

---

# Transfer Learning on Heterogeneous Feature Spaces for Treatment Effects Estimation

---

**Ioana Bica**

University of Oxford, Oxford, UK  
 The Alan Turing Institute, London, UK  
 ioana.bica@eng.ox.ac.uk

**Mihaela van der Schaar**

University of Cambridge, Cambridge, UK  
 University of California, Los Angeles, USA  
 The Alan Turing Institute, London, UK  
 mv472@cam.ac.uk

## Abstract

Consider the problem of improving the estimation of conditional average treatment effects (CATE) for a target domain of interest by leveraging related information from a source domain with a different feature space. This *heterogeneous transfer learning* problem for CATE estimation is ubiquitous in areas such as healthcare where we may wish to evaluate the effectiveness of a treatment for a new patient population for which different clinical covariates and limited data are available. In this paper, we address this problem by introducing several building blocks that use representation learning to handle the heterogeneous feature spaces and a flexible multi-task architecture with shared and private layers to transfer information between potential outcome functions across domains. Then, we show how these building blocks can be used to recover transfer learning equivalents of the standard CATE learners. On a new semi-synthetic data simulation benchmark for heterogeneous transfer learning we not only demonstrate performance improvements of our heterogeneous transfer causal effect learners across datasets, but also provide insights into the differences between these learners from a transfer perspective.

## 1 Introduction

Estimating the personalized effects of interventions from observational data is a fundamental problem in *causal inference* that is crucial for decision-making in many domains: in healthcare, for determining which treatments to give to patients [1], in education, for deciding which school curriculum is best for each student [2, 3], or in public policy for choosing who would benefit from job training programs [4]. Recently, a large number of machine learning methods have been proposed for estimating *conditional average treatment effects* (CATE) which enable such personalized policies [5–16].

Nevertheless, the good performance of these methods on a population of interest relies heavily on the availability of large enough observational datasets for training [17, 18]. In healthcare, for instance, this can be challenging when hospitals with few patients cannot collect enough data (e.g. a large proportion of hospitals in the USA have fewer than 100 beds, which for rare diseases can result in less than 80 training examples per year [19]). Moreover, in situations such as the COVID pandemic, each hospital will initially have very limited amount of data to learn the effectiveness of interventions from [20]. Compared to the predictive setting, this problem is exacerbated in the treatment effects setting where we need to observe both patients who are treated and not treated to be able to reliably train a model for CATE estimation to obtain personalized treatment recommendations for the intended patient population. While data from large national registries can be used to build global models for general use across hospitals, such models do not take into account the particularities of different patient populations (e.g. *different conditional outcome distributions*) and consequently can perform poorly during deployment [19, 21]. Moreover, various hospitals often record different (but overlapping) sets of patient covariates [19, 22] which makes this transfer learning problem [23]

even more challenging as we need to account for the *heterogeneous feature spaces*. Therefore, it becomes crucial to build methods for *heterogeneous transfer learning* that can leverage information from large source datasets with potentially different feature spaces to improve CATE estimation on the target datasets of interest.

We consider the Neyman-Rubin potential outcomes (PO) framework [24, 25], where for each individual we can define two *potential outcomes*, one with and one without the treatment. Out of these, we can only observe the *factual outcome*; the *counterfactual outcome* is never observed. Under the identifiability conditions of overlap and ignorability, observational data can be used to estimate the PO conditional on the patient’s characteristics, which can then be used to obtain CATE. In this paper, we address the problem of heterogeneous transfer for CATE estimation, where we aim to leverage related data from a source domain with different feature space (e.g. data from another hospital) to improve CATE estimation on a target domain from which we only have few training examples.

Due to the *fundamental problem of causal inference* of not being able to observe both PO for a patient [26], heterogeneous transfer learning in the context of CATE estimation becomes significantly more challenging than for supervised learning. In addition to the feature mismatch between domains, the PO may also have different conditional distributions, as covariate relationships and their impact on patients’ response to treatments cannot be expected to stay constant across hospitals/locations [19]. Moreover, as clinicians may use different criteria for assigning treatments for various patient populations, this selection bias may create discrepancies in the covariate shift induced by the treated and control populations in each domain. Consequently, we need to build an approach that can both handle the heterogeneous feature spaces, and also model the similarities and differences between both PO functions and treatment assignment mechanisms across the source and target domains.

