE error| out of sample results in IHDP setting B. **Lower is better**. The results in bold represent the best results from a one-way t-test with $p < 0.1$. Our method, PW FedAvg, is marked with *. For larger imbalances, PW FedAvg is better than Vanilla FedAvg in the balanced node and better than Isolated training in the unbalanced node.

	Imbalance 0			Imbalance 1			Imbalance 2			Imbalance 3		
	Node 1 33/148	Node 2 33/148	Node 3 33/148	Node 1 9/172	Node 2 9/172	Node 3 81/108	Node 1 1/180	Node 2 1/180	Node 3 97/92	Node 1 0/181	Node 2 0/181	Node 3 99/90
Centralized						0.35(0.26)						
PW FedAvg*	**0.47(0.33)**	**0.59(0.37)**	**0.50(0.39)**	**0.71(0.51)**	**0.75(0.52)**	**0.40(0.32)**	**0.73(0.86)**	**0.72(0.82)**	**0.44(0.34)**	**0.81(0.88)**	**0.80(0.96)**	**0.44(0.29)**
Vanilla FedAvg	**0.47(0.39)**	**0.54(0.43)**	**0.54(0.42)**	**0.75(0.59)**	**0.67(0.45)**	**0.48(0.31)**	1.20(0.98)	1.18(1.03)	0.68(0.54)	1.23(0.85)	1.24(0.92)	**0.44(0.33)**
Isolated	0.72(0.50)	0.72(0.52)	0.64(0.54)	1.29(0.94)	1.20(0.71)	**0.51(0.29)**	2.94(2.10)	2.72(2.29)	**0.53(0.30)**	3.16(1.00)	3.43(1.29)	**0.51(0.33)**
CausalRFF	1.16(0.61)	1.26(0.56)	1.32(0.50)	1.37(0.71)	1.28(0.62)	1.29(0.59)	1.17(0.82)	1.24(0.63)	1.13(0.56)	1.50(0.70)	1.36(0.70)	1.24(0.60)
FedCI	2.61(2.81)	2.37(2.05)	2.25(2.55)	2.48(2.45)	3.20(2.15)	2.25(2.06)	2.71(2.35)	3.56(3.27)	2.19(2.10)	2.63(0.85)	3.35(0.72)	1.88(0.60)
Fed MLE	2.22(0.80)	2.22(0.80)	2.22(0.80)	2.06(0.77)	2.06(0.77)	2.06(0.77)	2.11(0.80)	2.11(0.80)	2.11(0.80)	–	–	–

Table 9
|ATE error| (Mean(std)) out-of-sample results in ACIC2016: mean(std). **Lower is better**. Mean and standard deviation comes from the 10 replications of the dataset. The results in bold represent the best results from a one-way t-test with $p < 0.1$. Our method, PW FedAvg, is marked with *. For larger imbalances, PW FedAvg is the best distributed strategy.

	Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
	Node 1 400/400	Node 2 400/400	Node 1 25/775	Node 2 775/25	Node 1 1/799	Node 2 799/1	Node 1 0/800	Node 2 800/0
Centralized				0.18(0.14)				
PW FedAvg*	**0.35(0.23)**	**0.35(0.22)**	**0.50(0.68)**	**0.65(0.74)**	**0.98(0.95)**	**1.13(1.04)**	**1.10(1.10)**	**1.00(0.93)**
Vanilla FedAvg	**0.34(0.21)**	**0.34(0.21)**	0.90(0.39)	0.93(0.44)	1.55(0.87)	1.65(0.89)	1.76(1.00)	1.65(0.98)
Isolated	**0.29(0.26)**	**0.28(0.23)**	0.92(0.53)	0.88(0.48)	2.06(1.08)	2.10(1.15)	2.55(1.19)	2.62(1.42)
CausalRFF	**0.59(0.44)**	0.67(0.02)	1.45(0.82)	1.65(1.01)	1.32(0.91)	**1.60(0.86)**	1.48(0.98)	**1.41(0.83)**
FedCI	2.11(1.18)	1.96(0.95)	2.20(1.01)	1.99(1.30)	2.43(1.46)	2.40(1.49)	2.67(0.67)	2.44(0.43)
Fed MLE	3.10(0.94)	3.10(0.94)	3.08(1.04)	3.08(1.04)	2.85(0.93)	2.85(0.93)	–	–

A.1. IHDP 1. Two imbalanced nodes

A.1.1. Setting B

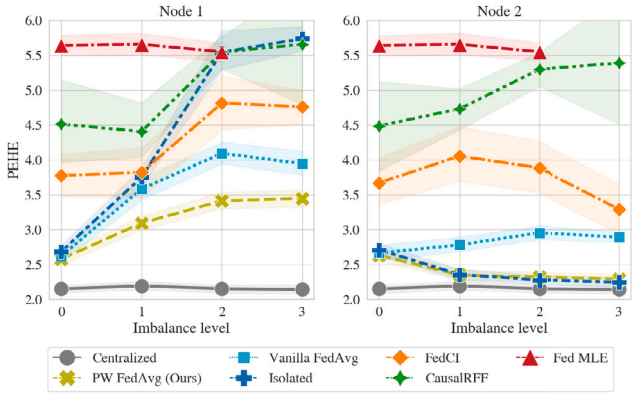
The PEHE results that compare our adaptation of FedAvg (TV FedAvg PA) with the isolated training, Vanilla FedAvg and state-of-the-art methods are shown in Fig. 11(a), which is the same figure that we can observe in Section 4 (see Fig. 10).

In Fig. 11(b) and Table 6, we can observe the mean absolute errors in the ATE estimates. We can observe similar conclusions to the PEHE evaluation: when the nodes are balanced, both FedAvg approaches performs similar and close to the isolated training in both nodes; however, as the imbalance increases, the ATE error of *Vanilla* FedAvg, and specially of isolated training, increases in node 1 and PW FedAvg achieves to reduce these errors. On the other hand, in node 2, PW FedAvg also achieves lower errors than *Vanilla* FedAvg and performs similar training to isolated training, very close to centralized training. All TEDVAE approaches achieve better metrics than SOTA methods.

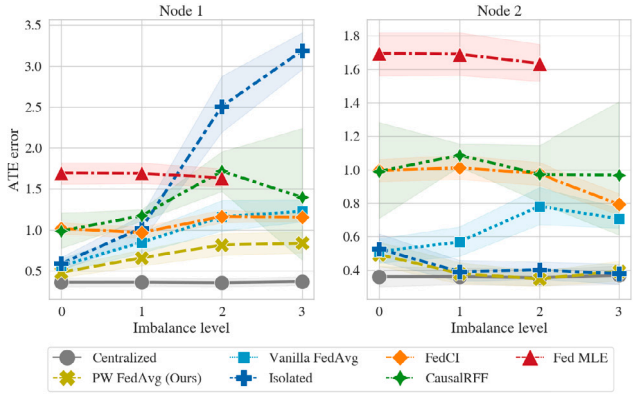
The results of ATE can also be observed in Table 7.

