FedICE - Federated Invariant Cause Effect Estimation

Sreya Francis
MILA - Quebec AI Institute
University of Montreal

Michael Schuster
CHUM - Hospital Research Centre
University of Montreal

***Abstract*—The use of causal effect estimation methods to estimate treatment outcomes from observational studies is widespread in the field of medicine, social sciences and econometrics. One of the main challenges in this field is data scarcity. Although these methods rely heavily on highly sensitive information, so far there has been a huge lack of privacy preserving approaches for the same. Another challenge is about generalization capacity of the causal inference model trained on different source domains for a specific case since the counterfactuals are never observed as well as the possibility of non identically distributed data in source/target domains. In this paper, we aim to tackle these challenges with the help of a federated invariant learning framework. Our approach can help exploit additional data sources to facilitate privacy-preserving causal effect estimation in an unseen target population. We evaluate our proposed framework on synthetic and semi-synthetic datasets and show that the empirical results with our distributed approaches are almost consistent with the current centralized approaches to treatment effect estimation with added advantage of better generalization.**

***Index Terms*—Federated Learning, Out of Distribution Generalization, Causal Machine Learning, Privacy, Treatment Estimation, Robustness**

I. Motivation

Large data sources such as electronic medical records present opportunities to study causal effects of interventions that are difficult to evaluate through experiments. Privacy is not a requirement confined to the medical field but extends to various other datasets pertaining to education, goverment records, legal information as well. All of these specified fields and more, often contain very sensitive personal information, due to which conclusions drawn from causal analysis made on such datasets are at immense risk of violating the privacy of participants involved in the studied observational datasets. This has been a very under-studied problem with little to no improvements proposed so far. The datasets compiled from these studies, namely Randomized Control Trial(RCT) data, serve as the base data for treatment effect estimation methods. But in real case scenarios, we only have access to observational data and not RCTs. More the data, better the estimation. To enhance this, we will need access to data throughout multiple hospitals within and across countries. As access to such health records are mostly restricted owing to privacy issues, the only way forward is to use federated learning or similar privacy preserving distributed learning approaches for this very purpose.

II. Background

A. Federated Learning

Federated Learning is an approach to Distributed Machine Learning developed by the Google AI team (10)(8); this approach allows users to keep ownership of their data during the model training process. In addition to keeping ownership of their data, users also have immediate access to the newest model after they have trained it on their data. The sender of the model benefits from distributed data, lower latency, and less local power consumption. Since Federated Learning ensures privacy, more users who don't share their data currently become more willing to partake in the training process and ultimately creating smarter models. The training process involves a central server sending out a model to a subset of the current users. This set downloads the model, trains the model on their device with their local data, and returns only the update that resulted in the new model. Each device returns their particular update and the central server aggregates these updates to improve the original model, which is the new global model (10)(8)(6).

B. Cause Effect Estimation

When studying treatment effects under the potential outcomes framework, we are interested in the effect a treatment T has on an individual, group or instance (called unit) i that has some features $X_i = \{x_{i1}, x_{i2}, \dots x_{id}\}$. Usually, we're interested in the outcome Y_i depending if treatment was given or not. That is to say $T_i \in \{0, 1\}$. Thus there are two potential outcomes for each individual: $Y_i(1)$ if i received treatment and $Y_i(0)$ if it did not receive treatment or received a control/placebo. It is obvious from this construction that in a realistic setting we can only ever see either $Y_i(1)$ or $Y_i(0)$ for any one unit and experiment (13). This is why it is called the potential outcomes framework. Formally, we can write this mutually exclusive relationships as

$$Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$$

The treatment effect τ is then defined as

$$\tau_i = Y_i(1) - Y_i(0)$$

The quantity we are aiming to estimate here is never observable, unlike usual machine learning scenario, where we usually

have ground truth for our data. This analysis is the analysis of a counterfactual world where in we are required to make statements about the counterfactual outcome that we did not observe to estimate the treatment effect . This framework is also used to define the individual treatment effect (ITE) which is defined as the difference between potential outcomes of a certain sample/patient under two different treatments. On arbitrary populations, this can further be extended to Average Treatment Effect(ATE)

1) Propensity Score Reweighting: The propensity score is the probability of treatment given the covariates of a unit.

$$p(x) = P(T = 1|X = x) = E[T|X = x].$$

This propensity score estimate can be used to create a pseudo-population by weighting each sample with the inverse propensity score (IPS-Weighting, or IPSW) (13). Essentially, this gives a higher weight to instances that are underrepresented. This means that treated/control instances whose predicted probability of treatment/control is very low, are weighted very high. The idea behind propensity score matching is that a logit model is used to estimate the probability that each observation in our dataset was in the treatment or control group. Then we use the predicted probabilities to prune out dataset such that, for every treated unit, there's a control unit that can serve as a viable counterfactual.

