Data Mining and Knowledge Discovery (2021) 35:1713–1738
https://doi.org/10.1007/s10618-021-00759-3

Adversarial balancing-based representation learning for causal effect inference with observational data

Xin Du[1] · Lei Sun[2] · Wouter Duivesteijn[1] · Alexander Nikolaev[2] · Mykola Pechenizkiy[1]

Received: 4 November 2019 / Accepted: 23 April 2021 / Published online: 17 May 2021
© The Author(s) 2021

Abstract

Learning causal effects from observational data greatly benefits a variety of domains such as health care, education, and sociology. For instance, one could estimate the impact of a new drug on specific individuals to assist clinical planning and improve the survival rate. In this paper, we focus on studying the problem of estimating the Conditional Average Treatment Effect (CATE) from observational data. The challenges for this problem are two-fold: on the one hand, we have to derive a causal estimator to estimate the causal quantity from observational data, in the presence of confounding bias; on the other hand, we have to deal with the identification of the CATE when the distributions of covariates over the treatment group units and the control units are imbalanced. To overcome these challenges, we propose a neural network framework called Adversarial Balancing-based representation learning for Causal Effect Inference (ABCEI), based on recent advances in representation learning. To ensure the identification of the CATE, ABCEI uses adversarial learning to balance the distributions of covariates in the treatment and the control group in the latent representation space, without any assumptions on the form of the treatment selection/assignment function.

Responsible editor: Panos Papapetrou and Myra Spiliopoulou.

✉ Xin Du
x.du@tue.nl

Lei Sun
leisun@buffalo.edu

Wouter Duivesteijn
w.duivesteijn@tue.nl

Alexander Nikolaev
anikolae@buffalo.edu

Mykola Pechenizkiy
m.pechenizkiy@tue.nl

[1] Eindhoven University of Technology, Eindhoven, The Netherlands

[2] University at Buffalo, Buffalo, NY, USA

In addition, during the representation learning and balancing process, highly predictive information from the original covariate space might be lost. ABCEI can tackle this information loss problem by preserving useful information for predicting causal effects under the regularization of a mutual information estimator. The experimental results show that ABCEI is robust against treatment selection bias, and matches/outperforms the state-of-the-art approaches. Our experiments show promising results on several datasets, encompassing several health care (and other) domains.

Keywords Causal inference · Observational study General Adversarial Networks · Representation learning

1 Introduction

Many domains of science require inference of causal effects, including healthcare (Casucci et al. 2017, 2019), economics and marketing (LaLonde 1986; Smith and Todd 2005), sociology (Morgan and Harding 2006), and education (Zhao and Heffernan 2017). For instance, medical scientists must know whether a new medicine benefits patients; teachers want to know whether their teaching plan significantly improves the grades of students; economists need to evaluate whether a policy can improve unemployment rates. Due to the broad application of machine learning models in these domains, properly estimating causal effects is an important task for machine learning research.

The classical method to estimate causal effects is Randomized Controlled Trials (RCTs) (Autier and Gandini 2007), where one must maintain two statistically identical groups, randomly assign treatments to each individual, and observe the outcomes. However, RCTs can be time-consuming, expensive, or unethical (e.g., for studying the effect of smoking on health). Hence, causal effect inference through observational studies is needed (Benson and Hartz 2000). The core issue of causal effect inference from observational data is the identification problem. That is: given a set of assumptions and non-experimental data, is it possible to derive a model that can correctly estimate the strength of a causal effect by certain quantities?

In this paper, our aim is to build a machine learning model that is able to estimate the Conditional Average Treatment Effect (CATE) (Abrevaya et al. 2015) from observational data. There are several challenges for this task. First, there might be spurious associations between the treatments and outcomes caused by confounding variables: variables that affect both treatment variables and the outcome variables. For example, patients with more personal wealth are in a better position to get new medicines, and at the same time their wealth increases the likelihood that they can survive. Due to the existence of confounding bias, it is nearly impossible to build an estimator by directly modeling the relations between treatments and outcomes. Strong ignorability in Rubin's Potential Outcome framework (Rubin 2005) provides a way to estimate the causal quantities. In order to satisfy ignorability in practical studies, people derive methods to match or balance the covariates using optimization techniques (Diamond and Sekhon 2013; Tam Cho et al. 2013; Zubizarreta 2012), e.g., based on mutual information between treatment variables and covariates (Sun and Nikolaev 2016), or based on propensity scores (Dehejia and Wahba 2002). However, these methods are only feasible for the estimation of Average Treatment Effect (ATE), or Average Treatment effect on the Treated (ATT), respectively. Pearl (2009) proposes a criterion based on graphical models to select admissible covariates for ignorability. Throughout this paper, we assume that all the variables in the causal system can be observed and measured, so that the causal effects we are interested in are identifiable from the observational data. This assumption allows us to build causal quantity estimators for each outcome system conditioning on the covariates.

Another challenge for CATE estimation is that in an observational study we can only observe the factual outcomes; the counterfactual outcomes can never be observed. In the presence of treatment selection bias, the imbalanced distributions of covariates in the treatment and the control groups would lead to bias in the estimation of the CATE due to generalization errors (Swaminathan and Joachims 2015). Several studies proposed various techniques to tackle this problem. Yao et al. (2018) propose to use hard samples to preserve local similarity information, which can be ported from covariate space to latent representation space. The hard sample mining process is highly dependent on the propensity score model, which is not robust when the propensity score model is misspecified. Imai and Ratkovic (2014) and Ning et al. (2020) propose estimators which are robust even when the propensity score model is not correctly specified. Kallus (2018, 2020) and Ozery-Flato et al. (2018) propose to generate balanced weights for data samples to minimize a selected imbalance measure in covariate space. Shalit et al. (2017) propose to derive upper bounds on the estimation error by considering both covariate balancing and potential outcomes. Highly predictive information might be lost in the reweighing or balancing processes of these methods.

To address these problems, we propose a framework (cf. Fig. 1), which generates balanced representations and preserves highly predictive information in the latent space without using propensity scores. We design a two-player adversarial game, between an encoder that transforms covariates to latent representations and a discriminator which distinguishes representations from the control and treatment groups. Unlike in the classical GAN framework, the ‘true distribution’ (latent representations of the control group[1]) in this game must also be generated by the encoder. To prevent losing useful information during the balancing process, we use a mutual information estimator to constrain the encoder to preserve highly predictive information (Hjelm et al. 2019). The outcome data are also considered in this unified framework to specify the causal effect predictor.

Technically, the unified framework encodes the input covariates into a latent representation space, and builds estimators to estimate the treatment outcomes with those representations. There are three components on top of the encoder in our model:

Mutual information estimation an estimator is specified to estimate and maximize the mutual information between representations and covariates;

Adversarial balancing the encoder plays an adversarial game with a discriminator, trying to fool the discriminator by minimizing the discrepancies between distributions of representations from the treatment and the control group;

[1] our method supports representations of either treatment/control group or both as the ‘true distribution’.

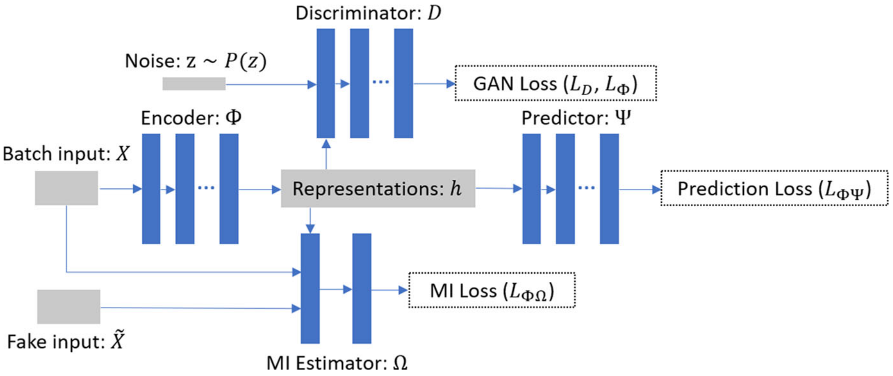


Fig. 1 Deep neural network architecture of ABCEI for causal effect inference

Treatment outcome prediction a predictor over the latent space is employed to estimate the treatment outcomes.

By jointly optimizing the three components via backpropagation, we can get a robust estimator for the CATE. The overarching architecture of our framework is shown in Fig. 1. As a summary, our main contributions are:

1. We propose a novel model: Adversarial Balancing-based representation learning for Causal Effect Inference (ABCEI) with observational data. ABCEI addresses information loss and treatment selection bias by learning highly informative and balanced representations in a latent space.
2. A neural network encoder is constrained by a mutual information estimator to minimize the information loss between representations and input covariates, which preserves highly predictive information for causal effect inference.
3. We employ an adversarial learning method to balance representations between the treatment and the control groups, which deals with the treatment selection bias problem without any assumption on the form of the treatment selection function, unlike, e.g., the propensity score method.
4. We conduct various experiments on synthetic and real-world datasets. ABCEI outperforms most of the state-of-the-art methods on benchmark datasets. We show that ABCEI is robust against various experimental settings. By supporting mini-batch, ABCEI can be applied on large-scale datasets.