A.1.2. Ablation study in IHDP 1 (setting B)

In addition, in this IHDP experiment with two nodes and imbalanced data, an ablation study has been performed, to unveil which

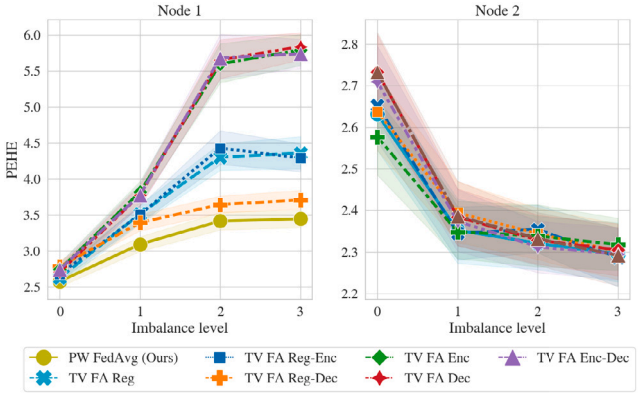


(a) PEHE

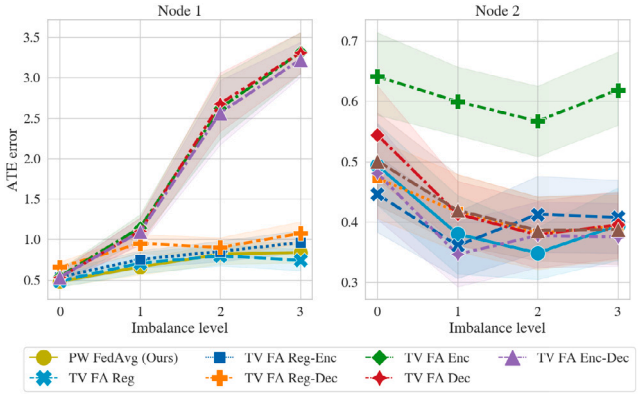


(b) |ATE error|

Fig. 11. Mean and 95% confidence interval of metrics across 100 realizations of IHDP setting B using two nodes. Our algorithm brings the gap between Centralized training and both *Vanilla* `FedAvg` and isolated training in both PEHE and ATE metrics.



(a) PEHE



(b) |ATE error|

Fig. 12. Metrics of Ablation study over 100 replications of IHDP. Sharing all the modules is the best option, especially in the unbalanced node. Sharing only Regressors and Decoder also shows a good performance, but significantly worse that sharing all modules in unbalanced node.

modules of the whole model are necessary to average. The measurements in $\sqrt{\text{PEHE}}$ performance for the ablation study are in Fig. 12(a). The names of each experiment correspond to the shared and averaged modules. For example, in the case of TV FA Reg-Enc, both the encoders and the regressors are shared and averaged in the central server, while the decoder is trained only locally. In cases where regressors weights are shared, the *Propensity Weighted* `FedAvg` has been used. PW FedAvg refers to the case in which the parameters of all modules are shared (All in Table 2). The behavior of these curves has been explained in Section 4. Fig. 12(b) shows the mean absolute error measurements in ATE for each combination of shared parameters, but the differences are not significant between the different configurations.

As we can see in Fig. 12(a), the performance of sharing all modules is better than in the cases where we only share some modules. Particularly, the cases in which we do not do the average in the regressors weights are the worst performers.

A.1.3. Setting A

Now, we include the PEHE and ATE error results for setting A of IHDP, with two nodes and increasing the imbalance. First, the PEHE line plot is presented in Fig. 13(a). As in setting B, setting A shows that the PEHE achieved by PW FedAvg falls between centralized training and *Vanilla* `FedAvg` at both node 1 and node 2 as the imbalances increase. The mean absolute errors in the ATE estimates can be observed in Fig. 13(b). As the imbalance increases, the errors of *Vanilla* `FedAvg` becomes greater in both nodes 1 and 2. Our algorithm achieves lower errors in both nodes and its errors are close to isolated training in node 2 (good node).

The information of ATE errors of Fig. 13(b) can also be consulted in Table 7 for both Setting A and Setting B.

A.2. IHDP 3. Three imbalanced nodes

The proposed imbalances for the three-node scenario can be observed in Fig. 14. There are two *bad nodes* (nodes 1 and 2), where the number of treated patients is becoming very low in larger imbalance levels, while the distribution of treated/control patients is more balanced in node 3.

The results are collected in Fig. 15 and Table 8. Regarding PEHE, we can observe that the incremental jumps of PEHE when the imbalances increase are higher due to the small number of data in each node. The *slope* of the PEHE and ATE error curve in PW FedAvg is greater than in the two-node case. However, it can be seen that it is a significant improvement compared to Vanilla FedAvg, the isolated case, and the state-of-the-art methods at nodes 1 and 2 in both PEHE and ATE. On the other hand, in node 3, the performance of PW FedAvg is similar to the isolated case and significantly better than *Vanilla* `FedAvg` and SOTA methods for larger imbalances. Note, again, that there is a slight reduction in PEHE and ATE errors in the most extreme imbalance because the regressors of treated patients (Reg1) are never updated locally at the bad nodes.

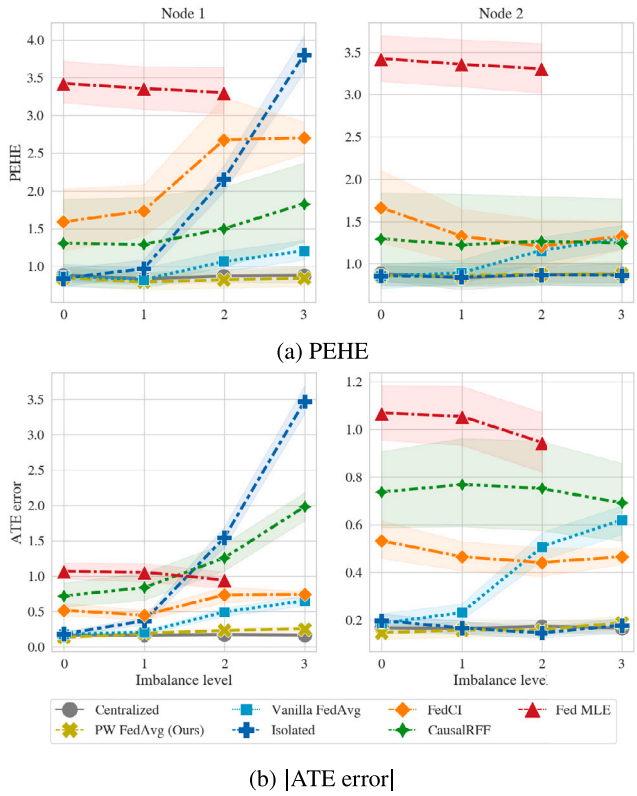


Fig. 13. Metrics (Mean and 95% CI) over 100 replications of IHDP setting A. PW FedAvg remains very close to the centralized case, being the best strategy in unbalanced scenarios.

A.3. ACIC2016. Two imbalanced nodes

The ACIC experiment is based on balanced datasets in terms of the number of patients: in total there are 800 treated patients and 800 control patients. Data are divided into two nodes with the same number of patients. In imbalance 0, both nodes have 400 treated patients and 400 untreated patients. As the imbalances become larger, node 1 loses treated patients and gains untreated patients, while the opposite is true for node 2. Therefore, in that case, there is no clearly good node that has a balanced distribution of treatment allocation at high levels of imbalances (see Fig. 16).

Regarding the analysis of the results, PEHE and the mean absolute error in ATE are presented in Fig. 17. The performance of both the isolated training and `FedAvg` approaches is similar to the centralized case in Imbalance 0 for both nodes, since the treatment distribution is balanced and the number of samples is relatively large. However, as the imbalance grows, the PEHE and ATE errors increase. PW FedAvg achieves better metrics in both PEHE and ATE errors than *Vanilla* `FedAvg`, isolated training and SOTA methods in both nodes for larger imbalances.

Numeric information on ATE errors can also be observed in Table 9.

A.4. IHDP 4. Five nodes IHDP

In this experiment there are five nodes, three of them are becoming worse due to the lack of treated patients, while the other two become more balanced. The performance of *Propensity Weighted* `FedAvg` is better than *Vanilla* `FedAvg` (see Fig. 18). However, we note that the performance is worse than in other experiments, since the number of samples in each node is very low due to the split of the data. A plot of the PEHE and |ATE error| is presented in Fig. 19. It can be observed, that, although the differences are smaller, PW `FedAvg` is still the best distributed alternative for larger imbalances (see Fig. 19 and Table 10).