For the binary treatment setting in a single patient population, a large number of different approaches for CATE estimation have been proposed where the main design choices involved modelling the PO functions and handling the selection bias present in observational datasets [10, 14–16, 27–29]. We discuss these in more details in Section 2. However, note that each different CATE learner has its own advantages and disadvantages in terms of the inductive biases they use for modelling the PO functions and the covariate shift induced by the selection bias, and thus, different learners will achieve better performance in various scenarios [15, 30]. Therefore, we propose a flexible approach for transfer learning, that (1) preserves the characteristics of each learner in a single domain, while (2) enabling heterogeneous feature spaces and (3) sharing information between PO functions across domains. Firstly, we introduce several building blocks that can be used to adapt the most common CATE learners [10, 15] to transfer information from a source to a target domain. These building blocks involve handling the heterogeneous feature spaces, sharing information between PO functions *across* domains and sharing information between PO functions *within* a single domain. Secondly, we show how these building blocks can be used to build heterogeneous transfer causal effect (HTCE-) learner equivalents of the most common and popular CATE learners based on neural networks [10, 15].

**Contributions.** Our contributions are three-fold (i) we define the problem of heterogeneous transfer learning in the context of CATE estimation and propose several building blocks that can be used to construct models to address this problem, (ii) we use these building blocks to construct HTCE-learner equivalents of the most common CATE learners, and (iii) we propose a new semi-synthetic data simulation and guidelines for evaluating CATE methods for heterogeneous transfer and perform extensive experiments that not only show that our HTCE-learners achieve improved performance, but also provide new insights into the differences between these learners from a transfer perspective.

## 2 Related works

We tackle the problem of heterogeneous transfer learning in the context of CATE estimation. Thus, our work straddles at the intersection of research in (1) causal inference methods for CATE estimation (2) leveraging multiple datasets for CATE estimation (3) multi-task/transfer learning and domain adaptation. Refer to Appendix A for further discussion of related works.

**CATE learners.** The estimation of CATE has received a lot of attention in the causal inference literature and several methods have been proposed to estimate the effects of binary treatments. Out of these, we consider the most popular approaches that involve using model agnostic learning strategies, also known as meta-learners, for CATE estimation [15, 31] or neural network-based models that build shared representations between the PO functions followed by outcome specific layers [10, 15, 27, 28].

The CATE meta-learners can be split into (a) one-step plug-in learners (indirect meta-learners) that estimate the PO from the observational data and then set CATE as the difference in the PO [31] and (b) two-step learners (direct meta-learners) that estimate the PO and/or the propensity score in the first step on the basis of which they build a pseudo-outcome and obtain CATE directly by regressing the input covariates on the pseudo-outcome in the second step [15, 31–33]. Refer to [15] for a more thorough classification of the different meta-learners. Alternatively, several methods based on representation learning with neural networks and multi-task learning have been proposed that involve having shared layers between the PO functions followed by outcome-specific layers. The most standard architecture for this is TARNet [10] which has been extended to allow for different types of information sharing between the PO and propensity score in [14, 15, 27]. To account for the confounding bias present in observational datasets, several approaches have been proposed to extend this model architecture by building balanced representations (treatment invariant representations) [10, 28] and/or incorporating propensity weighting to obtain unbiased estimates of the PO [29, 34]. These different approaches have their own benefits and drawback, which is why it is important to build a heterogeneous transfer learning approach that is general enough to extend all of them.

**Transfer and domain adaptation for CATE estimation.** While, the problem of transfer learning for CATE estimation has also been addressed by [35] the proposed approach considers shared feature spaces and consists of a two-stage training procedure that involves warm-start on the first domain and fine-tuning on the second domain. Alternatively, [36] proposes a CATE estimation method that can generalize to distribution shifts in the patient population in the unsupervised domain adaptation setting. However, they do not assume access to label information in the target domain and only consider a shared feature space between the two domains. In addition, [37] leverages data from multiple different environments, with shared feature spaces, to learn an invariant representation that removes the ‘bad controls’ which induce bias in the CATE estimation. Then, they use this invariant representation to learn shared PO functions across the different environments. Refer to Appendix A for more methods that use multiple datasets for causal inference, although for different purposes than ours.

**Multi-task/transfer learning and domain adaptation.** Methods to address these problems have been extensively studied in the predictive (supervised) setting. We describe here the works most related to ours that consider (a) shared feature space and (b) heterogeneous feature space. For shared feature spaces, methods in domain adaptation focus on handling the covariate shift, i.e. the distribution mismatch between the input features across the different domains and propose various approaches of learning domain invariant representations [38, 39] based on which they learn an outcome function shared between domains. Alternatively, multi-task/transfer learning methods propose various approaches for neural networks to learn from related tasks that involve using both shared and task (domain) specific layers [40, 41] to allow a flexible modelling of the different outcomes. To handle heterogeneous feature spaces, [22] proposes RadialGAN, a method that augments the target dataset with generated samples from the source datasets. However, RadialGAN involves training separate generators and discriminators for each domain and consequently also requires access to enough training data in the target domain. After the data generation, RadialGAN trains separate predictors in each domain that do not share information between each other. Alternatively, Wiens et al. [19] considers the problem of feature mismatch (in a specific healthcare application), but does not address the problem of distributional differences in the outcomes.