2 Problem setup

Assume an observational dataset $\{X, T, Y\}$, with covariate matrix $X \in \mathbb{R}^{n \times k}$, binary treatment vector $T \in \{0, 1\}^n$, and treatment outcome vector $Y \in \mathbb{R}^n$. Here, n denotes the number of observed units, and k denotes the number of covariates in the dataset. For each unit u, we have k covariates $x_1, \ldots, x_k$, as well as one observable outcome y corresponding to one specified value of treatment variable $t \in \{0, 1\}$. According to the Rubin-Neyman causal model (Rubin 2005), two potential outcomes y_0, y_1 exist for treatments $\{0, 1\}$, respectively. We call y_t the *factual outcome*, denoted by y_f, and

y_{1-t} the *counterfactual outcome*, denoted by y_{cf}. Assuming there is a joint distribution $P(x, t, y_0, y_1)$, we make the following assumptions:

Assumption 1 *(Strong Ignorability)* Conditioning on x, the potential outcomes y_0, y_1 are independent of t, which can be stated as: $(y_0, y_1) \perp\!\!\!\perp t|x$.

Assumption 2 *(No Interference)* The treatment outcome of each individual is not affected by the treatment assignment of other units, which can be formulated as: $Y^u(t^1, \cdots, t^n) = Y^u(t^u)$.

Assumption 3 *(Consistency)* The potential outcome y_t of each individual is equal to the observed outcome y, if the actual treatment received is $T = t$, which can be represented as: $y = y_t$, if $T = t, \forall t$.

Assumption 4 *(Positivity)* For all sets of covariates and for all treatments, the probability of treatment assignment will always be strictly larger than 0 and strictly smaller than 1, which can be expressed as: $0 < P(t|x) < 1, \forall t$ and $\forall x$.

Assumption 1 indicates that all the confounders are observed, i.e., *no unmeasured confounder is present*. This is a restrictive but much used assumption in a large subset of causal inference literature (Rosenbaum and Rubin 1983). Hence, by controlling on X, we can remove the confounding bias. Assumption 4 allows us to estimate the CATE for any x in the covariate space. Under these assumptions, we can formalize the definition of the CATE for our task:

Definition 1 The Conditional Average Treatment Effect (CATE) for unit u is: $CATE(u) := \mathbb{E}\left[\, y_1 \mid x^u \,\right] - \mathbb{E}\left[\, y_0 \mid x^u \,\right]$.

We can now define the Average Treatment Effect (ATE) and the Average Treatment effect on the Treated (ATT) as:

$$ATE := \mathbb{E}[\, CATE(u) \,], \qquad ATT := \mathbb{E}[\, CATE(u) \mid t = 1 \,].$$

Because the joint distribution $P(x, t, y_0, y_1)$ is unknown, we can only estimate $CATE(u)$ from observational data. A function over the covariate space $\mathcal{X}$ can be defined as $f : \mathcal{X} \times \{0, 1\} \rightarrow \mathcal{Y}$. The estimate of $CATE(u)$ can now be defined:

Definition 2 Given an observational dataset $\{X, T, Y\}$ and a function f, for unit u, the estimate of $CATE(u)$ is:

$$\widehat{CATE}(u) = f(x^u, 1) - f(x^u, 0).$$

In order to accomplish the task of CATE estimation, we need to find an optimal function over the covariate space for both systems (observable populations with $t = 1$ and $t = 0$, respectively).

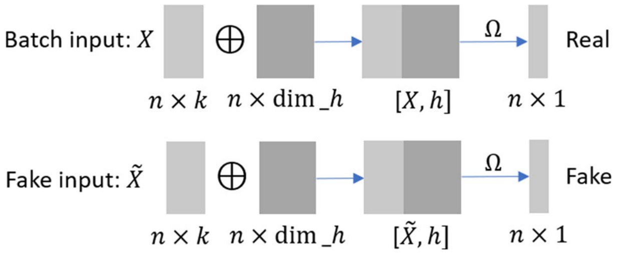


Fig. 2 Mutual information estimator between covariates and latent representations

3 Proposed method

In order to overcome the challenges in CATE estimation, we build our model on recent advances in representation learning. We propose to define function $\Phi : \mathcal{X} \to \mathcal{H}$ and function $\Psi : \mathcal{H} \to \mathcal{Y}$ so that we have $\widehat{Y_T} = f(X, T) = \Psi(\Phi(X), T) = \Psi(h, T)$. Instead of directly estimating the treatment outcome conditioned on covariates, we propose to use an encoder to learn latent representations of covariates. Thus, we simultaneously learn latent representations and estimate the treatment outcome. However, the function f could suffer from information loss and treatment selection bias, unless we constrain the encoder Φ to learn balanced representations while preserving useful information.

3.1 Mutual information estimation

Consider the information loss when transforming covariates into a latent space. The non-linear statistical dependencies between variables can be acquired by mutual information (MI) (Shannon 1948). Thus we use MI between latent representations and original covariates as a measure to account for information loss:

$$I(X; h) = \int_{\mathcal{X}} \int_{\mathcal{H}} P(x, h) \log\left(\frac{P(x, h)}{P(x)P(h)}\right) \mathrm{d}h \, \mathrm{d}x.$$

We denote the joint distribution between covariates and representations by $\mathbb{P}_{Xh}$ and the product of marginals by $\mathbb{P}_X \otimes \mathbb{P}_h$. Note that, consistent with Shannon's information theory, MI can be represented as the Kullback-Leibler (KL) divergence:

$$I(X; h) := H(X) - H(X|h) := D_{KL}(\mathbb{P}_{Xh} || \mathbb{P}_X \otimes \mathbb{P}_h).$$

It is hard to compute MI in continuous and high-dimensional spaces, but one can capture a lower bound of MI with the Donsker-Varadhan representation of KL-divergence (Donsker and Varadhan 1983):

Theorem 1 *(Donsker-Varadhan)*

$$D_{KL}(\mathbb{P}_{Xh} || \mathbb{P}_X \otimes \mathbb{P}_h) = \sup_{\Omega \in \mathcal{C}} \mathbb{E}_{\mathbb{P}_{Xh}}[\Omega(x, h)] - \log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_h}\left[e^{\Omega(x,h)}\right].$$

Here, $\mathcal{C}$ denotes the set of unconstrained functions Ω.

Proof Given a fixed function Ω, we can define distribution G as:

$$\mathrm{d}G = \frac{e^{\Omega(Z)}\mathrm{d}Q}{\int_{\mathcal{Z}} e^{\Omega(Z)}\mathrm{d}Q}$$

Equivalently, we have:

$$\mathrm{d}G = e^{(\Omega(Z)-S)}\mathrm{d}Q, \qquad S = \log \mathbb{E}_Q\left[e^{\Omega(Z)}\right].$$

Then by construction, we have:

$$\begin{aligned}\mathbb{E}_P[\Omega(Z)] - \log \mathbb{E}_Q\left[e^{\Omega(Z)}\right] &= \mathbb{E}_P[\Omega(Z)] - S\\ &= \mathbb{E}_P\left[\log \frac{\mathrm{d}G}{\mathrm{d}Q}\right]\\ &= \mathbb{E}_P\left[\log \frac{\mathrm{d}P\mathrm{d}G}{\mathrm{d}Q\mathrm{d}P}\right]\\ &= \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q} - \log \frac{\mathrm{d}P}{\mathrm{d}G}\right]\\ &= D_{KL}(P||Q) - D_{KL}(P||G)\\ &\leq D_{KL}(P||Q).\end{aligned}$$

When distribution G is equal to P, this bound is tight. □

Inspired by the Mutual Information Neural Estimation (MINE) idea (Belghazi et al. 2018), we propose to establish a neural network estimator for MI. Specifically, let Ω be a function $\mathcal{X} \times \mathcal{H} \rightarrow \mathbb{R}$ parameterized by a deep neural network. We have:

$$\begin{aligned}I(X;h) &:= D_{KL}\left(\mathbb{P}_{Xh}||\mathbb{P}_X \otimes \mathbb{P}_h\right)\\ &\geq \hat{I}_\Omega(X;h)\\ &:= \mathbb{E}_{\mathbb{P}_{Xh}}[\Omega(x,h)] - \log \mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_h}\left[e^{\Omega(x,h)}\right].\end{aligned} \tag{1}$$

By distinguishing the joint distribution and the product of marginals, the estimator Ω approximates the MI with arbitrary precision. In practice, as shown in Fig. 2, we concatenate the input covariates X with representations h one by one to create positive samples (as samples from the true joint distribution). Then, we randomly shuffle X on the batch axis to create fake input covariates $\tilde{X}$. Representations h are concatenated with fake input $\tilde{X}$ to create negative samples (as samples from the product of marginals). From Eq. (1) we can derive the loss function for the MI estimator:

$$L_{\Phi\Omega} = -\mathbb{E}_{x\sim X}\left[\Omega\left(x,h\right)\right] + \log \mathbb{E}_{x\sim\tilde{X}}\left[e^{\Omega(x,h)}\right].$$

Information loss can be decreased by simultaneously optimizing the encoder Φ and the MI estimator Ω to minimize $L_{\Phi\Omega}$ iteratively via gradient descent.