Appendix B. Full equations of local optimization and averaged parameters

Here are detailed the full equations of the optimization process of TEDVAE.

First, the local update of the parameters of each module of TEDVAE local models is in Eq. (14). Regressors $Reg0$ and $Reg1$ can be observed to be updated only for treated and control patients, respectively. On the other hand, the rest of the modules are updated for all patients.

$$
\begin{aligned}
\varphi^{(k)}_{Y0_{t+1}} &= \varphi^{(k)}_{Y0_t} - \eta\nabla_{\varphi_{Y0}} \frac{1}{N^{(k)}_C} \textstyle\sum_{i\in\mathcal{C}^{(k)}} l_{Y0}(\cdot)\\
\varphi^{(k)}_{Y1_{t+1}} &= \varphi^{(k)}_{Y1_t} - \eta\nabla_{\varphi_{Y1}} \frac{1}{N^{(k)}_T} \textstyle\sum_{i\in\mathcal{T}^{(k)}} l_{Y1}(\cdot)\\
\varphi^{(k)}_{Y_{t+1}} &= \varphi^{(k)}_{Y_t} - \eta\nabla_{\varphi_{Y}} \Big[\frac{1}{N^{(k)}_C} \textstyle\sum_{i\in\mathcal{C}^{(k)}} l_{Y0}(\cdot)\\
&\quad + \frac{1}{N^{(k)}_T} \textstyle\sum_{i\in\mathcal{T}^{(k)}} l_{Y1}(\cdot)\Big]\\
\theta^{(k)}_{t+1} &= \theta^{(k)}_t - \eta\nabla l_{\text{ELBO}}(\cdot)\\
\varphi^{(k)}_{T_{t+1}} &= \varphi^{(k)}_{T_t} - \eta\nabla_{\varphi_T} \frac{1}{N^{(k)}} \textstyle\sum_{i\in\mathcal{M}^{(k)}} l_T(\cdot)\\
\phi^{(k)}_{T_{t+1}} &= \phi^{(k)}_{T_t} - \eta\nabla_{\phi_T} \Big[\frac{1}{N^{(k)}} \textstyle\sum_{i\in\mathcal{M}^{(k)}} l_T(\cdot) + l_{\text{ELBO}}(\cdot)\Big]\\
\phi^{(k)}_{Y_{t+1}} &= \phi^{(k)}_{Y_t} - \eta\nabla_{\phi_Y} \Big[\frac{1}{N^{(k)}_C} \textstyle\sum_{i\in\mathcal{C}^{(k)}} l_{Y0}(\cdot)\\
&\quad + \frac{1}{N^{(k)}_T} \textstyle\sum_{i\in\mathcal{T}^{(k)}} l_{Y1}(\cdot) + l_{\text{ELBO}}(\cdot)\Big]\\
\phi^{(k)}_{C_{t+1}} &= \phi^{(k)}_{C_t} - \eta\nabla_{\phi_C} \Big[\frac{1}{N^{(k)}} \textstyle\sum_{i\in\mathcal{M}^{(k)}} l_T(\cdot)\\
&\quad + \frac{1}{N^{(k)}_C} \textstyle\sum_{i\in\mathcal{C}^{(k)}} l_{Y0}(\cdot) + \frac{1}{N^{(k)}_T} \textstyle\sum_{i\in\mathcal{T}^{(k)}} l_{Y1}(\cdot) + l_{\text{ELBO}}(\cdot)\Big]
\end{aligned}
\tag{14}
$$

Then, the equations resulting from the averaging process, applying *Vanilla* `FedAvg`, can be consulted in Eq. (15).

$$
\begin{aligned}
\varphi^{S}_{Y0_{t+1}} &= \varphi^{S}_{Y0_t} - \eta \textstyle\sum_k \nabla_{\varphi_{Y0}} \sum_{i\in\mathcal{C}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_C} l_{Y0}(\cdot)\\
\varphi^{S}_{Y1_{t+1}} &= \varphi^{S}_{Y1_t} - \eta \textstyle\sum_k \nabla_{\varphi_{Y1}} \sum_{i\in\mathcal{T}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_T} l_{Y1}(\cdot)\\
\varphi^{S}_{Y_{t+1}} &= \varphi^{S}_{Y_t} - \eta \textstyle\sum_k \nabla_{\varphi_{Y}} \Big[\sum_{i\in\mathcal{C}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_C} l_{Y0}(\cdot)\\
&\quad + \textstyle\sum_{i\in\mathcal{T}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_T} l_{Y1}(\cdot)\Big]\\
\varphi^{S}_{T_{t+1}} &= \varphi^{S}_{T_t} - \eta \textstyle\sum_k \nabla_{\varphi_{T}} \sum_{i\in\mathcal{M}^{(k)}} \frac{1}{N} l_T(\cdot)\\
\theta^{S}_{t+1} &= \theta^{S}_t - \eta \textstyle\sum_k \nabla_{\theta} \frac{N^{(k)}}{N} l_{\text{ELBO}}(\cdot)\\
\phi^{S}_{T_{t+1}} &= \phi^{S}_{T_t} - \eta \textstyle\sum_k \nabla_{\phi_T} \Big[\sum_{i\in\mathcal{M}^{(k)}} \frac{1}{N} l_T(\cdot) + \frac{N^{(k)}}{N} l_{\text{ELBO}}(\cdot)\Big]\\
\phi^{S}_{C_{t+1}} &= \phi^{S}_{C_t} - \eta \textstyle\sum_k \nabla_{\phi_C} \Big[\sum_{i\in\mathcal{M}^{(k)}} \frac{1}{N} l_T(\cdot)\\
&\quad + \textstyle\sum_{i\in\sum_k C^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_C} l_{Y0}(\cdot)\\
&\quad + \textstyle\sum_{i\in\mathcal{T}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_T} l_{Y1}(\cdot) + \frac{N^{(k)}}{N} l_{\text{ELBO}}(\cdot)\Big]\\
\phi^{S}_{Y_{t+1}} &= \phi^{S}_{Y_t} - \eta \textstyle\sum_k \nabla_{\phi_Y} \Big[\sum_{i\in\mathcal{C}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_C} l_{Y0}(\cdot)\\
&\quad + \textstyle\sum_{i\in\mathcal{T}^{(k)}} \frac{N^{(k)}}{N\cdot N^{(k)}_T} l_{Y1}(\cdot) + \frac{N^{(k)}}{N} l_{\text{ELBO}}(\cdot)\Big]
\end{aligned}
\tag{15}
$$

It is important to note the contribution of each patient to the averaged gradients. In Reg0 and Reg1 regressors (with parameters phi0 and phi1, respectively), the gradients corresponding to the control patients of node k are multiplied by a factor $N^{(k)}/N^{(k)}_C$, and the same is true for the treated group, multiplied by $N^{(k)}/N^{(k)}_T$. In contrast, we can observe that, in the rest of the modules, patients from all nodes contribute equally to the averaged gradient.