We are the first to address the problem of heterogeneous transfer for CATE estimation. We build HTCE-learners that use representation learning to handle the heterogeneous feature spaces and a multi-task architecture with shared and private layers to transfer information between PO across domains, thus also handling the case when different populations respond differently to treatments.

### 3 Problem formalism

Let random variable  $X_i \in \mathcal{X}$  denote a vector of pre-treatment covariates (confounders),  $W_i \in \{0, 1\}$  the assigned binary treatment and  $Y_i$  a categorical or continuous observed outcome for individual  $i$ . Let  $\pi(x) = p(W = 1 | X = x)$  denote the treatment assignment mechanism. As previously mentioned, we work in the Neyman-Rubin potential outcomes (PO) framework [24, 25] and we consider that each individual has two potential outcomes  $Y_i(1)$  and  $Y_i(0)$  for receiving and not receiving the treatment respectively. However, only one of these outcomes can be observed such that  $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ . Let  $\mu_1(x) = \mathbb{E}[Y(1) | X = x]$  and  $\mu_0(x) = \mathbb{E}[Y(0) | X = x]$  be the PO functions.

Our aim is to estimate the conditional average treatment effect (CATE):

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) | X = x] = \mu_1(x) - \mu_0(x) \quad (1)$$

which is the difference between expected outcomes for an individual with covariates  $X = x$ . Let  $\eta = (\mu_0(x), \mu_1(x), \pi(x))$  be the nuisance functions for this CATE estimation problem.

Assume access to a source dataset  $\mathcal{D}^R = \{(X_i^R, W_i, Y_i)\}_{i=1}^{N_R}$  and a target dataset  $\mathcal{D}^T = \{(X_i^T, W_i, Y_i)\}_{i=1}^{N_T}$ . Different domains in applications such as healthcare, have heterogeneous feature spaces such that  $X_i^R \in \mathbb{R}^{D_R}$  and  $X_i^T \in \mathbb{R}^{D_T}$ , where  $D_R \neq D_T$  are the dimensions of the feature spaces. We also consider that the source and target domains have different distributions  $p(X^R) \neq p(X^T)$  (due to their heterogeneous feature spaces), different treatment assignment mechanisms  $p(W = 1 | X^R) \neq p(W = 1 | X^T)$  and different conditional distributions for PO,  $p(Y(w) | X^R) \neq p(Y(w) | X^T)$ . This results in different joint distributions  $p(X^R, W, Y) \neq p(X^T, W, Y)$  which is representative of hospitals recording different types of patient data where the relationships between patient covariates, treatments and outcomes can change across diseases and locations [19, 42]. Nevertheless, we implicitly assume that there is a shared structure between these conditional distributions across domains to enable transfer.

Our aim is to estimate conditional average treatment effects (CATE) for the target domain:

$$\tau^T(x) = \mu_1^T(x^T) - \mu_0^T(x^T), \quad (2)$$

by using both the source  $\mathcal{D}^R$  and target dataset  $\mathcal{D}^T$ . In particular, we want to improve the estimation in the target domain by leveraging information from the source domain. This is useful in the setting where the target dataset is much smaller than the source one  $N_T \ll N_R$  and we can leverage shared structure between the source and target outcome response functions:  $\mu_0^R(x^R), \mu_0^T(x^T)$  and  $\mu_1^R(x^R), \mu_1^T(x^T)$  and treatment assignment mechanisms  $\pi^R(x^R), \pi^T(x^T)$ . To be able to identify the causal effects from observational data, we make the standard assumptions for both domains.

**Assumption 1.** (*Unconfoundedness*) *There are no unobserved confounders, such that the treatment assignment and PO are conditionally independent given the covariates:  $Y(0), Y(1) \perp\!\!\!\perp W | X^T$  and  $Y(0), Y(1) \perp\!\!\!\perp W | X^R$ .*

**Assumption 2.** (*Overlap*)  $\pi^T(x^T) = p(W = 1 | X^T = x^T) > 0, \forall x^T \in \mathcal{X}^T$  and  $\pi^R(x^R) = p(W = 1 | X^R = x^R) > 0, \forall x^R \in \mathcal{X}^R$ .

## 4 Building blocks for CATE transfer learners

In this section, we propose building blocks that enable a flexible transfer approach for CATE learners. The challenge in this setting is threefold as we need to (1) handle heterogeneous feature spaces between the source and target domains (2) share information between PO functions across source and target datasets  $(\mu_1^R, \mu_1^T)$  and  $(\mu_0^R, \mu_0^T)$  as well as (3) share information between PO functions within a single domain  $(\mu_0^R, \mu_1^R)$  and  $(\mu_0^T, \mu_1^T)$ .