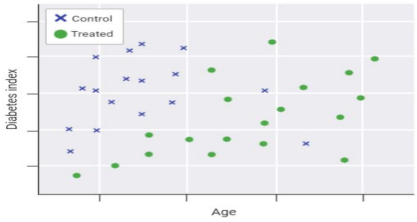

Fig. 1. Inverse Propensity Score Reweighting (p(x—T=0) != p(x—T=1))

Algorithm 1 PSRW - Propensity Score Reweighting

Estimate $p(T = t \mid x)$ using regression
Evaluate $\widehat{ATE} = \frac{1}{n}\sum_{i,t_i=1}\frac{y_i}{\bar{p}(t_i=1|x_i)} - \frac{1}{n}\sum_{i,t_i=0}\frac{y_i}{\bar{p}(t_i=0|x_i)}$

While evaluating ATE, we are multiplying (reweighting) each y_i by its' inverse propensity score. ATE is given by $\mathbb{E}_{x\sim p(x)}\left[\mathbb{E}\left[Y_1 \mid x, T = 1\right] - \mathbb{E}\left[Y_0 \mid x, T = 0\right]\right.$

In real case scenarios, we are limited to just samples for:

$$\mathbb{E}_{x\sim p(x|T=1)}\left[\mathbb{E}\left[Y_1 \mid x, T = 1\right]\right]$$
$$\mathbb{E}_{x\sim p(x|T=0)}\left[\mathbb{E}\left[Y_0 \mid x, T = 0\right]\right]$$

Applying Bayes Theorem, we have:

$$p(x) = p(x \mid T = 1) * \frac{p(T = 1)}{p(T = 1 \mid x)}$$

$$p(x) = p(x \mid T = 0) * \frac{p(T = 0)}{p(T = 0 \mid x)}$$

Here, $p(T = 0 \mid x)$ and $p(T = 1 \mid x)$ give us the propensity score.

2) Generic learners: **S Learner** Single-Learner(S Learner) employs a single supervised machine learning algorithm or regression for estimation of the combined response function

$$\mu(x, t) := \mathbb{E}[Y \mid X = x, T = t]$$

Given $\hat{\mu}$ as the estimator of μ, the conditional treatment effect can be estimated as:

$$\tau(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

When the control and treatment groups are very different in covariates, a single linear model is not sufficient to encode the different relevant dimensions and smoothness of features for the control and treatment groups(9).

T Learner Instead of a single estimator, T-Learners make use of two estimators to execute the same task(9). In this case, we split the dataset into two, namely treated and control so as to learn an outcome regression on each of the subsets.

Let estimate learnt using observations in the treated group be:

$$\mu_0(x) = \mathbb{E}[Y \mid X = x, T = 1]$$

Similarly let estimates learnt using observations in the control group be:

$$\mu_1(x) = \mathbb{E}[Y \mid X = x, T = 0]$$

The above learnt estimates $\hat{\mu}_1$ and $\hat{\mu}_0$ can be used to get

$$2\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

X Learner In X Learner, the outcome functions, $\hat{\mu}_0$ and $\hat{\mu}_1$, are estimated and the individual treatment effects are imputed:

$$\hat{\tau}_i^1 := Y(1) - \hat{\mu}_0\left(X_i\right) \quad \text{and} \quad \hat{\tau}_i^0 := \hat{\mu}_1\left(X_i\right) - Y_i(0)$$

In the second stage, estimators for the CATE are derived by regressing the features X on the imputed treatment effects(9). In the X-learner, the estimators of the first stage are held fixed without an update in the second stage. This is helpful for machine learning algorithms such as RF and BART which cannot be updated once they have been trained.