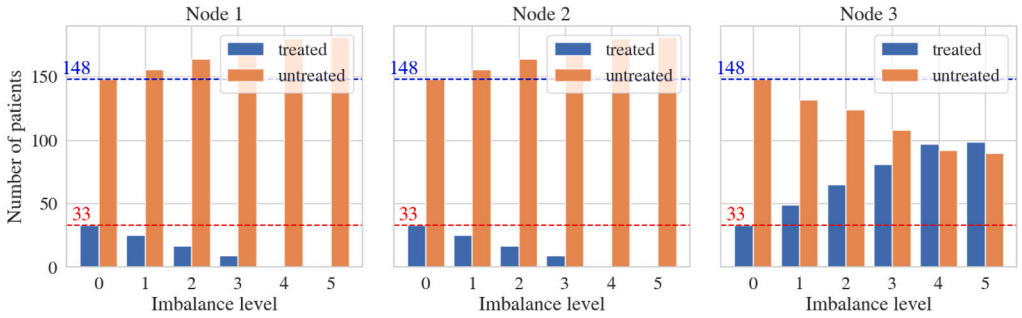
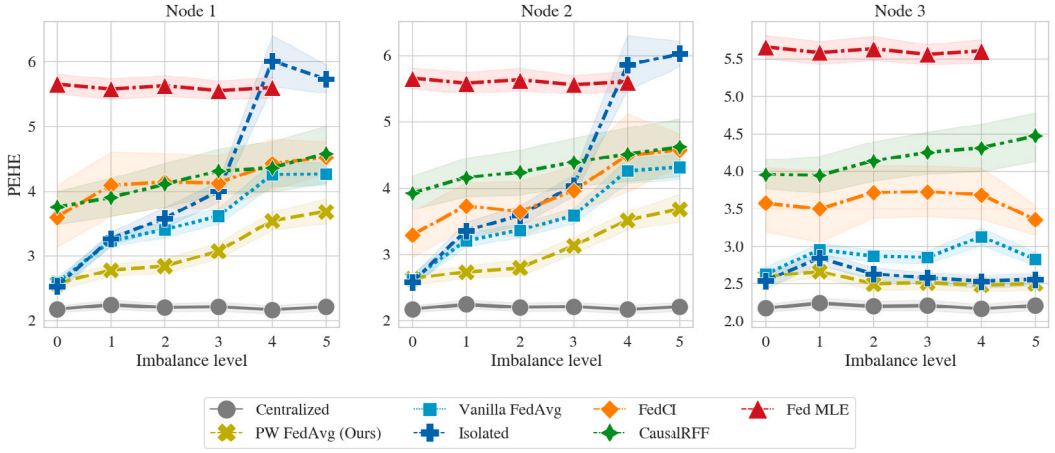
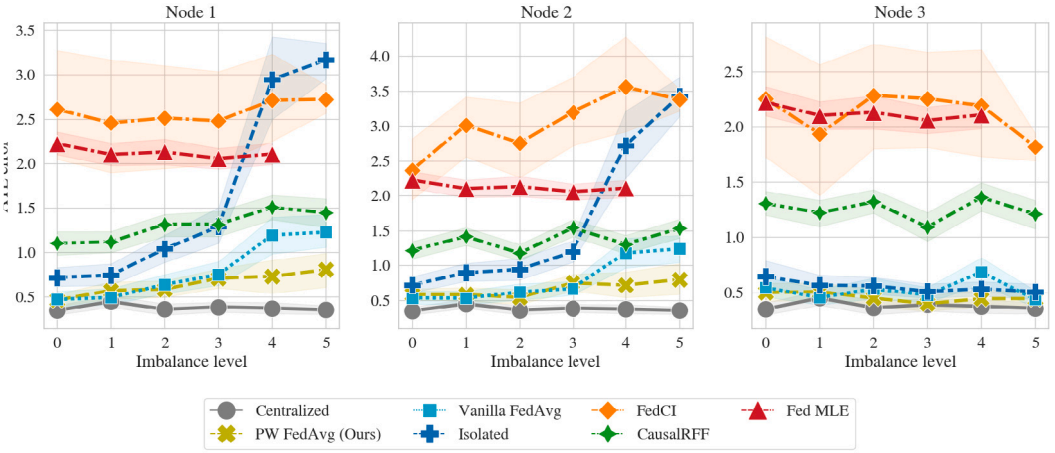


Fig. 14. Imbalances levels in the experiment of three nodes. In this case, five levels of imbalances are presented. **Level 0** (balanced): 33/148 treated/control patients in each node. **Level 1**: 25/156 treated/control patients in node 1 and 2, 49/132 treated/control patients in node3. **Level 2**: 17/164 treated/control patients in node 1 and 2, 65/124 treated/control patients in node3. **Level 3**: 9/172 treated/control patients in node 1 and 2, 81/108 treated/control patients in node3. **Level 4**: 1/180 treated/control patients in node 1 and 2, 97/92 treated/control patients in node3. **Level 5**: 0/181 treated/control patients in node 1 and 2, 99/90 treated/control patients in node3. The imbalance is increasing, while the number of patients in each node remains the same.



(a) PEHE



(b) |ATE error|

Fig. 15. Metrics (Mean and 95% CI) in three nodes experiment. Larger imbalance level correspond to a larger different of treated/control patients in each node. PW FedAvg (Ours) outperforms other distributed methods in larger imbalances.

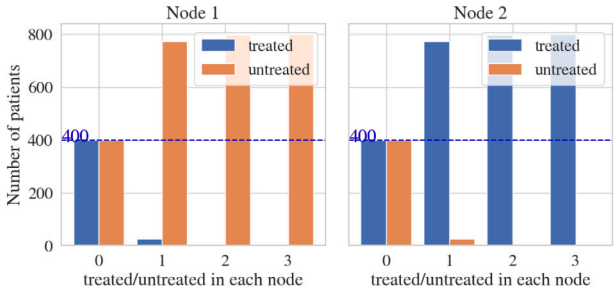
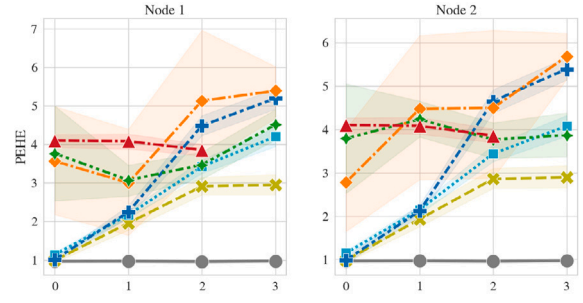
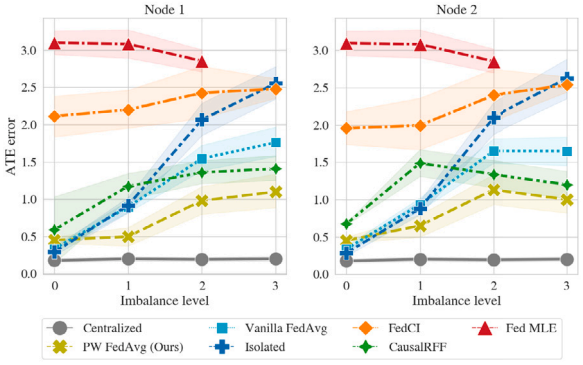


Fig. 16. Imbalances of the two-nodes ACIC experiment. **Level 0** (balanced): 400/400 treated/control patients in each node. **Level 1**: 25/775 treated/control patients in node 1, 775/25 treated/control patients in each node. **Level 2**: 1/799 treated/control patients in node 1, 799/1 treated/control patients in node 2. **Level 3**: 0/800 treated/control patients in node 1, 800/0 treated/control patients in node 2.



(a) PEHE



(b) |ATE error |

Fig. 17. Metrics in ACIC 2016 dataset. Our *Propensity Weighted* `FedAvg` brings the gap between centralized and *Vanilla* `FedAvg` in both nodes in both PEHE and ATE errors.