We start by addressing (1) and (2) and show how the proposed building blocks can be used to obtain transfer approaches for the most common meta-learning strategies in the treatment effects literature [15]. Then, we propose a building block for addressing (3) to obtain transfer CATE learners that use shared layers and outcome specific layers for the potential outcome functions in each domain [10].

### 4.1 Handling heterogeneous feature spaces between source and target domains

Consider the following split for the source and target covariates  $X^R = (X^s, X^{p_R})$  and  $X^T = (X^s, X^{p_T})$  such that we have a set of features private (specific) to the source dataset  $X^{p_R} \in \mathbb{R}^{D_{p_R}}$ , a set of features private to the target dataset  $X^{p_T} \in \mathbb{R}^{D_{p_T}}$  and a set of shared features between the two datasets  $X^s \in \mathbb{R}^{D_s}$ . To handle the heterogeneous features spaces between the source and target datasets we propose using several encoders to create a common representation that can be used as input to the different transfer CATE learners.

Let  $\phi^{p_R}(x^R) : \mathbb{R}^{D_R} \rightarrow \mathbb{R}^{D_p}$  and  $\phi^{p_T}(x^T) : \mathbb{R}^{D_T} \rightarrow \mathbb{R}^{D_p}$  be domain-specific (private) encoders that map the heterogeneous input features to a representation of size  $D_p$ , such that  $\phi^{p_T}(x^T) = z^{p_T}$  and  $\phi^{p_R}(x^R) = z^{p_R}$ . Moreover, let  $\phi^s(x^s) : \mathbb{R}^{D_s} \rightarrow \mathbb{R}^{D_s}$  be a shared encoder that maps the shared features between the source and target domains into a representation of size  $D_s$  such that  $\phi^s(x^s) = z^s$ .

As illustrated in Figure 1, a source example  $x^R$  is encoded to  $[z^s || z^{p_R}]$  and a target example  $x^T$  to  $[z^s || z^{p_T}]$ , where  $||$  denotes concatenation and where both representations have size  $D_s + D_p$ . Note that an alternative approach would have been to use the domain-specific encoders  $\phi^p$  only for the private features  $x^{p_R}$  and  $x^{p_T}$ . However, inputting the shared features through both types of encoders allows us to learn relationships between them that are shared across the different domain, as well as interactions which are domain-specific.

To discourage redundancy and ensure that  $z^p$  and  $z^s$  encode different information from the input features, we propose using a regularization loss that enforces their orthogonality [39]:

$$\mathcal{L}_{\text{orth}_z} = \|\zeta^{s^\top} \zeta^{p_R}\|_F^2 + \|\zeta^{s^\top} \zeta^{p_T}\|_F^2 \quad (3)$$

where  $\zeta^{p_R}, \zeta^{p_T}$  and  $\zeta^s$  are matrices whose rows are the private  $z^{p_R}$ ,  $z^{p_T}$  and shared  $z^s$  representations for the source and target examples respectively, and  $\|\cdot\|_F^2$  is the squared Frobenius norm.

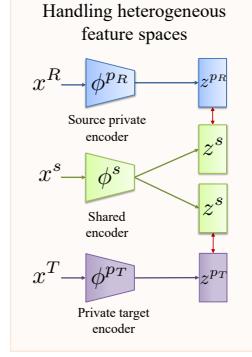


Figure 1: Building block for handling the heterogeneous feature space of the source and target domains.

## 4.2 Sharing information between potential outcomes response functions across domains

As treatment responses can vary between different patient populations, it is important to build a transfer approach that enables learning target-specific outcome functions, while also sharing information from the source domain. We propose a building block for sharing information between PO functions across domains that is inspired by the FlexTENet architecture [14] and by works in multitask learning [41] and that involves having private layers (subspaces) for each domain as well as shared layers.

As shown in Figure 2, for each treatment  $w \in \{0, 1\}$ , we consider a model architecture for estimating its PO functions in the source and target domains  $\mu_w^R$  and  $\mu_w^T$  that consists of  $L$  layers, each having a shared and two private subspaces (one for each domain). For simplicity, we consider the same number of hidden dimensions for each shared and private subspace. Let  $\tilde{h}_{w,l}^{p_R}$ ,  $\tilde{h}_{w,l}^{p_T}$ ,  $\tilde{h}_{w,l}^s$  be the inputs and  $h_{w,l}^{p_R}$ ,  $h_{w,l}^{p_T}$ ,  $h_{w,l}^s$  the outputs of the  $l^{th}$  layer. For  $l > 1$ , similarly to [14], the inputs to the  $(l+1)^{th}$  layer are obtained as follows:  $\tilde{h}_{w,l+1}^{p_R} = [h_{w,l}^s || h_{w,l}^{p_R}]$ ,  $\tilde{h}_{w,l+1}^{p_T} = [h_{w,l}^s || h_{w,l}^{p_T}]$ ,  $\tilde{h}_{w,l+1}^s = [h_{w,l}^s]$ . For  $l = 1$ , we set  $\tilde{h}_{w,1}^{p_R} = \Phi^R(x^R)$ ,  $\tilde{h}_{w,1}^{p_T} = \Phi^T(x^T)$ , and  $\tilde{h}_{w,1}^s = \tilde{h}_{w,1}^{p_R}$  when using an example from the source domain or  $\tilde{h}_{w,1}^s = \tilde{h}_{w,1}^{p_T}$  when using an