3.2 Adversarial balancing

The representations of the treatment and the control groups are denoted by $h(t=1)$ and $h(t=0)$, respectively. The discrepancy between the covariate distributions within the treatment and the control groups is the issue to be addressed. To decrease this discrepancy, we propose an adversarial learning method to constrain the encoder to learn treatment and control representations that are balanced distributions. We build an adversarial game between a discriminator D and the encoder Φ, in line with the framework of Generative Adversarial Networks (GAN) (Goodfellow et al. 2014). In the classical GAN framework, a source of noise is mapped to a generated image by a generator. A discriminator is trained to distinguish whether an input sample is from the true or the synthetic image distribution generated by the generator. The logic of GANs lies in training a reliable discriminator to distinguish fake and real images, and then, using the discriminator to train the generator, which in turn generates images constructed so as to try to fool the discriminator.

In our adversarial game:

1. We draw a noise vector $z \sim P(z)$ which has the same length as the latent representations, where $P(z)$ can be a spherical Gaussian distribution or a Uniform distribution;
2. We separate representation by treatment assignment, and form two distributions: $P_{h(t=1)}$ and $P_{h(t=0)}$;
3. We train a discriminator D to distinguish concatenated vectors from the treatment and the control group ($[z, h(t=1)]$ and $[z, h(t=0)]$);
4. We optimize the encoder Φ to generate balanced representations to fool the discriminator.

According to the architecture of ABCEI, the encoder is associated with the MI estimator Ω, treatment outcome predictor Ψ, and adversarial discriminator D. This means that the training process is iteratively adjusting each of the components. The instability of GAN training will become serious in this context. To stabilize the GAN training, we propose to use the framework of Wasserstein GAN with gradient penalty (Gulrajani et al. 2017). By removing the sigmoid layer and applying the gradient penalty to the data between the distributions of the treatment and the control groups, we can find a function D which satisfies the 1-Lipschitz inequality:

$$\left|\left|D\left(x^1\right) - D\left(x^2\right)\right|\right| \leq \left|\left|x^1 - x^2\right|\right|.$$

We can write down the form of our adversarial game:

$$\min_{\Phi} \max_{D} \mathbb{E}_{h \sim P_{h(t=0)}}[D([z,h])] - \mathbb{E}_{h \sim P_{h(t=1)}}[D([z,h])] - \\ \beta\, \mathbb{E}_{h \sim P_{\text{penalty}}}\left[(||\nabla_{[z,h]} D([z,h])||_2 - 1)^2\right],$$

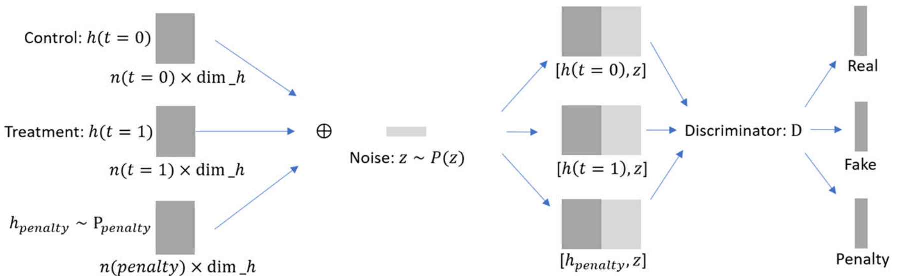


Fig. 3 Adversarial learning structure for representation balancing

where P_{penalty} is the distribution acquired by uniformly sampling along the straight lines between pairs of samples from $P_{h(t=0)}$ and $P_{h(t=1)}$. The adversarial learning process is depicted in Fig. 3.

The encoder Φ will now be smoothly trained to generate balanced representations. We can write down the training objectives for discriminator D and encoder Φ, respectively:

$$\begin{aligned} L_D &= -\mathbb{E}_{h \sim P_{h(t=0)}}[D([z,h])] + \mathbb{E}_{h \sim P_{h(t=1)}}[D([z,h])] \\ &\quad + \beta\, \mathbb{E}_{h \sim P_{\text{penalty}}}\left[(||\nabla_{[z,h]} D([z,h])||_2 - 1)^2\right], \\ L_\Phi &= \mathbb{E}_{h \sim P_{h(t=0)}}[D([z,h])] - \mathbb{E}_{h \sim P_{h(t=1)}}[D([z,h])]. \end{aligned}$$

3.3 Treatment outcome prediction

The final step for CATE estimation is to predict the treatment outcomes with learned representations. We establish a neural network predictor, which takes latent representations and treatment assignments of units as the input, to conduct outcome prediction: $\widehat{y_t} = \Psi(h, t)$. We can write down the loss function of the training objective as:

$$L_{\Phi\Psi} = \mathbb{E}_{(h,t,y_t) \sim \{h,T,Y_T\}}\left[(\Psi\,(h,t) - y_t)^2\right] + \lambda\, R(\Psi).$$

Here, R is a regularization on Ψ for the model complexity.

3.4 Learning optimization

With respect to the architecture in Fig. 1, we minimize $L_{\Phi\Omega}$, L_Φ, and $L_{\Phi\Psi}$, respectively, to iteratively optimize parameters in the global model. The optimization steps are handled with the stochastic method due to Adam (Kingma and Ba 2015), and the training of the model is done per Algorithm 1.

Algorithm 1 ABCEI

```
Input: Observational dataset $\{X, T, Y\}$; loss function $L_{\Phi\Omega}$, $L_{\Phi}$ and $L_{\Phi\Psi}$, $L_D$; Neural Networks $\Phi$, $\Omega$, $D$, $\Psi$; parameters $\Theta_{\Phi}$, $\Theta_{\Omega}$, $\Theta_D$, $\Theta_{\Psi}$
repeat
    Draw mini-batch $\{X_b, T_b, Y_b\} \subset \{X, T, Y\}$
    Compute representations $h = \Phi(X_b)$
    Draw fake input $\tilde{X}_b \sim \tilde{\mathbb{P}}$
    Draw noise $z \sim \mathcal{N}(0, I)$
    Set $\Theta_{\Phi}, \Theta_{\Omega} \leftarrow \text{Adam}(L_{\Phi\Omega}(X_b, \tilde{X}_b, h), \Theta_{\Phi}, \Theta_{\Omega})$
    for $i = 1$ to 3 do
        Set $\Theta_D \leftarrow \text{Adam}(L_D(h, z, T_b), \Theta_D)$
    Set $\Theta_{\Phi} \leftarrow \text{Adam}(L_{\Phi}(h, z, T_b), \Theta_{\Phi})$
    Set $\Theta_{\Phi}, \Theta_{\Psi} \leftarrow \text{Adam}(L_{\Phi\Psi}(h, T_b, Y_b), \Theta_{\Phi}, \Theta_{\Psi})$
until convergence
```

4 Experiments

Due to the lack of counterfactual treatment outcomes in observational data, it is difficult to validate and test the performance of causal effect inference methods. In this paper, we adopt two ways to construct datasets for validating and testing the performance of causal inference methods: the one is to use simulated or semi-simulated treatment outcomes, in particular based on dataset IHDP (Hill 2011); the other is to use RCT datasets and add a non-randomized component to generate imbalanced datasets, in particular based on dataset Jobs (LaLonde 1986; Smith and Todd 2005). We employ five benchmark datasets: IHDP, Jobs, Twins (Louizos et al. 2017), ACIC (Dorie et al. 2019) and MIMIC-III (Johnson et al. 2016, 2019). For IHDP, Jobs, Twins, ACIC, and MIMIC-III, the experimental results are averaged over 1000, 100, 100, 7700, 100 train/validation/test sets, respectively, with split sizes 60%/30%/10%. The implementation of our method is based on Python and Tensorflow (Abadi et al.2016). All the experiments in this paper are conducted on a cluster with 1x Intel Xeon E5 2.2GHz CPU, 4x Nvidia Tesla V100 GPU and 256GB RAM. The source code of our algorithms is available on GitHub.[2]

4.1 Details of datasets

Metadata on the employed datasets can be found in Table 1.