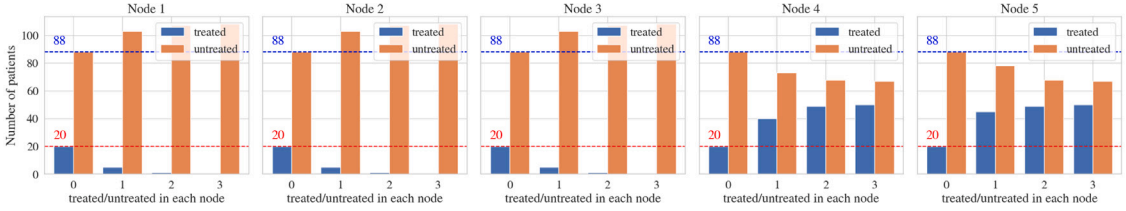
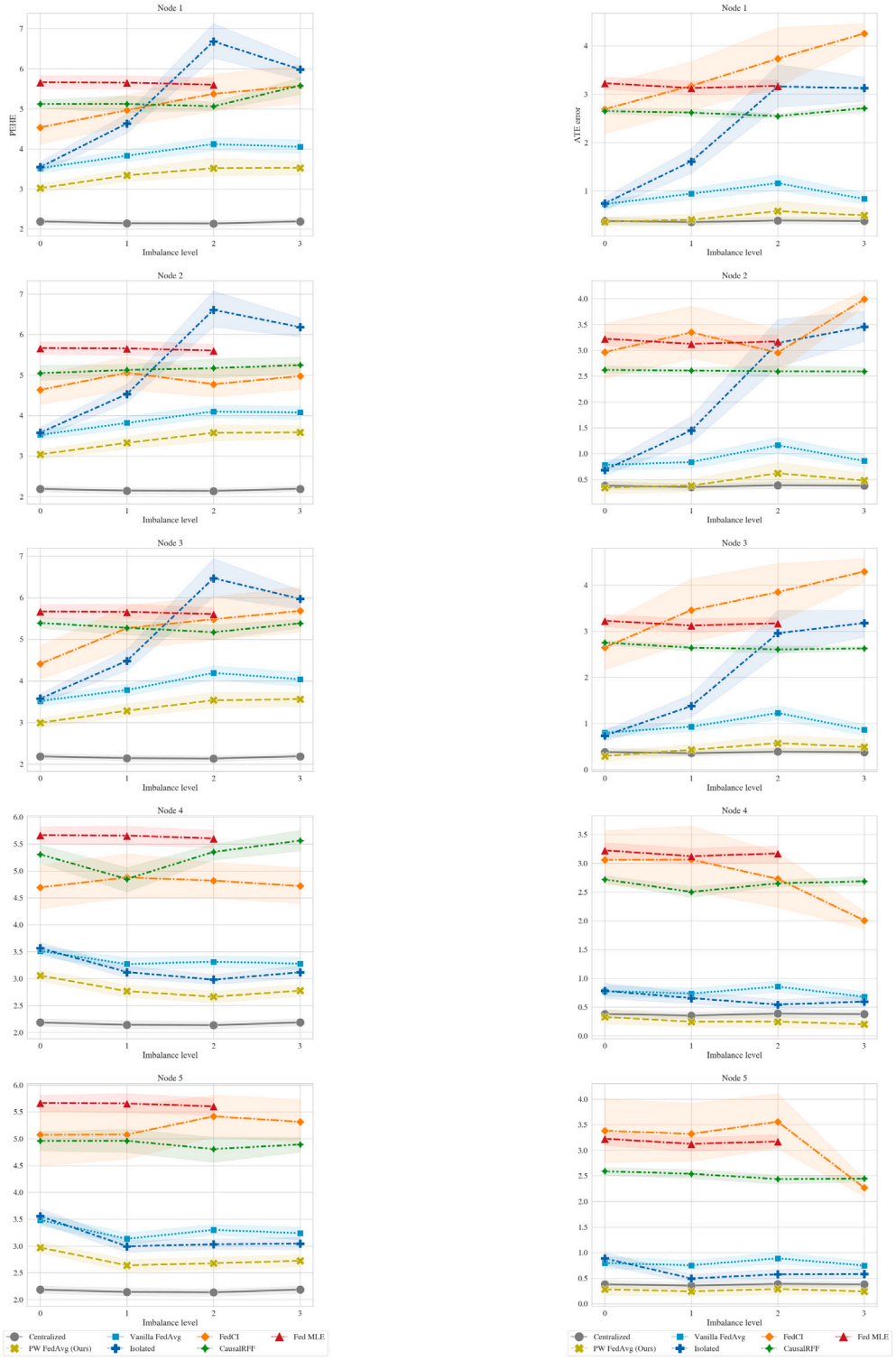


Fig. 18. Imbalances for five nodes experiment with IHDP Setting B. **Level 0** (balanced): 20/88 treated/control patients in each node. **Level 1**: 5/103 treated/control patients in node 1, 2 and 3; 40/73 treated/control patients in nodes 4 and 5. **Level 2**: 1/107 treated/control patients in node 1, 2 and 3; 49/68 treated/control patients in nodes 4 and 5. **Level 3**: 0/108 treated/control patients in node 1, 2 and 3; 50/67 treated/control patients in nodes 4 and 5.

Table 10
PEHE and |ATE error| (Mean(std)) Out-of-sample results of five nodes experiment with IHDP Setting B. **Lower is better**. The results in bold represent the best results of a one-way t-test with $p < 0.05$ in the PEHE table and $p < 0.1$ in the ATE table. Our method, PW `FedAvg`, is marked with *. Note that, because the total training data have been spread over more nodes, the metrics of the distributed algorithms are worse than in the other experiments already at the equilibrium position. Therefore, the differences between *Vanilla* and *Propensity Weighted* `FedAvg` are not as noticeable as in other experiments because the amount of data at each node is small and the errors committed are greater.