example from the target domain, where  $\Phi^R(\cdot)$  and  $\Phi^T(\cdot)$  are input representations. When sharing the encoders from Section 4.1 for both treatments, we set  $\Phi^R(x^R) = [z^s || z^{p_R}]$  and  $\Phi^T(x^T) = [z^s || z^{p_T}]$ . However, as we will see in Section 5.1, this input representation is CATE learner specific and can be extended (see Section 5.2) by adding more representation layers to share information between PO functions within each domain. For the last layer  $L$ , we build  $h_{w,L}^s, h_{w,L}^{p_R}, h_{w,L}^{p_T}$  to each have the same dimension as the potential outcome  $y$ .

Overall, let  $g_w^R, g_w^T$  be the hypothesis functions estimating the potential outcomes in the source and target domains respectively, such that  $g_w^R(\Phi^R(x^R)) = \psi(h_{w,L}^{p_R} + h_{w,L}^s)$  and  $g_w^T(\Phi^T(x^T)) = \psi(h_{w,L}^{p_T} + h_{w,L}^s)$ , where  $\psi$  is the linear function for continuous outcomes and sigmoid function for binary ones. This allows us to define the following loss function for estimating the PO:

$$\mathcal{L}_y = \sum_{i=1}^{N_R} l(y_i, g_{w_i}^R(\Phi_{w_i}^R(x_i^R))) + \sum_{i=1}^{N_T} l(y_i, g_{w_i}^T(\Phi_{w_i}^T(x_i^T))) \quad (4)$$

where  $l(\cdot, \cdot)$  can be the mean squared error for continuous outcomes and the cross-entropy for binary outcomes. Moreover, to discourage redundancy between the shared and private layers, we apply an orthogonal regularization loss, similar to [14, 41]. Let  $m_{w,l-1}^s, m_{w,l-1}^{p_R}, m_{w,l-1}^{p_T}$  be the dimensions of  $h_{w,l-1}^s, h_{w,l-1}^{p_R}, h_{w,l-1}^{p_T}$  respectively, i.e. the outputs of the  $(l-1)^{th}$  layer. Let the weights in the  $l^{th}$  layer be  $\theta_{w,l}^s \in \mathbb{R}^{m_{w,l-1}^s \times m_{w,l}^s}, \theta_{w,l}^{p_R} \in \mathbb{R}^{(m_{w,l-1}^s + m_{w,l-1}^{p_R}) \times m_{w,l}^{p_R}}$  and  $\theta_{w,l}^{p_T} \in \mathbb{R}^{(m_{w,l-1}^s + m_{w,l-1}^{p_T}) \times m_{w,l}^{p_T}}$ . This allows us to use the following orthogonal regularizer:

$$\mathcal{L}_{\text{orthpo}} = \sum_{w \in \{0,1\}} \sum_{l=1}^L \|\theta_{w,l}^s^\top \theta_{w,l,1:m_{w,l-1}^s}^{p_R}\|_F^2 + \|\theta_{w,l}^s^\top \theta_{w,l,1:m_{w,l-1}^s}^{p_T}\|_F^2 \quad (5)$$

A similar approach can be used to share information between the propensity estimation functions ( $\hat{\pi}(x)$ ) that are used by some of the meta-learners. See appendix B for details.

## 5 Heterogeneous transfer causal effect learners

Using these building blocks of handling the heterogeneous feature spaces and sharing information about PO functions across domains, we now propose a transfer learning alternative for the standard meta-learners and neural networks (NNs) based CATE estimators, which we refer to as Heterogeneous Transfer Causal Effect (HTCE) learners. See Appendix C for the pseudo-code for the HTCE-learners.

### 5.1 Heterogeneous transfer learning for CATE meta-learners

Consider NN-based implementations for the nuisance functions  $\eta = (\mu_0(x), \mu_1(x), \pi(x))$  of each CATE meta-learner. Based on the taxonomy of meta-learners described in [15] we divide them into *one-step plug-in learners* and *two-step learners* and provide HTCE-equivalents for both.