ACIC (Dorie et al. 2019) The *Atlantic Causal Inference Conference* (ACIC) dataset is derived from real-world data with 4802 observations with 58 covariates. There are 77 datasets which are simulated with different treatment selection and outcome functions. Each dataset is generated in 100 random replications independently. In this benchmark, different settings like degrees of non-linearity, treatment selection bias and magnitude of treatment outcome are considered.

[2] https://github.com/octeufer/Adversarial-Balancing-based-representation-learning-for-Causal-Effect-Inference.

Table 1 Metadata on employed datasets

Dataset	Observations	Control/treatment	Covariates	Reference
ACIC	4802	–/–	58	Dorie et al. (2019)
IHDP	747	608/139	25	Hill (2011)
Twins	25656	12828/12828	43	Louizos et al. (2017)
MIMIC-III	7413	–/–	25	Johnson et al. (2016)
Jobs	3122	2825/297	7	LaLonde (1986)

On the ACIC dataset, the number of control and treatment units varies across replications; on the MIMIC-III dataset, the numbers of control and treatment units are simulated. Both procedures can be found in the main text, in the respective paragraphs of Sect. 4.1 where the corresponding datasets are introduced. Note that control units pool in the Jobs dataset consists of two components (cf. Jobs paragraph of Sect. 4.1)

IHDP (Hill 2011) The *Infant Health and Development Program* (IHDP) studies the impact of specialist home visits on future cognitive test scores. Covariates in the semi-simulated dataset are collected from a real-world randomized experiment. The treatment selection bias is created by removing a subset of the treatment group. We use the setting ‘A’ in (Dorie 2016) to simulate treatment outcomes. This dataset includes 747 units (608 control and 139 treated) with 25 covariates associated with each unit.

Twins (Louizos et al. 2017) The *Twins* dataset is created based on the “Linked Birth / Infant Death Cohort Data” by NBER.[3] Inspired by Almond et al. (2005), we employ a matching algorithm to select twin births in the USA between 1989 and 1991. By doing this, we get units associated with 43 covariates including education, age, race of parents, birth place, marital status of mother, the month in which pregnancy prenatal care began, total number of prenatal visits, and other variables indicating demographic and health conditions. We only select twins that have the same gender who both weigh less than $2000g$. For the treatment variable, we use $t = 0$ indicating the lighter twin and $t = 1$ indicating the heavier twin. We take the mortality of each twin in their first year of life as the treatment outcome, inspired by Louizos et al. (2017). Finally, we have a dataset consisting of 12,828 pairs of twins whose mortality rate is 19.02% for the lighter twin and 16.54% for the heavier twin. Hence, we have observational treatment outcomes for both treatments. In order to simulate the selection bias, we selectively choose one of the twins to observe with regard to the covariates associated with each unit as follows: $t|x \sim \text{Bernoulli}(\sigma(w^T x + n))$, where $w^T \sim \mathcal{N}(0, 0.1 \cdot I)$ and $n \sim \mathcal{N}(1, 0.1)$.

MIMIC-III (Johnson et al. 2016, 2019) This benchmark is created based on *MIMIC-III*, a database comprised of de-identified profile and health outcome data for critical care unit patients. We select patient samples with their demographic information as well as various observed laboratory measurements by chemistry or hematology. After filtering samples with missing values, the benchmark consists of 7413 samples with 25 covariates. We investigate the effect of prescription amount in the first day of critical care unit on the length of stay in the ICU: for the binary treatment, we let 0 represent a small prescription amount and 1 a large prescription amount. The treatment outcomes are simulated by $y|x, t \sim (w^T x + \beta t + n)$, where $n \sim \mathcal{N}(0, 1)$, $w \sim \mathcal{N}(0^{25}, 0.5 \cdot$

[3] https://nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html.

$(\Sigma + \Sigma^T))$, and $\Sigma \sim \mathcal{U}((-1, 1)^{25\times 25})$. The treatment assignments are simulated as $t|x \sim Bernoulli(\sigma(s^T x + m))$, where $m \sim \mathcal{N}(0, 0.1)$ and $s \sim \mathcal{N}(0^{25}, 0.1 \cdot I)$.

Jobs (LaLonde 1986; Smith and Todd 2005) The *Jobs* dataset studies the effect of job training on employment status. It consists of a non-randomized component from observational studies and a randomized component based on the National Supported Work program. The randomized component includes 722 units (425 control and 297 treated) with seven covariates, and the non-randomized component (PSID comparison group) includes 2490 control units.

4.2 Evaluation metrics

Since the ground truth CATE for the IHDP dataset and MIMIC-III benchmark is known, we can employ Precision in Estimation of Heterogeneous Effect (PEHE) (Hill 2011), as the evaluation metric of CATE estimation:

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{u=1}^{n}((\mathbb{E}[y_1|x^u] - \mathbb{E}[y_0|x^u]) - (f(x^u, 1) - f(x^u, 0)))^2.$$

Subsequently, we can evaluate the precision of ATE estimation based on the estimated CATE.

For the Jobs dataset, because we only know parts of the ground truth (the randomized component), we cannot evaluate the performance of ATE estimation. Following Shalit et al. (2017), we evaluate the precision of ATT estimation and policy risk estimation, where

$$R_{pol}(\pi) = 1 - \left[\mathbb{E}\left(y_1\middle|\pi\left(x^u\right) = 1\right) \cdot P(\pi = 1) + \mathbb{E}\left(y_0\middle|\pi\left(x^u\right) = 0\right) \cdot P(\pi = 0)\right].$$

In this paper, we consider $\pi(x^u) = 1$ when $f(x^u, 1) - f(x^u, 0) > 0$.

For the Twins dataset, because we only know the observed treatment outcome for each unit, we follow Louizos et al. (2017) in using the Area Under the ROC Curve (AUC) as the evaluation metric. For the ACIC dataset, we follow Ozery-Flato et al. (2018) in using the RMSE ATE as performance metric.

4.3 Baseline methods

We perform the comparisons with the following baselines: least square regression using treatment as a feature (OLS/LR_1); separate least square regressions for each treatment (OLS/LR_2); balancing linear regression (BLR) and balancing neural network (BNN) (Johansson et al. 2016); k-nearest neighbor (k-NN) as suggested by Crump et al. (2008); Bayesian additive regression trees (BART) (Sparapani et al. 2016); random forests (RF) (Breiman 2001); causal forests (CF) (Wager and Athey 2018); treatment-agnostic representation networks (TARNet) and counterfactual regression with Wasserstein distance (CFR-Wass) (Shalit et al. 2017); causal effect variational autoencoders (CEVAE) (Louizos et al. 2017); local similarity preserved individual treatment effect (SITE) (Yao et al. 2018); MMD measure using RBF kernel

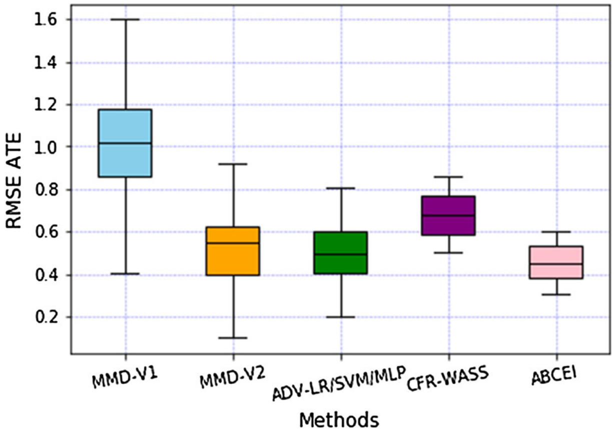


Fig. 4 Boxplots per method of the Root Mean Square Error in the Average Treatment Effect on the ACIC dataset (lower = better)

(MMD-V1, MMD-V2) (Kallus 2020, 2018); and adversarial balancing with cross-validation procedure (ADV-LR/SVM/MLP) (Ozery-Flato et al. 2018). We show a quantitative comparison between our method and the state-of-the-art baselines. In the comparison, we include two variants of ABCEI: by ABCEI* we denote ABCEI without the mutual information estimation component, and by ABCEI** we denote ABCEI without the adversarial learning component. All baseline methods are parameterized according to the recommended settings in the original papers.

4.4 Results

Figure 4 displays the relative performance of ABCEI and recent balancing methods, on the ACIC benchmark dataset. As we can see, representation learning methods display a lower variance than methods based on reweighing samples on covariate space. Also, adversarial balancing methods have a lower (= better) mean RMSE in ATE estimation. ABCEI combines the benefits of both: it has the advantage of adversarial balancing as well as preserving predictive information in latent space. As a consequence, it outperforms counterfactual regression with Wasserstein distance and MMD-V1, and compared to the other baselines, performs similarly in mean but better in variance.