Y Learner In Y-learner, instead of first deriving an estimator for the control response functions and then an estimator for the ATE function, these functions are optimized jointly(9).

C. Domain Generalization

It is next to impossible to observe client level data to be distributed in an independent and identical manner across different client domains in a federated setting. When training a federated learning system with the i.i.d. assumption, an implicit assumption is made on the underlying data generating process defining an envrionment for the client level data which is definitely not applicable in real world deployment.(5)(3) This issue is also prevalent in the field of cause effect estimation wherein we might want to do causal inference with data collected from multiple environments, as in the hospitals.

For a distribution $P(X, T, Y)$, if a representation is invariant across all valid environments then the information in that representation is exactly the information in the causal parents of Y which can be considered the main connection between causality and invariance. Also a representation capturing only the causal parents will be invariant when the causal structure relevant to the outcome is invariant across participating client domains(15)

To obtain an invariant representation, we rely on invariant risk minimization (IRM)(1), a popular domain generalization framework which enhances learning features that are invariant across domains on one hand and ignores the differing features across domains on the other. To employ IRM in a federated setting, (3) has proposed a federated invariant learning framework.

III. Problem

The problem under consideration demands doing causal inference with data coming from several clients/hospitals with rich observed covariates that includes all the causal parents of the outcome in a federated setting. Eventhough there are no unobserved confounders in this setting, there is no guarantee on identifiability (11) or strong ignorability (12) owing to the possible existence of bad controls. Our target here is to make use of these multiple client environments to find a representation of the covariates for valid causal estimation in a federated setting. We address this problem with a new federated learning framework - FedICE for invariant causal estimation, an estimation framework for causal inference from observational data where the data comes from multiple clients/users. Each client dataset is drawn from distinct environments consisting of distinct covariate distributions.

IV. Contribution

In this work, we explore the idea of a distributed/federated invariant learning approach to cause effet estimation to enhance access to otherwise sensitive datasets with added advantage of generalization to unseen domains.We try to bridge the fields of distributed learning, treatment effect estimation and invariant representation learning. To the best of our knowledge, this is the first attempt aiming to bridge these important fields. We have an initial set of experiments that validate our proposed approach. However, analyzing the domain generalization capabilities in the proposed framework is still ongoing work.Our analysis is based on previous work on invariant cause effect estimation by (15)

V. Proposed Approach - FedICE

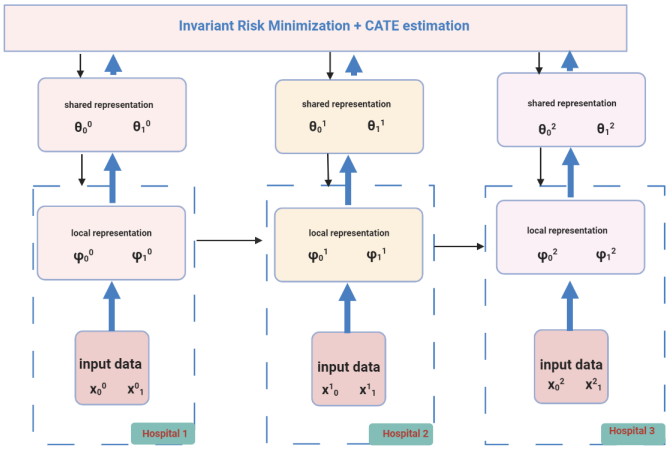

Fig. 2. FedICE
Architecture Overview - In this depiction, we show 2 representation per client which is just one sample scenario. With most generic learners, we need to learn just one common representation per client

A. Setup

Consider multiple datasets $D_e := \{(X_k^e, Y_k^e, T_k^e)\}_{k=1}^{n_e}$, from multiple client domains $e \in \mathcal{E}_{tr}$ consisting of observations of the feature vector $(x \in X)$ and correspondig treatment response $(y \in Y)$. Our goal is to produce a predictor that is robust to changes in the client/server test domain with the application of Invariant Risk Minimization (IRM) (1) for causal adjustment. The data from multiple client environments is used to learn an invariant representation $\Phi(T, X)$, a function such that the outcome Y and the representation of the treatment and covariates $\Phi(T, X)$ have the same relationship in each client environment. Predictors built on top of this representation will have the desired robustness as is already proved in (15). Keeping the individual client data private, each participating client trains the base neural network layers for extracting base level features from their respective input data which is then passed on to the Global Server layer which uses them as input to learn a predictor $f : \Phi(T, X) \to Y$ that minimizes the maximum risk over all the domains $\mathcal{E}$ which implies $\min_f \max_{e \in \mathcal{E}} R_e(f)$ where $R_e(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_e}[l(f(\mathbf{x}), y)]$ is the risk under domain e for a convex and differentiable loss function l(1). The equation is given by