**One-step plug-in learners.** One-step plug-in learners estimate  $\hat{\mu}_1$  and  $\hat{\mu}_0$  and then compute CATE as  $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$ . The most common strategies are the T-learner which does not share information between PO functions and the S-learner which does [31]. Here, we consider the simplest S-learner that uses the treatment as an additional feature, while in Section 5.2 we explore more complex strategies for sharing information between PO functions within a single domain.

For our *HTCE-T-learner* (Figure 3 (a)), we build a model where there is no parameter sharing between  $(g_0^R, g_1^R)$  and between  $(g_0^T, g_1^T)$ . We consider treatment-specific encoders  $\phi_w^{p_R}, \phi_w^{p_T}, \phi_w^s$  for the heterogeneous feature spaces, such that  $\Phi_w^R(x^R) = [\phi_w^s(x^s) || \phi_w^{p_R}(x^R)]$  and  $\Phi_w^T(x^T) = [\phi_w^s(x^s) || \phi_w^{p_T}(x^T)]$ . To enable transfer, we use the approach described in Section 4.2 to share information between the hypothesis functions  $(g_0^R(\Phi_0^R(x^R)), g_0^T(\Phi_0^T(x^T)))$  and  $(g_1^R(\Phi_1^R(x^R)), g_1^T(\Phi_1^T(x^T)))$ . Alternatively, for the *HTCE-S-learner* (Figure 3 (b)), we propose concatenating the treatment  $w$  to the shared input features  $x^s$  and using shared layers to obtain the PO functions in each domain. Thus, we build input representations  $\Phi^R(x^R, w) = [\phi^s(x^s, w) || \phi^{p_R}(x^R)]$ ,  $\Phi^T(x^T, w) = [\phi^s(x^s, w) || \phi^{p_T}(x^T)]$  and we enable transfer by having private and shared layers between the hypothesis functions  $g^R(\Phi^R(x^R, w))$  and  $g^T(\Phi^T(x^T, w))$  that share parameters for both treatments within a single domain.

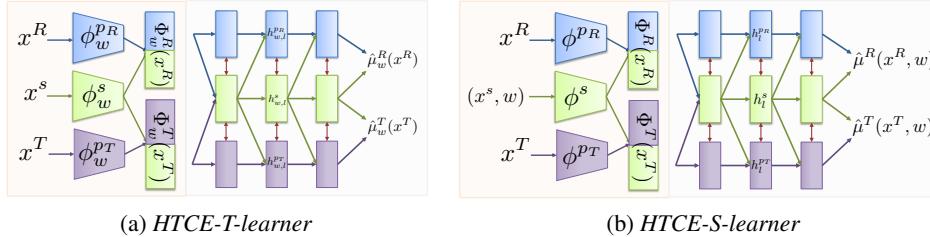


Figure 3: HTCE-one-step plug-in learners. Notice that the HTCE-T-learner uses treatment specific feature encoders and hypothesis functions in each domain, while the HTCE-S-learner concatenates the treatment to the shared features and shares its components for both treatments in each domain.

**Two-step learners.** The two-step (direct) learners consist of a first stage that involves obtaining plug-in estimates  $\hat{\eta}$  of the nuisance parameters  $\eta = (\mu_0, \mu_1, \pi)$  which are used to compute the pseudo-outcome  $\tilde{Y}_{\hat{\eta}}$  and a second stage that involves regressing  $\tilde{Y}_{\hat{\eta}}$  on X directly to obtain  $\hat{\tau}$ . Several

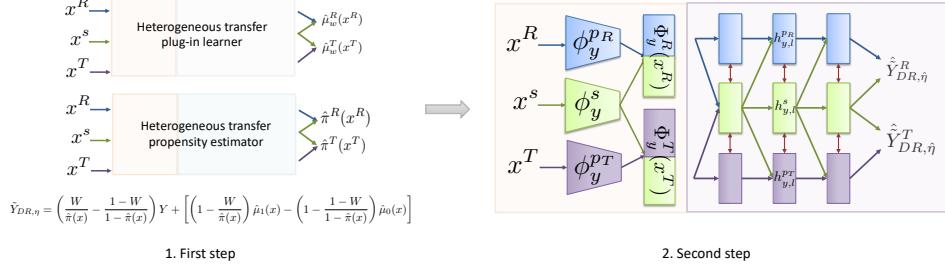


Figure 4: *HTCE-DR-learner*. The first step obtains estimates of  $\hat{\eta}$  using our heterogeneous transfer learning approach that shares information between  $(\hat{\mu}_w^R, \hat{\mu}_w^T)$  and  $(\hat{\pi}^R, \hat{\pi}^T)$ , which are then used to compute the pseudo-outcomes  $\tilde{Y}_{DR,\eta}$ . The second stage regresses the input on the pseudo-outcomes and builds a heterogeneous transfer approach to share information between  $(\tilde{Y}_{DR,\eta}^R, \tilde{Y}_{DR,\eta}^T)$ .

strategies based on regression adjustment (RA), propensity weighting (PW) or doubly robust (DR) meta-learner [15] have been proposed to compute pseudo-outcomes  $\tilde{Y}_{\hat{\eta}}$  which result in unbiased estimates of CATE when  $\eta$  is known:  $\tau(x) = \mathbb{E}[\tilde{Y}_{\hat{\eta}} | X = x]$ . These meta-learners have different theoretical properties depending on the sample size and the selection bias present in observational datasets [15], which is why it is important to build a transfer approach to extend all of them.