Experimental results on the other four datasets are shown in Tables 2, 3, 4, and 5. It would be unsound to report aggregated statistical test results over the results reported in these tables. Due to varying (un-)availability of ground truth, we must resort to reporting varying evaluation measures per dataset. It would not be appropriate to aggregate over these measures in a single statistical hypothesis test. However, one can see that ABCEI performs best in fourteen out of sixteen cases, not only by the best number in the column, but often also by a non-overlapping empirical confidence interval ($\mu \pm \sigma$, so 68% confidence intervals) with that of the best competitor (cf. reported standard deviations). This provides evidence that ABCEI is a substantial improvement over the state of the art.

Table 2 In-sample and out-of-sample results with mean and standard errors on the IHDP dataset (lower = better)

Methods	In-sample		Out-sample	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
OLS/LR_1	5.8 ± .3	.73 ± .04	5.8 ± .3	.94 ± .06
OLS/LR_2	2.4 ± .1	.14 ± .01	2.5 ± .1	.31 ± .02
BLR	5.8 ± .3	.72 ± .04	5.8 ± .3	.93 ± .05
BART	2.1 ± .1	.23 ± .01	2.3 ± .1	.34 ± .02
k-NN	2.1 ± .1	.14 ± .01	4.1 ± .2	.79 ± .05
RF	4.2 ± .2	.73 ± .05	6.6 ± .3	.96 ± .06
CF	3.8 ± .2	.18 ± .01	3.8 ± .2	.40 ± .03
BNN	2.2 ± .1	.37 ± .03	2.1 ± .1	.42 ± .03
TARNet	.88 ± .0	.26 ± .01	.95 ± .0	.28 ± .01
CFR-Wass	.71 ± .0	.25 ± .01	.76 ± .0	.27 ± .01
CEVAE	2.7 ± .1	.34 ± .01	2.6 ± .1	.46 ± .02
SITE	**.69 ± .0**	.22 ± .01	.75 ± .0	.24 ± .01
ABCEI*	.74 ± .0	.12 ± .01	.78 ± .0	.11 ± .01
ABCEI**	.81 ± .1	.18 ± .03	.89 ± .1	.16 ± .02
ABCEI	.71 ± .0	**.09 ± .01**	**.73 ± .0**	**.09 ± .01**

Additionally, we observe that the two out of the sixteen cases, in which ABCEI does not perform best, have several similar characteristics. They are based on in-sample measurements on the Jobs and IHDP datasets, and as Table 1 shows, these are the datasets with the smallest numbers of observations, the (joint) smallest numbers of covariates, and a relative imbalance between control and treatment group size. Among these five datasets, ABCEI always performs better on datasets with more observations, with more covariates, and with more balance between control and treatment group size, although the number of datasets is too small to claim the *significance* of these effects.

Due to the existence of treatment selection bias, regression based methods suffer from a high generalization error. Nearest neighbor based methods consider unit similarity to overcome selection bias, but cannot achieve balance globally. Recent advances in representation learning bring improvements in causal effect estimation. Unlike CFR-Wass, BNN, and SITE, ABCEI considers information loss and balancing problems. The mutual information estimator ensures that the encoder learns representations preserving useful information from the original covariate space. The adversarial learning component constrains the encoder to learn balanced representations. This causes ABCEI to achieve better performance than the baselines. We also report the performance of our model without mutual information estimator or adversarial learning, respectively, as ABCEI*, ABCEI**. From the results we can see that performance suffers when either of these components is left out, which demonstrates the importance of combining adversarial learning and mutual information estimation in ABCEI.

Table 3 In-sample and out-of-sample results with mean and standard errors on the Twins dataset (AUC: higher = better, ϵ_{ATE}: lower = better)

Methods	In-sample		Out-sample	
	AUC	ϵ_{ATE}	AUC	ϵ_{ATE}
OLS/LR_1	.660 ± .005	.004 ± .003	.500 ± .028	.007 ± .006
OLS/LR_2	.660 ± .004	.004 ± .003	.500 ± .016	.007 ± .006
BLR	.611 ± .009	.006 ± .004	.510 ± .018	.033 ± .009
BART	.506 ± .014	.121 ± .024	.500 ± .011	.127 ± .024
k-NN	.609 ± .010	.003 ± .002	.492 ± .012	.005 ± .004
BNN	.690 ± .008	.006 ± .003	.676 ± .008	.020 ± .007
TARNet	.849 ± .002	.011 ± .002	.840 ± .006	.015 ± .002
CFR-Wass	.850 ± .002	.011 ± .002	.842 ± .005	.028 ± .003
CEVAE	.845 ± .003	.022 ± .002	.841 ± .004	.032 ± .003
SITE	.862 ± .002	.016 ± .001	.853 ± .006	.020 ± .002
ABCEI*	.861 ± .001	.005 ± .001	.851 ± .001	.006 ± .001
ABCEI**	.855 ± .001	.005 ± .001	.849 ± .001	.006 ± .001
ABCEI	**.871 ± .001**	**.003 ± .001**	**.863 ± .001**	**.005 ± .001**

Table 4 In-sample and out-of-sample results with mean and standard errors on the MIMIC-III benchmark (lower = better)

Methods	In-sample		Out-sample	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
OLS/LR_1	7.1 ± .2	.92 ± .15	8.2 ± .2	.97 ± .15
OLS/LR_2	2.7 ± .1	.24 ± .11	3.3 ± .2	.29 ± .13
BLR	7.3 ± .1	.90 ± .09	8.5 ± .3	.97 ± .09
BART	2.4 ± .2	.31 ± .09	3.1 ± .2	.37 ± .12
k-NN	2.8 ± .1	.32 ± .11	3.6 ± .1	.36 ± .11
RF	4.6 ± .3	.88 ± .10	5.3 ± .3	.89 ± .11
CF	4.1 ± .1	.22 ± .13	4.9 ± .1	.24 ± .14
BNN	2.5 ± .1	.45 ± .11	3.3 ± .1	.49 ± .11
TARNet	1.91 ± .0	.25 ± .16	2.11 ± .1	.31 ± .16
CFR-Wass	1.06 ± .0	.19 ± .14	1.09 ± .0	.21 ± .14
CEVAE	2.71 ± .0	.23 ± .11	2.72 ± .0	.23 ± .12
SITE	1.29 ± .0	.21 ± .14	1.35 ± .0	.25 ± .14
ABCEI*	.89 ± .0	.13 ± .13	.92 ± .0	.16 ± .14
ABCEI**	.96 ± .0	.15 ± .12	.99 ± .0	.16 ± .14
ABCEI	**.85 ± .0**	**.11 ± .12**	**.89 ± .0**	**.12 ± .14**

Table 5 In-sample and out-of-sample results with mean and standard errors on the Jobs dataset (lower = better)

Methods	In-sample		Out-sample	
	R_{pol}	ϵ_{ATT}	R_{pol}	ϵ_{ATT}
OLS/LR_1	.22 ± .0	**.01** ± **.00**	.23 ± .0	.08 ± .04
OLS/LR_2	.21 ± .0	.01 ± .01	.24 ± .0	.08 ± .03
BLR	.22 ± .0	.01 ± .01	.25 ± .0	.08 ± .03
BART	.23 ± .0	.02 ± .00	.25 ± .0	.08 ± .03
k-NN	.23 ± .0	.02 ± .01	.26 ± .0	.13 ± .05
RF	.23 ± .0	.03 ± .01	.28 ± .0	.09 ± .04
CF	.19 ± .0	.03 ± .01	.20 ± .0	.07 ± .03
BNN	.20 ± .0	.04 ± .01	.24 ± .0	.09 ± .04
TARNet	.17 ± .0	.05 ± .02	.21 ± .0	.11 ± .04
CFR-Wass	.17 ± .0	.04 ± .01	.21 ± .0	.08 ± .03
CEVAE	.15 ± .0	.02 ± .01	.26 ± .1	.03 ± .01
SITE	.17 ± .0	.04 ± .01	.21 ± .0	.09 ± .03
ABCEI*	.14 ± .0	.04 ± .01	.18 ± .0	.04 ± .01
ABCEI**	.15 ± .0	.05 ± .01	.19 ± .0	.04 ± .01
ABCEI	**.13** ± **.0**	.02 ± .01	**.17** ± **.0**	**.03** ± **.01**

4.5 Training details

We adopt ELU (Clevert et al. 2016) as the non-linear activation function if there is no specification. We employ various numbers of fully-connected hidden layers with various sizes across networks: four layers with size 200 for the encoder network; two layers with size 200 for the mutual information estimator network; three layers with size 200 for the discriminator network; and finally, three layers with size 100 for the predictor network, following the structure of TARnet (Shalit et al. 2017). The gradient penalty weight β is set to 10.0, and the regularization weight is set to 0.0001.