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R_e(\Phi(T, X)) + \lambda \left\|\nabla_{w|w=1.0} R_e(w \cdot \Phi(T, X))\right\|^2$$

where $\lambda \in [0, \infty)$ is a regularizer balancing between the first term(standard ERM), and the invariance of the predictor $1 \cdot \Phi(x)$. $\Phi : \mathcal{X} \to \mathcal{Y}$ is an invariant predictor, $w = 1.0$ is just a dummy classifier, the gradient norm penalty measures the optimality of the dummy classifier at each domain e. The invariant representation learnt with IRM suffices for causal adjustment due to the fact that a representation is invariant if

and only if it is informationally equivalent to the causal parents of the outcome Y (1). In our setting, the causal parents of Y constitute an adjustment set that suffices for causal adjustment following the assumption that there are no variables on the causal path between the treatment and outcome in the covariate set resulting in little to zero impact on overlap as well as exclusion of all bad controls. Hence, adjusting for an invariant representation is a safe way to estimate the causal effect.

In our framework, $\left[\theta, \phi_0^{(k)}\right]$ and $\left[\theta, \phi_1^{(k)}\right]$ are trainable parameters used to predict μ_0^k and $\mu_1^{(k)}$ denoted as $\omega_0^{(k)}$ and $\omega_1^{(k)}$ respectively. Using the invariant predictor obtained, we estimate the Conditional Average Treatment Effect as $\hat{\mu}_1 - \hat{\mu}_0$. In our experiments, we have followed the Rubin-Neyman potential outcomes framework (13). See Algorithm for details.

B. Algorithm

Algorithm 2 FedICE - Server Side Code

ServerCausalUpdate:

Initialize $\mathbf{W}_0^s$
for each server epoch, t = 1,2,..k **do**
  Select random set of S clients
  Share initial model with the selected clients
  for each client $k \in S$ **do**
    $(\boldsymbol{\phi}(x_t^k), \mathbf{Y}^k) \leftarrow ClientRepresentation(k, \mathbf{W}_t^k)$
    Sample $\boldsymbol{\phi}(x_0^k)$ and $\boldsymbol{\phi}(x_1^k)$: control and treatment units from client k
    for $j = [0, 1]$ **do**
      $\mathrm{L}_j^{(k)} = \left\| \pi_{\omega_j^{(k)}} \left(\boldsymbol{\phi}(x_j^k) \right) - \mu_j \left(\boldsymbol{\phi}(x_j^k) \right) \right\|$
    end for
  end for
  $\mathcal{L}_s = \sum_k^S \sum_j \mathcal{L}_j^{(k)} + \lambda \sum_k^S \sum_j \left\| \nabla \mathcal{L}_j^{(k)} \right\|^2$
  Compute $\nabla_\Omega \mathcal{L}_s$
  $\mathbf{W}_{t+1}^s \leftarrow \mathbf{W}_t^s - \eta \nabla \mathcal{L}_s$
end for
$\mathbf{W}_t^k \leftarrow ClientUpdate(\nabla \mathcal{L}_s)$

Algorithm 3 FedICE - Client Side Code

ClientRepresentation($\mathbf{W}_t^k$):

if k is first client to start training **then**
  $\mathbf{W}_t^k \leftarrow$ initial weights from server
else
  $\mathbf{W}_t^k \leftarrow \mathbf{W}_{t-1}^{k-1}$from the previous $ClientUpdate(\nabla \mathcal{L}_s)$
end if
for each local client epoch, i=1,2,..k **do**
  Calculate hidden representation $\boldsymbol{\phi}(x_t^k)$
end for
return $\boldsymbol{\phi}(x_t^k)$ and $\mathbf{Y^k}$ to server