		Imbalance 0					Imbalance 1					Imbalance 2					Imbalance 3				
		Node1 20/88	Node2 20/88	Node3 20/88	Node4 20/88	Node5 20/88	Node1 5/103	Node2 5/103	Node3 5/103	Node4 40/73	Node5 40/73	Node1 1/107	Node2 1/107	Node3 1/107	Node4 49/68	Node5 49/68	Node1 0/108	Node2 0/108	Node3 0/108	Node4 50/67	Node5 50/67
PEHE	Centralized	2.10(0.18)																			
	PW FedAvg*	**3.32(0.53)**	**3.34(0.53)**	**3.30(0.49)**	**3.36(0.44)**	**3.27(0.45)**	**3.64(0.68)**	**3.62(0.71)**	**3.58(0.72)**	**3.07(0.54)**	**2.94(0.53)**	**3.82(1.00)**	**3.87(0.92)**	**3.83(0.87)**	**2.96(0.48)**	**2.98(0.50)**	**3.83(0.76)**	**3.88(0.80)**	**3.86(0.78)**	**3.08(0.49)**	**3.02(0.51)**
	Vanilla FedAvg	**3.32(0.54)**	**3.32(0.49)**	**3.32(0.52)**	**3.31(0.44)**	**3.29(0.51)**	3.83(0.66)	3.82(0.74)	3.78(0.69)	**3.27(0.54)**	**3.12(0.73)**	4.10(0.69)	4.19(0.76)	**3.32(0.49)**	**3.30(0.47)**	3.49(0.51)	4.05(0.74)	4.08(0.75)	4.04(0.76)	**3.28(0.51)**	**3.24(0.53)**
	Isolated	3.55(0.61)	3.58(0.61)	3.57(0.53)	3.57(0.51)	3.56(0.65)	4.63(1.11)	4.53(1.04)	4.48(1.14)	3.12(0.48)	3.00(0.51)	6.68(1.91)	6.61(2.06)	6.46(2.13)	2.98(0.43)	3.03(0.43)	5.99(1.11)	6.18(1.10)	5.97(1.08)	3.12(0.53)	3.04(0.52)
	CausalRFF	5.20(1.01)	4.95(1.21)	5.28(0.73)	5.15(0.91)	4.90(0.76)	4.77(0.71)	5.36(0.63)	5.20(1.26)	5.05(0.84)	5.08(1.19)	5.07(1.12)	5.35(1.16)	5.25(1.08)	5.14(0.54)	5.42(0.97)	5.42(1.03)	5.18(0.92)	5.21(0.85)	5.28(0.81)	5.47(1.65)
	FedCI	4.54(1.84)	4.63(1.58)	4.41(1.69)	4.70(1.65)	5.07(2.28)	4.97(1.69)	5.06(1.78)	5.27(2.31)	4.88(1.87)	5.08(2.17)	5.38(2.05)	4.77(1.45)	5.48(2.32)	4.82(1.50)	5.42(1.72)	5.58(2.05)	4.97(1.45)	5.68(2.32)	4.72(1.50)	5.32(1.72)
	Fed MLE	5.67(0.95)	5.67(0.95)	5.67(0.95)	5.67(0.95)	5.67(0.95)	5.66(1.09)	5.66(1.09)	5.66(1.09)	5.66(1.09)	5.66(1.09)	5.60(0.97)	5.60(0.97)	5.60(0.97)	5.60(0.97)	5.60(0.97)	–	–	–	–	–
ATE	Centralized	0.35(0.26)																			
	PW FedAvg*	**0.66(0.45)**	**0.64(0.54)**	**0.59(0.52)**	**0.63(0.57)**	**0.58(0.51)**	**0.70(0.53)**	**0.68(0.58)**	**0.73(0.51)**	**0.55(0.44)**	**0.54(0.44)**	**0.88(0.94)**	**0.92(0.88)**	**0.87(0.80)**	**0.55(0.48)**	**0.59(0.49**	**0.79(0.61)**	**0.78(0.64)**	**0.79(0.72)**	**0.50(0.38)**	**0.54(0.38)**
	Vanilla FedAvg	**0.63(0.37)**	**0.68(0.48)**	**0.70(0.49)**	**0.68(0.44)**	**0.69(0.55)**	0.94(0.56)	0.84(0.56)	0.93(0.52)	0.74(0.42)	0.75(0.40)	3.16(2.12)	3.14(2.16)	2.96(2.14)	**0.54(0.41)**	**0.57(0.45)**	0.83(0.50)	0.86(0.58)	0.86(0.53)	0.68(0.39)	0.75(0.38)
	Isolated	**0.74(0.65)**	**0.69(0.53)**	0.73(0.53)	0.78(0.62)	0.88(0.59)	1.61(1.21)	1.45(1.17)	1.38(1.09)	**0.66(0.44)**	**0.49(0.39)**	0.74(0.65)	0.69(0.53)	0.73(0.53)	0.78(0.62)	0.88(0.59)	3.12(1.24)	3.45(1.29)	3.18(1.32)	**0.60(0.46)**	**0.58(0.43)**)
	CausalRFF	2.66(0.42)	2.57(0.48)	2.70(0.31)	2.68(0.38)	2.56(0.32)	2.47(0.30)	2.71(0.25)	2.64(0.51)	2.58(0.35)	2.59(0.50)	2.57(0.43)	2.67(0.47)	2.62(0.43)	2.57(0.24)	2.71(0.41)	2.66(0.43)	2.55(0.38)	2.57(0.35)	2.60(0.33)	2.67(0.68)
	FedCI	2.69(2.29)	2.97(2.16)	2.65(2.18)	3.06(2.01)	3.38(2.81)	3.17(2.21)	3.35(2.23)	3.46(2.86)	3.07(2.38)	3.32(2.61)	3.73(2.66)	2.95(1.86)	3.85(2.76)	2.73(2.13)	3.56(2.34)	4.25(0.89)	3.99(0.63)	4.30(1.01)	2.01(0.65)	2.27(0.75)
	Fed MLE	3.23(0.80)	3.23(0.80)	3.23(0.80)	3.23(0.80)	3.23(0.80)	3.12(0.87)	3.12(0.87)	3.12(0.87)	3.12(0.87)	3.12(0.87)	3.17(0.82)	3.17(0.82)	3.17(0.82)	3.17(0.82)	3.17(0.82)	–	–	–	–	–



(a) PEHE (b) |ATE error|

Fig. 19. Metrics of five nodes experiment with IHDP Setting B. Although the performance of PW FedAvg is close to the rest of other algorithms due to the sparsity of samples in each node, it is still the best distributed strategy for imbalanced data.

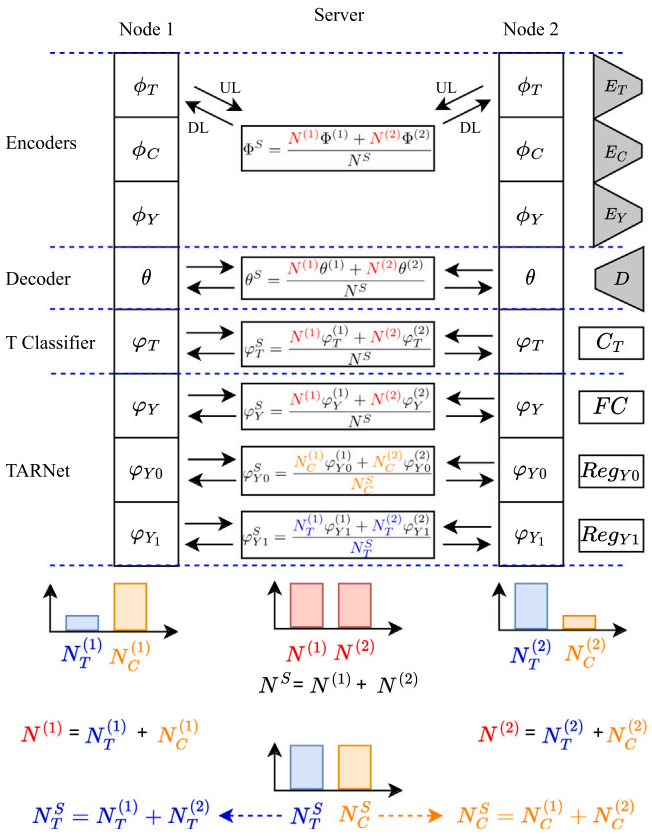


Fig. 20. Schema of one averaging process of *Propensity Weighted* FedAvg in server for two nodes.

For that reason, our adaptation, *Propensity Weighted* FedAvg, modify the average of the parameters of the regressors $Reg0$ ad $Reg1$, by applying Eq. (9). The average in the rest of parameters is not modified (see Table below).

Vanilla FedAvg	*Propensity Weighted* FedAvg
$\phi^S_{T_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \phi^{(k)}_{T_{t+1}}$	$\phi^S_{T_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \phi^{(k)}_{T_{t+1}}$
$\phi^S_{C_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \phi^{(k)}_{C_{t+1}}$	$\phi^S_{C_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \phi^{(k)}_{C_{t+1}}$
$\phi^S_{Y_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \phi^{(k)}_{Y_{t+1}}$	$\phi^S_{Y_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \phi^{(k)}_{Y_{t+1}}$
$\varphi^S_{T_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{T_{t+1}}$	$\varphi^S_{T_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{T_{t+1}}$
$\theta^S_{t+1} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \theta^{(k)}_{t+1}$,	$\theta^S_{t+1} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \theta^{(k)}_{t+1}$,
$\varphi^S_{Y_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{Y_{t+1}}$	$\varphi^S_{Y_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{Y_{t+1}}$
$\varphi^S_{Y1_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{Y1_{t+1}}$	$\varphi^S_{Y0_{t+1}} = \sum_{k=1}^{K} \frac{N_T^{(k)}}{N_T^S} \varphi^{(k)}_{Y1_{t+1}}$
$\varphi^S_{Y0_{t+1}} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \varphi^{(k)}_{Y0_{t+1}}$	$\varphi^S_{Y0_{t+1}} \sum_{k=1}^{K} \frac{N_C^{(k)}}{N_C^S} \varphi^{(k)}_{Y0_{t+1}}$

Finally, an illustration of the averaging process performed by our *Propensity Weighted* FedAvg in a two-node scenario is shown in Fig. 20.