In the training step, firstly we minimize $L_{\Phi\Omega}$ by simultaneously optimizing Φ and Ω with one-step gradient descent. Then the representations h are passed to the discriminator to minimize L_D by optimizing D with 3-step gradient descent, in order to find a stable discriminator. Next, we use discriminator D to train encoder Φ by minimizing L_Φ with one-step gradient descent. Finally, encoder Φ and predictor Ψ are optimized simultaneously by minimizing $L_{\Phi\Psi}$.

4.6 Hyper-parameter optimization

Due to the fact that we cannot observe counterfactuals in observational datasets, standard cross-validation methods are not feasible. We follow the hyper-parameter optimization criterion in (Shalit et al. 2017), with an early stopping with regard to the lower bound on the validation set. Hyper-parameter search space details are displayed in Table 6. The optimal hyper-parameter settings for each benchmark dataset are reported in Table 7.

Table 6 Search spaces of hyper-parameters

Hyper-parameter	Range
λ	1*e*-3,1*e*-4,5*e*-5
β	1.0,5.0,10.0,15.0
Optimizer	RMSProp, Adam
Depth of encoder layers	1, 2, 3, 4, 5, 6
Depth of discriminator layers	1, 2, 3, 4, 5, 6
Depth of predictor layers	1, 2, 3, 4, 5, 6
Dimension of encoder layers	50, 100, 200, 300, 500
Dimension of discriminator layers	50, 100, 200, 300, 500
Dimension of MI estimator layers	50, 100, 200, 300, 500
Dimension of predictor layers	50, 100, 200, 300, 500
Batch size	65, 80, 100, 200, 300, 500

Table 7 Optimal hyper-parameters for each benchmark dataset

Hyper-parameters	Datasets			
	IHDP	Jobs	Twins	ACIC
λ	1*e*–4	1*e*–4	1*e*–4	1*e*–4
β	10.0	10.0	10.0	10.0
Optimizer	Adam	Adam	Adam	Adam
Depth of encoder layers	4	5	5	4
Depth of discriminator layers	3	3	3	3
Depth of predictor layers	3	3	3	3
Dimension of encoder layers	200	200	300	200
Dimension of discriminator layers	200	200	200	200
Dimension of MI estimator layers	200	200	200	200
Dimension of predictor layers	100	100	200	100
Batch size	65	100	300	100

4.7 Computational complexity

Assuming that the size of mini-batch is n, and the number of epochs is m, the computational complexity of our model is $\mathcal{O}(n \cdot m \cdot ((\Phi_h - 1)\Phi_w^2 + (\Omega_h - 1)\Omega_w^2 + (D_h - 1)D_w^2 + (\Psi_h - 1)\Psi_w^2))$. Here $\Phi_h, \Omega_h, D_h, \Psi_h$ indicate the numbers of layers, and $\Phi_w, \Omega_w, D_w, \Psi_w$ indicate the numbers of neurons in each layer, in Neural Networks Φ, Ω, D, Ψ.

4.8 Robustness analysis on treatment selection bias

To investigate the performance of our model when varying the level of treatment selection bias, we generate toy datasets by varying the discrepancy between the treatment

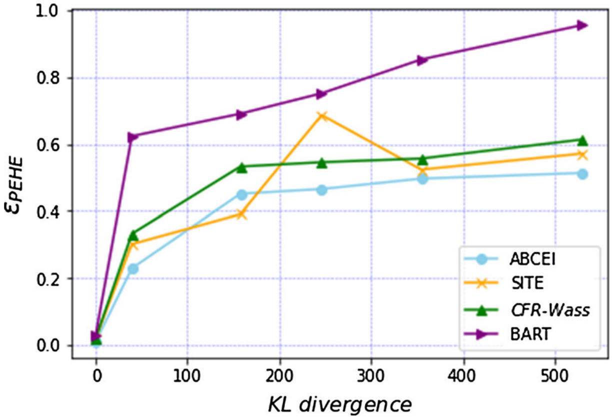


Fig. 5 ϵ_{PEHE} on datasets with varying treatment selection bias. ABCEI is comparatively robust

and control group. We draw 8 000 samples with ten covariates $x \sim \mathcal{N}(\mu_0, 0.5 \cdot (\Sigma + \Sigma^T))$ as control group, where $\Sigma \sim \mathcal{U}((-1, 1)^{10\times10})$. Then we draw 2 000 samples from $x \sim \mathcal{N}(\mu_1, 0.5 \cdot (\Sigma + \Sigma^T))$. By adjusting μ_1, we generate treatment groups with varying treatment selection bias, which can be measured by KL-divergence. For the outcomes, we generate $y|x \sim (w^T x + n)$, where $n \sim \mathcal{N}(0^{2\times1}, 0.1 \cdot I^{2\times2})$ and $w \sim \mathcal{U}((-1, 1)^{10\times2})$.

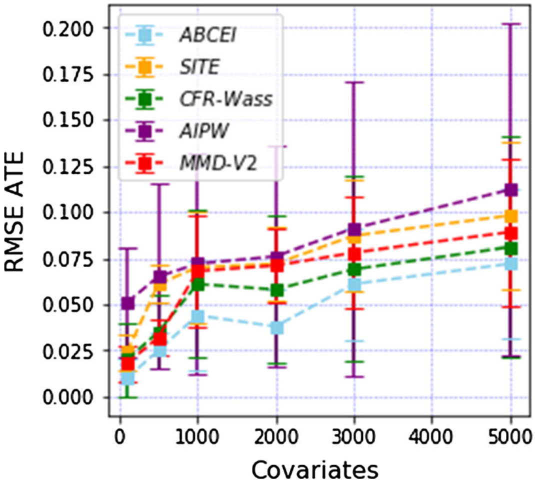
In Fig. 5, we can see the robustness of ABCEI, in comparison with CFR-Wass, BART, and SITE. The reported experimental results are averaged over 100 test sets. From the figure, we can see that with increasing KL-divergence, our method achieves the most stable performance. We do not visualize standard deviations as they are negligibly small.

4.9 Robustness analysis on number of covariates

To investigate how our model performs in regards to the numbers of covariates or confounders, we generate synthetic datasets in which we control the number of covariates, according to the procedure outlined in Ning et al. (2020). The outcome functions are also simulated following Ning et al. (2020). The results are displayed in Fig. 6. As one would expect, all methods experience higher errors when the number of covariates increases; the error increase in ABCEI's performance is in line with the error increase for competing methods. ABCEI has a consistently though not necessarily significantly lower error than the competitors, but more importantly for this specific experiment: ABCEI does not suffer more or less than the competitors from the increase in the number of covariates.

As shown in Fig. 6, our method, without estimating propensity scores, can perform robustly in high-dimensional settings, while preserving the valuable predictive information during covariate balancing.

Fig. 6 Performance comparison (RMSE; lower is better) when varying the number of covariates



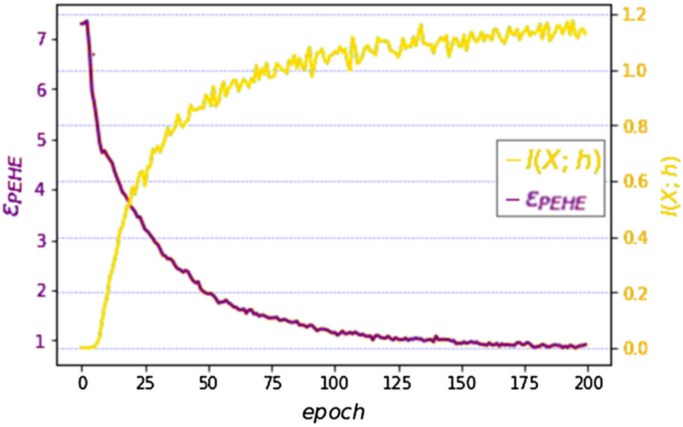
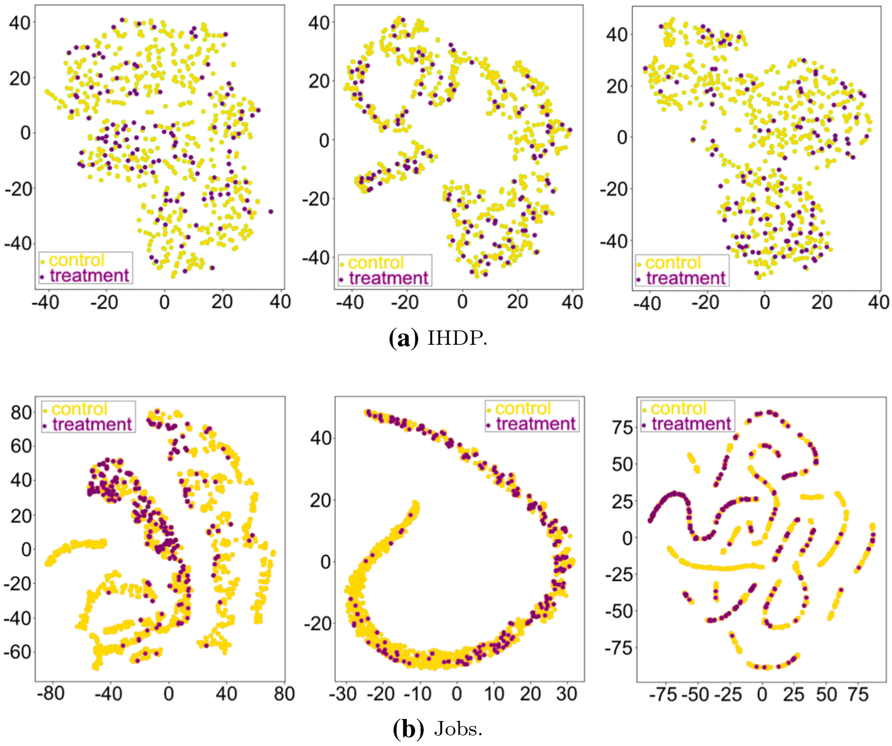