ClientUpdate:

for each client $k \in S$ **do**
  $\mathbf{W}_{t+1}^k \leftarrow \mathbf{W}_t^k - \eta \nabla \mathcal{L}_s$
end for
return $\mathbf{W}_{t+1}^k$ to server

CATE estimation:

for each client $k \in S$ **do**
  Sample X : test units from client k
end for
$\hat{\mu}_0 = \pi_{\omega_0^{(k)}}(X)$
$\hat{\mu}_1 = \pi_{\omega_1^{(k)}}(X)$
return ATE estimate $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$

VI. Experiments

A. Dataset

1) Semi Synthetic Data: This is experimental data from the Infant Health and Development Program (IHDP), a randomized experiment that began in 1985, targeted low-birth-weight, premature infants, and provided the treatment group with both intensive high-quality child care and home visits from a trained provider. The program was highly successful at significantly raising cognitive test scores of the treated children relative to controls at the end of the intervention (2) IHDP is a real-world dataset with 25 covariates describing 747 children and their mothers, derandomised binary treatments and synthetic continuous outcomes that can be used to compute a ground truth ATE (4) In (4), a total of six continuous and nineteen binary pretreatment variables in a randomised control trial. Using the covariates of all instances in both treatment groups, the potential outcomes are generated synthetically. In our experiments, we use the train and test sets from (14). We create balanced client datasets by sampling with replacement 100 units with T = 1 and T = 0 and 50 units with T = 1 and T = 0 respectively from the above train and test sets.

2) Synthetic Data: Here we resort to synthetic settings to evaluate the generalizability of our approach. We generate simulated data for 5 client domains each ranging in size from 3,000 units to 5,000 units and the covariate vector is 10 dimensional : Let N be the total units sampled from the underlying experiment to get the features and the treatment assignment of client samples $(X_i, T_i)_{i=1}^N$. Sample $\beta^0 = \left(\beta_1^0, \ldots, \beta_d^0\right) \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\beta^1 = \left(\beta_1^1, \ldots, \beta_d^1\right) \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then we have $\mu_0(x) = \text{logistic}\left(x\beta^0\right)$ and $\mu_1(x) = \text{logistic}\left(x\beta^1\right)$. Once sampling is done, true ATE for each unit is generated using $\tau_i = \tau(X_i) = \mu_1(X_i) - \mu_0(X_i)$. This is followed by generation of observed outcome by sampling a Bernoulli distributed variable around mean μ_i.

$$Y_i^{obs} \sim \text{Bern}(\mu_i)$$

B. Inference

Evaluation metric used here is the Precision in Estimation of Heterogeneous Effects(PEHE), which is a mean squared error comparing individual treatment effect estimates with true effects, defined as follows:

$$\epsilon_{\text{PEHE}} = \frac{1}{n} \sum_i^n (\hat{\tau}(x_i) - \tau(x_i))^2$$

C. Centralized Vs Federated Setting Analysis

We experimented our proposed approach on 3 causal effect estimators namely Y learner, S learner and T learner. In the experiment setting details, the specifications pertain to T learner. But with slight modifications, the same approach can be followed for S learner and other learners with just the adjustment to be done based on the number of estimators for training as well as the loss function. We observed that our proposed approach almost near the performance of centralized approaches for cause effect estimation with added advantage of retaining client data privacy and out of distribution generalization.

TABLE I
CENTRALIZED CAUSE EFFECT ESTIMATION RESULTS

Data	Causal Estimation Method	$\sqrt{\epsilon_{\text{PEHE}}}$
IHDP	T-Learner	2.4180 ± 0.1
IHDP	S-Learner	5.6920 ± 0.3
IHDP	Y-Learner	2.3973 ± 0.3
Synth	T-Learner -	3.378 ± 0.4
Synth	S-Learner	9.457 ± 0.5
Synth	Y-Learner	2.9273 ± 0.4

TABLE II
FEDERATED CAUSE EFFECT ESITMATION RESULTS

Data	Causal Estimation Method	$\sqrt{\epsilon_{\text{PEHE}}}$
IHDP	FedICE T-Learner	2.9560 ± 0.5
IHDP	FedICE S-Learner	6.2520 ± 0.4
IHDP	FedICE Y-Learner	2.5879 ± 0.1
Synth	FedICE T-Learner	3.213 ± 0.3
Synth	FedICE S-Learner	9.245 ± 0.3
Synth	FedICE Y-Learner	2.8437 ± 0.2

We observed that the estimated treatment scores rarely exactly match the true ones with extreme covariate imbalance in participating client subgroups resulting in a potential bias in federated client causal effect estimation. This specific causal effect analysis in a federated setup exposed a high bias-variance tradeoff which in turn resulted in an estimated score with high complexity. A higher score complexity can lead to better covariate balance within client groups but with a higher variance.