Appendix C. Hyperparameters, data shared and time consumption

C.1. Hyperparameters

Hyperparameters of TEDVAE:

- Number of layers of each submodule: 3
- Number of neurons of each layer: 128
- Non-linear activation functions (in hidden layers): ReLU
- Number of epochs: 200 (20 rounds)
- Number of epochs between averaging processes: 10
- Hyperparameters $\alpha_t = \alpha_y = 100$

C.2. Amount of data transmitted

These hyperparameters for IHDP datasets, with 25 input covariates, make a total of 224181 trainable parameters. Each parameter is coded using 8B. Therefore, each node has to share with the server 1.71 MB in each averaging process. In the same way, the server transmits 1.71 MB to each node after the averaging process.

We define the direction nodes→server as upload and the inverse as download, for understanding.

Taking into account the number of epochs of the training process and the federation intervals, this sharing process must be performed 20 times. Therefore, the total amount of data transmitted in the training of our algorithm is:

- upload: 34.2 MB × n_{nodes}
- download: 34.2 MB

However, the comparison methods share information only once. The total amount of data transmitted for each method is:

- CausalRFF: The total number of parameters of the network, for IHDP dataset is 305687. Then, in each iteration are shared: 305687 × 8 B = 2.33 MB. Using 10000 iterations, as defined in the original code:
  - upload: 22.77 GB × n_{nodes}
  - download: 22.77 GB
- FedCI: First of all, it is needed to share the first four moments of each variable ($Y0$ and $Y1$ separately) in each node: (n_{cov} + 1 + 1 + 1) × 4 × 8 B
  In the training process, in each iteration, all the parameters of the net are shared. With the architecture of the model available and IHDP dataset, each source has 310017 parameters. Then, in each iteration, the following is shared: 310017 × 8 B = 2.36 MB
  Using 2000 iterations, the total amount of data shared is:
  - pretrain: 896 B
  - upload: 4.61 GB × n_{nodes}
  - download: 4.61 GB
- Fed MLE: From nodes to central server, the Hessian and the coefficients for treatment and outcome regression are shared. The Hessian is a matrix of shape (n_{cov} + 1 + 1) × (n_{cov} + 1 + 1). There are (n_{cov} + 1) coefficients for treatment regression and (n_{cov} + 1 + 1) coefficients for outcome regression. Then, the central server computes the set of adjusted coefficients for treatment and outcome regressions and send them to the nodes. This model only share the parameters once. Therefore, the total amount of data transmitted for IHDP, where $n_{cov} = 25$ is:
  - upload: (27 × 27 + 26 + 27) × 8 B × n_{nodes} = 9.11 kB × n_{nodes}
  - download: (26 + 27) × 8 B = 53 B

C.3. Time consumption

As can be expected, the federated training in TEDVAE takes more time in training than the centralized training, and also than isolated training, where sharing and averaging parameters are not needed.

However, the training and inference time of Federated TEDVAE is much shorter than the CausalRFF and FedCI training time. Finally, training and inference in Fed MLE are much faster than in the other algorithms (see Table 11).

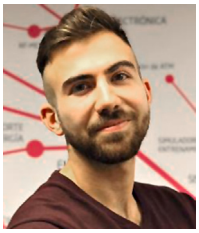
Table 11
Training and inference time in seconds (s). Mean(std) come from 100 realizations of IHDP. Measures of single-thread execution on CPU AMD Ryzen 9 5950X 16-Core Processor.

	Training time	Inference time
Cen TV	18.75(1.44)	3.80(0.11)
Iso TV	33.53(1.51)	8.50(0.15)
Fed TV	49.24(0.80)	8.49(0.18)
CausalRFF	566.1(56.3)	216.7(19.6)
FedCI	252.0(20.8)	3.21(0.83)
Fed MLE	3.02(0.05)	0.00(0.00)