Fig. 7 Mutual information ($I(X;h)$) between representations and original covariates, as well as ϵ_{PEHE} in each epoch. With increasing MI, ϵ_{PEHE} decreases. Best viewed in color; when viewed in black and white, notice that ϵ_{PEHE} corresponds to the generally left-to-right decreasing line and the axis values on the left-hand side, while $I(X;h)$ corresponds to the generally left-to-right increasing line and the axis values on the right-hand side

4.10 Robustness analysis on mutual information estimation

To investigate the impact of minimizing the information loss on causal effect learning, we block the adversarial learning component and train our model on the IHDP dataset. We record the values of the estimated MI and ϵ_{PEHE} in each epoch. In Fig. 7, we report the experimental results averaged over 1 000 test sets. We can see that with increasing MI, the mean square error decreases and reaches a stable region. But without the adversarial balancing component, the ϵ_{PEHE} cannot be further lowered due to the selection bias. This result indicates that even though the estimators benefit from highly predictive information, they will still suffer if imbalance is ignored.



(a) IHDP.

(b) Jobs.

Fig. 8 t-SNE visualization of the treatment and the control groups, on the IHDP and Jobs datasets. The dark (purple) dots are treated units, and the light (yellow) dots are control units. The left figures are the units in original covariate space, the middle figures are the representations learned by ABCEI, and the right figures are the representations learned by CFR-Wass; notice how the latter has control unit clusters unbalanced by treatment observations

4.11 Balancing performance of adversarial learning

In Fig. 8, we visualize the learned representations on the IHDP and Jobs datasets using t-SNE (van der Maaten and Hinton 2008). Such visualizations display high-dimensional datapoints in a two-dimensional embedding, such that similar datapoints find themselves close together and dissimilar datapoints find themselves far apart. As such, the t-SNE visualization represents natural grouping behavior within the dataset. We use this visualization to interpret how well the encoders Φ, learned by the adversarial learning methods, are capable of generating a balanced representation between the control and the treatment groups; if the encoder works well, the t-SNE visualization of the learned representation should display a consistent mix of the control and the treatment group units. It is of no importance in these visualizations what the distribution of the full dataset looks like in particular: the encoder is free to choose a representation that naturally falls apart into an arbitrary number of clusters. Of importance, however, is that the distribution of the control group naturally spans the same locations as the distribution of the treatment group: if there are no parts of the space where the control group appears but the treatment group doesn't, and there are no parts of the space where the treatment group appears but the control group doesn't, then the encoder has rather successfully balanced the control and the treatment groups.

Figure 8 contains t-SNE visualizations of two datasets: IHDP in the top row, and Jobs in the bottom row. For each dataset, we show three visualizations: the left column displays the t-SNE visualizations in the original covariate space, the middle column illustrates the representations learned by ABCEI, and the right column illustrates the representations learned by CFR-Wass. In the top row we see that on the IHDP dataset, both methods achieve a reasonably good matching. The biggest unmatched area is found in the CFR-Wass figure, where one can draw a circle with midpoint $(5, 17)$ featuring quite a few control units and no treatment units. However, in the ABCEI figure, the area around midpoint $(16, -20)$ is not much better matched; there is little difference between the two. On the Jobs dataset, however, the results are quite a bit different. The two methods choose strikingly different embeddings: the ABCEI encoder goes for one long and thick connected component, while the CFR-Wass encoder goes for multiple shorter and thinner components. In itself, this is interesting but not necessarily a mark of quality: the one is not intrinsically better than the other. However, throughout ABCEI's connected component, we can almost always find both control and treatment units nearby. Some of the components of the CFR-Wass encoder also decently balance the control and treatment unit pockets, but there are other components featuring very few treatment units, and three components featuring no treatment units at all. These clusters of control units are badly or not at all balanced by treatment units, and hence we can conclude that ABCEI achieves a better coverage of the treatment group over the control group in the learned representation space than CFR-Wass does. This showcases the degree to which adversarial balancing improves the performance of ABCEI, especially in population causal effect (ATE, ATT) inference.

5 Related work

Studies on causal effect inference give us insight on the true data generating process and allow us to answer what-if questions. The core issue of causal effect inference is the identifiability problem given some data and set of assumptions (Tian and Pearl 2002). Such data includes experimental data from Randomized Controlled Trials (RCTs) and non-experimental data collected from historic observations. Due to the difficulties of conducting RCTs, we mainly focus on the study of causal effect inference based on observational data.

One could split the task of causal effect inference into two parts: given variables, what is the direction of their causal relation, and what is the strength of their causal relation? Mooij et al. (2016) and Marx and Vreeken (2019) have recently provided answers to the first part: determining cause from effect. In this paper, we focus on the second part: the study of assessing the strength of the causal effect, assuming causal relations. Confounding bias might create spurious correlations between variables and would lead to difficulties for the identification of causal effect with observational data. The strong ignorability assumption in the Potential Outcomes framework (Rubin 2005) provides a way to remove the confounding-driven bias and makes causal effect inference possible with observational data. For practical applications, there are some studies focusing on matching-based methods (Ho et al. 2011; Nikolaev et al. 2013) to create comparable groups for causal effect inference. Various similarity measures are applied to achieve better matching results and reduce the estimation error, e.g., Mahalanobis distance and propensity score matching methods are proposed for population causal effect inference (Rubin 2001; Diamond and Sekhon 2013). An information theory-driven approach is proposed by using mutual information as the similarity measure (Sun and Nikolaev 2016).

Recent studies employ deep representation learning methods to derive models that satisfy the conditional ignorability assumption (Li and Fu 2017), in order to make the Conditional Average Treatment Effect identifiable. For instance, Johansson et al. (2016) propose to use a single neural network with the concatenation of representations and treatment variable as the input to predict the potential outcomes. Shalit et al. (2017) propose to train separate models for different treatment outcome systems associating with a measure based on probabilistic integral metric to bound the generalization error. Yao et al. (2018) propose to employ hard samples to preserve local similarity in order to achieve better balancing results. The main difference between ABCEI and the state-of-the-art representation learning-based methods are two-fold: on the one hand, by employing adversarial learning, our balancing method does not need any assumptions on the treatment selection functions; on the other hand, the transformation between original covariate space and the latent space might lead to information loss, but it turns out that this loss can be controlled. In our framework, a mutual information estimator is employed to steer the encoder towards preserving as much highly predictive information about cause and effect as possible.

From the view of graphical interpretation, there are some other difficulties for the identification of causal effect, e.g., selection bias (Correa et al. 2019). Bareinboim and Pearl (2012) propose the use of an instrumental variable for the identification of causal effect. In this paper, we assume there exists only the confounding-driven bias, i.e., that the removal of the confounding-driven bias can make the causal effect accurately measurable. Louizos et al. (2017) propose to estimate causal effect by using proxy variables. A modified variational autoencoder structure is employed to identify the causal effect from observational data. In this paper, we assume that all the confounders can be measured, so that our method is sufficient for the identifiability of the CATE.

6 Conclusions

We propose a novel model for causal effect inference with observational data, called *ABCEI*, which is built on deep representation learning methods. ABCEI focuses on balancing latent representations from the treatment and the control groups by designing a two-player adversarial game. We use a discriminator to distinguish the representations from these two groups. By adjusting the encoder parameters, our aim is to find an encoder that can fool the discriminator, which ensures that the distributions of treatment and control representations are as similar as possible. Our balancing method does not make any assumption on the form of the treatment selection function. With the mutual information estimator, we preserve the highly predictive information going from the original covariate space to latent space. Experimental results on benchmark datasets and synthetic datasets demonstrate that ABCEI is able to achieve robust and substantially better performance than that of the pre-existing state of the art methods.