We describe here how to build the *HTCE-DR-learner* (Figure 4) and note that a similar approach can be used for to obtain the *HTCE-RA-learner*, and the *HTCE-PW-learner* which use a subset of nuisance parameters to compute the pseudo-outcomes. For the first stage of the *HTCE-DR-learner* of estimating the nuisance parameters  $\hat{\eta}$ , we use an approach similar to the *HTCE-T-learner* where each nuisance function has its own parameters. To handle the heterogeneous feature spaces, we consider treatment-specific feature encoders for the potential outcome functions  $\phi_w^{PR}, \phi_w^{PT}, \phi_w^s$  and additional feature encoders for the propensity estimation  $\phi_\pi^{PR}, \phi_\pi^{PT}, \phi_\pi^s$  such that  $\Phi_y^R(x^R) = [\phi_w^s(x^s) || \phi_w^{PR}(x^R)]$ ,  $\Phi_y^T(x^T) = [\phi_w^s(x^s) || \phi_w^{PT}(x^T)]$  and  $\Phi_\pi^R(x^R) = [\phi_\pi^s(x^s) || \phi_\pi^{PR}(x^R)]$ ,  $\Phi_\pi^T(x^T) = [\phi_\pi^s(x^s) || \phi_\pi^{PT}(x^T)]$ . Using the approach described in Section 4.2 and Appendix B, we share information across domains between each nuisance function:  $(g_w^R(\Phi_y^R(x^R)), g_w^T(\Phi_y^T(x^T)))$  and  $(g_\pi^R(\Phi_\pi^R(x^R)), g_\pi^T(\Phi_\pi^T(x^T)))$  to estimate  $\hat{\eta}$ . We then compute the pseudo-outcomes  $\tilde{Y}_{DR,\hat{\eta}}$  as follows [15]:

$$\tilde{Y}_{DR,\eta} = \left( \frac{W}{\hat{\pi}(x)} - \frac{1-W}{1-\hat{\pi}(x)} \right) Y + \left[ \left( 1 - \frac{W}{\hat{\pi}(x)} \right) \hat{\mu}_1(x) - \left( 1 - \frac{1-W}{1-\hat{\pi}(x)} \right) \hat{\mu}_0(x) \right] \quad (6)$$

For the second stage of regressing the input on the pseudo-outcomes  $\tilde{Y}_{DR,\hat{\eta}}^R, \tilde{Y}_{DR,\hat{\eta}}^T$  directly, we build a similar transfer architecture with feature encoders  $\phi_y^{PR}, \phi_y^{PT}, \phi_y^s$  to handle the heterogeneous feature, followed by hypothesis functions with private and shared layers  $(g_y^R(\Phi_y^R(x^R)), g_y^T(\Phi_y^T(x^T)))$ , where  $\Phi_y^R(x^R) = [\phi_y^s(x^s) || \phi_y^{PR}(x^R)]$ ,  $\Phi_y^T(x^T) = [\phi_y^s(x^s) || \phi_y^{PT}(x^T)]$ .

## 5.2 Sharing information between potential outcome functions within each domain

In previous sections, we focused on sharing information between PO functions *across* domains  $(\mu_1^R, \mu_1^T)$  and  $(\mu_0^R, \mu_0^T)$ . However, many methods in the causal inference literature for estimating treatment effects propose different strategies for also sharing information between PO functions *within* a single domain  $(\mu_0^R, \mu_1^R)$  and  $(\mu_0^T, \mu_1^T)$  [10, 15, 27]. The most common one is the TARNet architecture [10] which uses several layers to build a shared representation, followed by outcome-specific layers.