D. Federated domain invariance analysis

1) Linear Setting: Our data setting for this analysis is entirely inspired from (15). We simulate data with the three causal graphs wherein each intervention e generates a new environment e with interventional distribution $P\left(X^e, T^e, Y^e\right)$ with T^e denoting binary treatment, Y^e denoting outcome and X^e denoting a 10 -dimensional covariate set that differs across DGPs. For the hyper-parameter λ used in IRMv1, we use $\lambda = 10$. $X^e = (X_1^e, X_2^e)$, where X_1^e is a five-dimensional confounder. X_2^e is either noise, a descendant, or a collider in each DGP. The DGPs are:

$$
\begin{aligned}
X_1^e &\leftarrow N\left(0, e^2\right) \\
T^e &\leftarrow \text{Bern}\left(\text{sigmoid}\left(X_1^e \cdot w_{xt^e} + N(0,1)\right)\right) \\
\tau &\leftarrow 5 + N\left(0, \sigma^2\right) \\
Y^e &\leftarrow X_1^e \cdot w_{xy^e} + T^e \cdot \tau + N\left(0, e^2\right)
\end{aligned}
$$

In (a) $X_2^e \leftarrow N(0,1)$, in (b) $X_2^e \leftarrow e * Y^e + N(0,1)$, and in (c) $X_2^e \leftarrow e * Y^e + T^e + N(0,1)$. For evaluation, following (Arjovsky et al., 2019), we create three environments $E = \{0.2, 2, 5\}$. We ran 20 simulations. We draw 1000 samples from each environment in each simulation. In our initial setting where $X_2^e \leftarrow N(0,1)$, FedICE performs almost as good as adjusting for all covariates. In the second setting where $X_2^e \leftarrow e * Y^e + N(0,1)$, we observe a reduction in estimation bias. In the third setting $X_2^e \leftarrow e * Y^e + T^e + N(0,1)$, we observe a similar result as the previous one with reduced estimation bias.

2) Non linear setting: The experiment setting is same as (15). In the nonlinear settings, we use TARNet (14) which is a double headed model with a shared representation $\Phi(X)$ and two heads for the treated and control representation. The network has 4 layers for the shared representation and 3 layers for each expected outcome head. The hidden layer size is 250 for the shared representation layers and 100 for the expected outcome layers. We use Adam (7) as the optimizer, set the learning rate as 0.001, and an 12 regularization rate of 0.0001.For the hyper-parameter λ used in IRMv1, we use $\lambda = 100$.

We first draw three client environments $\{C^{e_1}, C^{e_2}, C^{e_3}\}$ that are diverse. To control the level of environment variation, we construct three new environments $\{C_1', C_2', C_3'\}$ that are mixtures of the three source environments. Respectively, $C^{e_1'}, C^{e_2'}, C^{e_3'}$ draw (c_1, c_2, c_3) proportions from C^{e_1}, $(c_2, c_3, , c_1)$ proportions from C^{e_2}, and (c_2, c_3, c_1) proportions from C^{e_3}. The proportions (c_1, c_2, c_3) sum to one.

The diversity of the environments is approximated by the diversity of the proportions with a diversity measure which is equated by $\frac{1}{3}\sum_{ij}|c_i - c_j|$. In our federated setting, we draw 26 set of new client environments, induced by different combination of the mixture probabilities and use them to compare the estimation quality of FedICE when given a covariate set that include bad controls against adjusting for a valid covariate set.

As shown in the plot, more the number/diversity of the client environments in the federated setting, higher the probability that bad controls are stripped out with reduced bias. When environments are sufficiently diverse, the learned representation is equivalent to a valid adjustment set.