In future work, we will explore more connections between relevant methods in domain adaptation (Daume III and Marcu 2006) and counterfactual learning (Swaminathan and Joachims 2015) with the methods in causal inference. A proper extension would be to consider multiple treatment assignments or the existence of hidden confounders.

Funding No funding was received to assist with the preparation of this manuscript.

Code availability Publicly available at https://github.com/octeufer/Adversarial-Balancing-based-representation-learning-for-Causal-Effect-Inference.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker PA, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016) Tensorflow: a system for large-scale machine learning. In: Keeton K, Roscoe T (eds) 12th USENIX symposium on operating systems design and implementation, OSDI 2016, Savannah, GA, USA, November 2–4, 2016, USENIX Association, pp 265–283

Abrevaya J, Hsu YC, Lieli RP (2015) Estimating conditional average treatment effects. J Bus Econ Stat 33(4):485–505

Almond D, Chay KY, Lee DS (2005) The costs of low birth weight. Q J Econ 120(3):1031–1083

Autier P, Gandini S (2007) Vitamin D supplementation and total mortality: a meta-analysis of randomized controlled trials. Arch Internal Med 167(16):1730–1737

Bareinboim E, Pearl J (2012) Controlling selection bias in causal inference. In: Lawrence ND, Girolami MA (eds) Proceedings of the fifteenth international conference on artificial intelligence and statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21–23, 2012, JMLR Proceedings, vol 22, pp 100–108

Belghazi MI, Baratin A, Rajeswar S, Ozair S, Bengio Y, Hjelm RD, Courville AC (2018) Mutual information neural estimation. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, PMLR, Proceedings of Machine Learning Research, vol 80, pp 530–539

Benson K, Hartz AJ (2000) A comparison of observational studies and randomized, controlled trials. New England J Med 342(25):1878–1886

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Casucci S, Lin L, Hewner S, Nikolaev A (2017) Estimating the causal effects of chronic disease combinations on 30-day hospital readmissions based on observational medicaid data. J Am Med Inform Assoc 25(6):670–678

Casucci S, Zhou Y, Bhattacharya B, Sun L, Nikolaev A, Lin L (2019) Causal analysis of the impact of homecare services on patient discharge disposition. Home Health Care Serv Q 38(3):162–181

Clevert D, Unterthiner T, Hochreiter S (2016) Fast and accurate deep network learning by exponential linear units (ELUs). In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings

Correa JD, Tian J, Bareinboim E (2019) Identification of causal effects in the presence of selection bias. In: the Thirty-Third AAAI conference on artificial intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, pp 2744–2751

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2008) Nonparametric tests for treatment effect heterogeneity. Rev Econ Stat 90(3):389–405

Daume H III, Marcu D (2006) Domain adaptation for statistical classifiers. J Artif Intell Res 26:101–126

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 84(1):151–161

Diamond A, Sekhon JS (2013) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev Econ Stat 95(3):932–945

Donsker MD, Varadhan SRS (1983) Asymptotic evaluation of certain Markov process expectations for large time: IV. Commun Pure Appl Math 36(2):183–212

Dorie V (2016) NPCI: non-parametrics for causal inference. https://github.com/vdorie/npci

Dorie V, Hill J, Shalit U, Scott M, Cervone D et al (2019) Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. Stat Sci 34(1):43–68

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y, (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27: annual conference on neural information processing systems 2014(December), pp. 8–13, (2014) Montreal. Quebec, Canada, pp 2672–2680

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC, (2017) Improved training of Wasserstein GANs. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017(December), pp. 4–9, (2017) Long Beach. CA, USA, pp 5767–5777

Hill JL (2011) Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 20(1):217–240

Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2019) Learning deep representations by mutual information estimation and maximization. In: 7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019

Ho DE, Imai K, King G, Stuart EA et al (2011) Matchit: nonparametric preprocessing for parametric causal inference. J Stat Softw 42(8):1–28

Imai K, Ratkovic M (2014) Covariate balancing propensity score. J R Stat Soc Ser B (Stat Methodol) 76(1):243–263

Johansson FD, Shalit U, Sontag DA (2016) Learning representations for counterfactual inference. In: Balcan M, Weinberger KQ (eds) Proceedings of the 33nd international conference on machine learning, ICML 2016, New York City, NY, USA, June 19–24, 2016, JMLR Workshop and Conference Proceedings, vol 48, pp 3020–3029

Johnson A, Pollard T, Mark R (2019) MIMIC-III clinical database demo (version 1.4). PhysioNet. https://doi.org/10.13026/C2HM2Q

Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:160035

Kallus N (2018) Balanced policy evaluation and learning. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp 8909–8920

Kallus N (2020) Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In: Proceedings of the 37th International conference on machine learning, ICML 2020, 13–18 July 2020, Virtual Event, PMLR, Proceedings of Machine Learning Research, vol 119, pp 5067–5077

Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings

LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev 76(4):604–620

Li S, Fu Y, (2017) Matching on balanced nonlinear representations for treatment effects estimation. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017(December), pp. 4–9, (2017) Long Beach. CA, USA, pp 929–939

Louizos C, Shalit U, Mooij JM, Sontag DA, Zemel RS, Welling M, (2017) Causal effect inference with deep latent-variable models. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017(December), pp. 4–9, (2017) Long Beach. CA, USA, pp 6446–6456

Marx A, Vreeken J (2019) Identifiability of cause and effect using regularized regression. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G (eds) Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019, ACM, pp 852–861

Mooij JM, Peters J, Janzing D, Zscheischler J, Schölkopf B (2016) Distinguishing cause from effect using observational data: methods and benchmarks. J Mach Learn Res 17(1):1103–1204

Morgan SL, Harding DJ (2006) Matching estimators of causal effects: prospects and pitfalls in theory and practice. Sociol Methods Res 35(1):3–60

Nikolaev AG, Jacobson SH, Cho WKT, Sauppe JJ, Sewell EC (2013) Balance optimization subset selection (boss): an alternative approach for causal inference with observational data. Oper Res 61(2):398–412

Ning Y, Sida P, Imai K (2020) Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. Biometrika 107(3):533–554

Ozery-Flato M, Thodoroff P, El-Hay T (2018) Adversarial balancing for causal inference. Preprint arXiv:1810.07406

Pearl J (2009) Causality. Cambridge University Press

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rubin DB (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv Outcomes Res Methodol 2(3–4):169–188

Rubin DB (2005) Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc 100(469):322–331

Shalit U, Johansson FD, Sontag DA (2017) Estimating individual treatment effect: generalization bounds and algorithms. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, PMLR, Proceedings of Machine Learning Research, vol 70, pp 3076–3085

Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27(3):379–423

Smith JA, Todd PE (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? J Econom 125(1–2):305–353

Sparapani RA, Logan BR, McCulloch RE, Laud PW (2016) Nonparametric survival analysis using Bayesian additive regression trees (BART). Stat Med 35(16):2741–2753

Sun L, Nikolaev AG (2016) Mutual information based matching for causal inference with observational data. J Mach Learn Res 17(1):6990–7020

Swaminathan A, Joachims T (2015) Counterfactual risk minimization: learning from logged bandit feedback. In: Bach FR, Blei DM (eds) Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille, France, 6–11 July 2015, JMLR Workshop and Conference Proceedings, vol 37, pp 814–823

Tam Cho WK, Sauppe JJ, Nikolaev AG, Jacobson SH, Sewell EC (2013) An optimization approach for making causal inferences. Stat Neerlandica 67(2):211–226

Tian J, Pearl J (2002) A general identification condition for causal effects. In: Dechter R, Kearns MJ, Sutton RS (eds) Proceedings of the eighteenth national conference on artificial intelligence and fourteenth conference on innovative applications of artificial intelligence, July 28–August 1, 2002, Edmonton, Alberta, Canada, AAAI Press/The MIT Press, pp 567–573

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242

Yao L, Li S, Li Y, Huai M, Gao J, Zhang A (2018) Representation learning for treatment effect estimation from observational data. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp 2638–2648

Zhao S, Heffernan NT (2017) Estimating individual treatment effect from educational studies with residual counterfactual networks. In: Hu X, Barnes T, Hershkovitz A, Paquette L (eds) Proceedings of the 10th international conference on educational data mining, EDM 2017, Wuhan, Hubei, China, June 25–28, 2017, International Educational Data Mining Society (IEDMS)

Zubizarreta JR (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. J Am Stat Assoc 107(500):1360–1371

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.