References

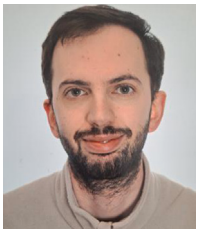
[1] G.W. Imbens, D.B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, Cambridge University Press, 2015.
[2] J. Pearl, An introduction to causal inference, Int. J. Biostat. 6 (2) (2010) 1–62.
[3] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, 2nd Edition, MIT Press Books, vol. 1, The MIT Press, 2001.
[4] B. Schölkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proc. IEEE 109 (5) (2021) 612–634.
[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.
[6] W. Zhang, L. Liu, J. Li, Treatment effect estimation with disentangled latent factors, in: AAAI Conference on Artificial Intelligence, 2020.
[7] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, J. Qadir, Privacy-preserving artificial intelligence in healthcare: Techniques and applications, Comput. Biol. Med. 158 (2023) 106848.
[8] C. Dwork, Differential privacy, in: International Colloquium on Automata, Languages, and Programming, Springer, 2006, pp. 1–12.
[9] N. Truong, K. Sun, S. Wang, F. Guitton, Y. Guo, Privacy preservation in federated learning: An insightful survey from the GDPR perspective, Comput. Secur. 110 (2021) 102402.
[10] P.C. Austin, An introduction to propensity score methods for reducing the effects of confounding in observational studies, Multivar. Behav. Res. 46 (3) (2011) 399–424.
[11] P.R. Rosenbaum, D.B. Rubin, Reducing bias in observational studies using subclassification on the propensity score, J. Amer. Statist. Assoc. 79 (387) (1984) 516–524.
[12] J.K. Lunceford, M. Davidian, Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study, Stat. Med. 23 (19) (2004) 2937–2960.
[13] Y. Xie, J.E. Brand, B. Jann, Estimating heterogeneous treatment effects with observational data, Sociol. Methodol. 42 (1) (2012) 314–347, PMID: 23482633.
[14] S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests, J. Amer. Statist. Assoc. 113 (523) (2018) 1228–1242.
[15] S. Athey, S. Wager, Estimating treatment effects with causal forests: An application, Observational Stud. 5 (2) (2019) 37–51.
[16] S. Athey, J. Tibshirani, S. Wager, Generalized random forests, Ann. Statist. 47 (2) (2019) 1148–1178.
[17] E.H. Kennedy, Towards optimal doubly robust estimation of heterogeneous causal effects, Electron. J. Stat. 17 (2) (2023) 3008–3049.
[18] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins, Double/debiased machine learning for treatment and structural parameters, Econom. J. 21 (1) (2018) C1–C68.
[19] M. van der Laan, S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, in: Springer Series in Statistics, Springer New York, 2011.
[20] M.S. Schuler, S. Rose, Targeted maximum likelihood estimation for causal inference in observational studies, Am. J. Epidemiol. 185 (1) (2017) 65–73.
[21] M.A. Luque-Fernandez, M. Schomaker, B. Rachet, M.E. Schnitzer, Targeted maximum likelihood estimation for a binary treatment: A tutorial, Stat. Med. 37 (16) (2018) 2530–2546.
[22] J.L. Hill, Bayesian nonparametric modeling for causal inference, J. Comput. Graph. Statist. 20 (2011) 217–240.
[23] P.R. Hahn, J.S. Murray, C.M. Carvalho, Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion), Bayesian Anal. 15 (2020) 965–1056.
[24] S.R. Künzel, J.S. Sekhon, P.J. Bickel, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, Proc. Natl. Acad. Sci. 116 (10) (2019) 4156–4165.
[25] S. Li, Y. Fu, Matching on balanced nonlinear representations for treatment effects estimation, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, 2017.
[26] U. Shalit, F.D. Johansson, D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms, in: International Conference on Machine Learning, PMLR, 2017, pp. 3076–3085.
[27] F.D. Johansson, U. Shalit, N. Kallus, D. Sontag, Generalization bounds and representation learning for estimation of potential outcomes and causal effects, J. Mach. Learn. Res. 23 (166) (2022) 1–50.
[28] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, A. Zhang, Representation learning for treatment effect estimation from observational data, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, 2018.
[29] C. Shi, D. Blei, V. Veitch, Adapting neural networks for the estimation of treatment effects, Adv. Neural Inf. Process. Syst. 32 (2019).
[30] C. Louizos, U. Shalit, J.M. Mooij, D. Sontag, R. Zemel, M. Welling, Causal effect inference with deep latent-variable models, Adv. Neural Inf. Process. Syst. 30 (2017).
[31] P. Wu, K. Fukumizu, Intact-VAE: Estimating treatment effects under unobserved confounding, 2022, arXiv preprint arXiv:2101.06662.
[32] I. Khemakhem, D. Kingma, R. Monti, A. Hyvarinen, Variational autoencoders and nonlinear ica: A unifying framework, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2207–2217.
[33] J. Yoon, J. Jordon, M. van der Schaar, GANITE: Estimation of individualized treatment effects using generative adversarial nets, in: International Conference on Learning Representations, 2018.
[34] M. Kocaoglu, C. Snyder, A.G. Dimakis, S. Vishwanath, CausalGAN: Learning causal implicit generative models with adversarial training, in: International Conference on Learning Representations, 2018.
[35] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R.R. Colen, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (1) (2020) 12598.
[36] X. Li, K. Huang, W. Yang, S. Wang, Z. Zhang, On the convergence of FedAvg on non-IID data, in: International Conference on Learning Representations, 2020.
[37] F. Sattler, S. Wiedemann, K.-R. Müller, W. Samek, Robust and communication-efficient federated learning from non-iid data, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (2019) 3400–3413.
[38] R. Xiong, A. Koenecke, M. Powell, Z. Shen, J.T. Vogelstein, S. Athey, Federated causal inference in heterogeneous observational data, Stat. Med. 42 (24) (2023) 4418–4439.
[39] M. Hu, X. Shi, P.X.-K. Song, Collaborative causal inference with a distributed data-sharing management, 2022, arXiv preprint arXiv:2204.00857.
[40] L. Han, J. Hou, K. Cho, R. Duan, T. Cai, Federated adaptive causal estimation (FACE) of target treatment effects, 2022, arXiv preprint arXiv:2112.09313.
[41] T.V. Vo, Y. Lee, T.N. Hoang, T.-Y. Leong, Bayesian federated estimation of causal effects from observational data, in: Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 2024–2034.
[42] T.V. Vo, A. Bhattacharyya, Y. Lee, T.-Y. Leong, An adaptive kernel approach to federated learning of heterogeneous causal effects, Adv. Neural Inf. Process. Syst. 35 (2022) 24459–24473.
[43] A. Almodóvar, J. Parras, S. Zazo, Federated learning for causal inference using deep generative disentangled models, in: Deep Generative Models for Health Workshop NeurIPS 2023, 2023.
[44] J. Pearl, Causality: Models, Reasoning and Inference, second ed., Cambridge University Press, USA, 2009.
[45] D.B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, J. Educ. Psychol. 66 (5) (1974) 688.
[46] A. Curth, M. van der Schaar, Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1810–1818.
[47] A. Curth, D. Svensson, J. Weatherall, M. van der Schaar, Really doing great at estimating CATE? A critical look at ML benchmarking practices in treatment effect estimation, in: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
[48] M.A. Hernán, J.M. Robins, Causal Inference: What If, Chapman & Hall/CRC, Boca Raton, 2020.
[49] J. Peters, D. Janzing, B. Schlkopf, Elements of Causal Inference: Foundations and Learning Algorithms, The MIT Press, 2017.
[50] J. Hahn, On the role of the propensity score in efficient semiparametric estimation of average treatment effects, Econometrica (1998) 315–331.
[51] A. Abadie, G.W. Imbens, Large sample properties of matching estimators for average treatment effects, Econometrica 74 (1) (2006) 235–267.
[52] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, 2022, arXiv preprint arXiv:1312.6114.
[53] R. McDonald, K. Hall, G. Mann, Distributed training strategies for the structured perceptron, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 456–464.

[54] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, V. Chandra, Federated learning with non-iid data, 2018, arXiv preprint arXiv:1806.00582.
[55] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.
[56] V. Dorie, NPCI: Non-parametrics for causal inference, 2016, https://github.com/vdorie/npci/, Online, access April 23th 2023.
[57] F.D. Johansson, Fredik D. Johansson's personal webpage, 2017, https://www.fredjo.com/, Online, access April 23th 2023.
[58] V. Dorie, J. Hill, U. Shalit, M. Scott, D. Cervone, Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, Statist. Sci. 34 (1) (2019) 43–68.
[59] J. Hill, 2016 Atlantic causal inference conference competition: Is your satt where it's at?, 2016, https://jenniferhill7.wixsite.com/acic-2016/competition, Online, access April 23th 2023.
[60] S. Assaad, S. Zeng, C. Tao, S. Datta, N. Mehta, R. Henao, F. Li, L. Carin, Counterfactual representation learning with balancing weights, in: International Conference on Artifi- cial Intelligence and Statistics, 2020.
[61] A. Curth, M. van der Schaar, On inductive biases for heterogeneous treatment effect estimation, Adv. Neural Inf. Process. Syst. 34 (2021) 15883–15894.
[62] D. Cheng, Y. Xie, Z. Xu, J. Li, L. Liu, J. Liu, Y. Zhang, Z. Feng, Disentangled latent representation learning for tackling the confounding M-bias problem in causal inference, in: 2023 IEEE International Conference on Data Mining, ICDM, IEEE, 2023, pp. 51–60.
[63] Y. Liu, J. Wang, B. Li, EDVAE: Disentangled latent factors models in counterfactual reasoning for individual treatment effects estimation, Inform. Sci. 652 (2024) 119578.
[64] N. Sturma, C. Squires, M. Drton, C. Uhler, Unpaired multi-domain causal representation learning, in: Thirty-Seventh Conference on Neural Information Processing Systems, 2023.
[65] J. Yoon, J. Jordon, M. Schaar, Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2018, pp. 5699–5707.



Alejandro Almodóvar received the B.S. degree and the M.Sc. degree in telecommunications engineering from Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2020 and 2022, respectively.

Currently, he is a Ph.D. student whose work focuses on Causal inference from a machine learning perspective, with applications in medicine and healthcare.



Juan Parras received his B.S. in Telecommunications Engineering from Universidad de Jaén in 2014, and MSC and Ph.D. in Telecommunications Engineering from Universidad Politécnica de Madrid (UPM) in 2016 and 2020. He is currently an Assistant Professor at UPM and his research interests include deep generative models, deep reinforcement learning, game theory and optimization with health and communications applications.

Santiago Zazo received the Dr.Eng. degree from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1995.

He joined UPM in 1998, where he is currently a Professor of Signal Theory and Communications. He is the author/coauthor of more than 40 journal papers and about 200 conference papers. His main research activities are in the field of signal processing. More recently, he has been mostly focused on distributed optimization, optimum control, game theory, and reinforcement learning.