VII. CONCLUSION

To summarize the overall effect, we observed that fitting a single causal model to the whole data (not federated) perform well in overall balancing of covariates with some worse cases of severe imbalance in some clients. On the other

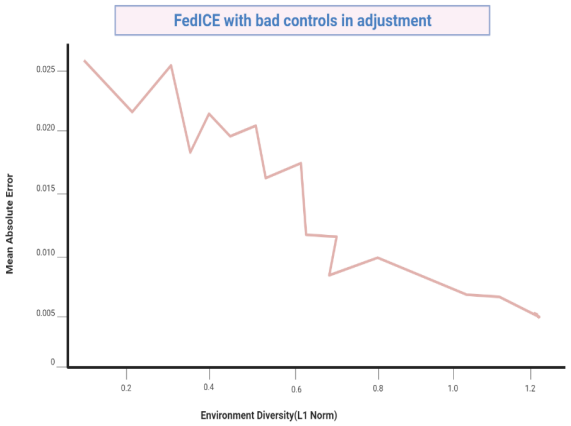

Fig. 3. FedICE mitigates bad controls more with access to more diverse environments just as in (15)

hand, in a federated setting with a separate causal model per client, we get much better client subgroup balance. But in this case, we observed larger variance in the subsequent treatment effects estimates which can be owed to the smaller sample size per client. Anaysis of FedICE performance on non identical client/server domains is ongoing work. We look forward to simulating more data with features drawn from different distributions to analyze the generalization property of the proposed approach in non linear settings. However, domain invariance of IRM in treatment effect estimation has already been validated in prior work(15). We hope that this work provides the base for much needed further research on adapting causal effect estimation methods as well as invariant learning techniques in a federated setting in observational studies to enhance user data privacy and robustness which is of utmost importance when it comes to working with sensitive health datasets.

REFERENCES

[1] Martin ARJOVSKY, Léon BOTTOU, Ishaan GULRAJANI et David Lopez PAZ : Invariant risk minimization. *arXiv:1907.02893*, 2019.

[2] J Liaw BROOKS-GUNN et Klebanov P : Effects of early intervention on cognitive function of lowbirth weight preterm infants. 1991.

[3] Sreya FRANCIS, Irene TENISON et Irina RISH : Towards causal federated learning for enhanced robustness and privacy, 2021.

[4] Jennifer Lynn HILL : Bayesian nonparametric modeling for causal inference. 2011.

[5] Peter KAIROUZ, Brendan MCMAHAN, Brendan AVENT, Aurélien BELLET, Mehdi BENNIS, Arjun Nitin BHAGOJI, Keith BONAWITZ, Zachary CHARLES, Graham CORMODE et Rachel CUMMINGS : Advances and open problems in federated learning. *arxiv*, 2019.

[6] Hyesung KIM, Jihong PARK, Mehdi BENNIS et Seong-Lyun KIM : On-device federated learning via blockchain and its latency analysis. *arXiv preprint arXiv:1808.03949*, 2018.

[7] Diederik P. KINGMA et Jimmy BA : Adam: A method for stochastic optimization, 2017.

[8] Jakub KONEČNÝ, H Brendan MCMAHAN, Felix X YU, Peter RICHTÁRIK, Ananda Theertha SURESH et Dave BACON : Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[9] Sören R KÜNZEL, Jasjeet S SEKHON, Peter J BICKEL et Bin YU : Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 2019.

[10] Brendan MCMAHAN et Daniel RAMAGE : Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 2017.

[11] Judea PEARL : Causality: Models, reasoning and inference, 2000.

[12] PAUL R. ROSENBAUM et DONALD B. RUBIN : The central role of the propensity score in observational studies for causal effects. *In The central role of the propensity score in observational studies for causal effects*. Biometrika, 1983.

[13] Donald B. RUBIN : Causal inference using potential outcomes: Design, modeling, decisions. *In Causal Inference Using Potential Outcomes: Design, Modeling, Decisions*. American Statistical Association Journal of the American Statistical Association, 2005.

[14] Uri SHALIT, Fredrik D. JOHANSSON et David SONTAG : Estimating individual treatment effect: generalization bounds and algorithms, 2017.

[15] Claudia SHI, Victor VEITCH et David BLEI : Invariant representation learning for treatment effect estimation. 2021.