To obtain *HTCE-TARNet* (Figure 5), we propose using feature encoders  $\phi^{PR}, \phi^{PT}, \phi^s$  that are shared across the different treatments in each

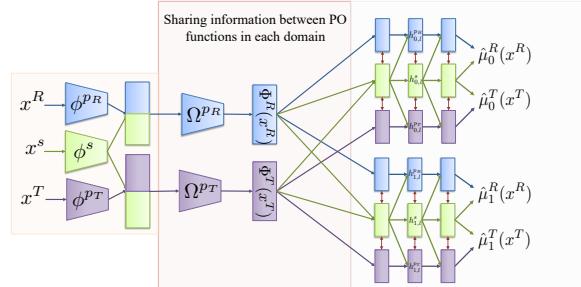


Figure 5: *HTCE-TARNet*.

domain. Then, we construct another building block that involves adding several representation layers on top of the feature embeddings in each domain using domain-specific encoders  $\Omega^{p_T}$  and  $\Omega^{p_R}$ . These representation layers are domain specific as their purpose is to share information between the PO functions within a single domain. Thus, we obtain  $\Phi^R(x^R) = \Omega^{p_R}([\phi^s(x^s) || \phi^{p_R}(x^R)])$ ,  $\Phi^T(x^T) = \Omega^{p_T}([\phi^s(x^s) || \phi^{p_T}(x^T)])$ , which are used as input to the corresponding (treatment-specific) hypothesis functions  $g_w^R(\Phi^R(x^R))$  and  $g_w^T(\Phi^T(x^T))$ .

This approach can be easily extended to recover HTCE-equivalents of other SNet architecture [15] that build additional representation layers to share information between PO (and propensity) estimation within each single domain [27]. Moreover, several methods have been proposed to account for the selection bias in the observational dataset by building balancing representations [10, 27, 28]. This can also be achieved for our HTCE-learners by enforcing regularization on top of the learnt domain-specific representations:  $\Omega^{p_R}(\cdot)$  and  $\Omega^{p_T}(\cdot)$ . Overall, we have built the HTCE-learners to preserve the characteristics of the different CATE-learners in a single domain, but to handle the heterogeneous feature spaces and to share information between outcome functions across domains.

## 6 Evaluation

### 6.1 Experimental set-up

**Semi-synthetic data generation.** Given that counterfactuals, and subsequently ground truth causal effects are never observed in real-datasets, semi-synthetic data is needed to be able to evaluate methods for estimating CATE. While several benchmarks have been proposed for the standard CATE estimation problem [15, 17] in a single domain, none exist for the heterogeneous transfer learning setting. Thus, we propose a new semi-synthetic data simulation to evaluate our HTCE-learners against the benchmarks. Let  $X^R = (X^s, X^{p_R})$  and  $X^T = (X^s, X^{p_T})$  be patient features from real source and target datasets respectively. We simulate outcomes as follows:

$$Y^R(w) = \alpha \sum_{j=1}^{D_S} (v_{w,j}^s X_j^s) / D_S + (1 - \alpha) \left[ \beta \sum_{j=1}^{D_{p_R}} (v_{w,j}^{p_R} X_j^{p_R}) / D_{p_R} + (1 - \beta) \sum_{j=1}^{D_R} (v_j^R X_j^R) / D_R \right] + \epsilon_R \quad (7)$$

$$Y^T(w) = \alpha \sum_{j=1}^{D_S} (v_{w,j}^s X_j^s) / D_S + (1 - \alpha) \left[ \beta \sum_{j=1}^{D_{p_T}} (v_{w,j}^{p_T} X_j^{p_T}) / D_{p_T} + (1 - \beta) \sum_{j=1}^{D_T} (v_j^T X_j^T) / D_T \right] + \epsilon_T \quad (8)$$

where  $v_{w,j}^s, v_{w,j}^{p_R}, v_{w,j}^{p_T}, v_j^R, v_j^T \sim \mathcal{N}(-10, 10)$ ,  $\epsilon_R, \epsilon_T \sim \mathcal{N}(0, 0.1)$  and the parameters  $\alpha$  and  $\beta$  control the amount of sharing between PO across and within domains respectively. For each domain, we simulate the treatment assignment using  $W | X \sim \text{Bernoulli}(\text{Sigmoid}(\kappa(Y(1) - Y(0))))$ , where for  $\kappa = 0$  there is no selection bias and the treatments are assigned randomly, while a high  $\kappa$  indicates high selection bias. Unless otherwise specified, we set  $\alpha = 0.5$ ,  $\beta = 0.5$  and  $\kappa_R = 1$ ,  $\kappa_T = 1$ .

**Datasets.** We perform experiments using patient features from three real datasets with diverse characteristics (e.g. number of features, proportion of categorical features): Twins [43], TCGA [44] and MAGGIC [45]. From the real patient features, the treatment assignments and corresponding outcomes are simulated as described. Refer to Appendix D for details of the datasets and of the way the source and target domain and the shared and private features are obtained. Due to space limitation, we report here the results on Twins. See Appendix E for the results on the other datasets.

**CATE learners and benchmark methods.** The following CATE learners are used in our experiments: S-Learner, T-Learner, DR-Learner and TARNet. We consider the following benchmarks: (1) training the CATE learners only on the target dataset, (2) using only the shared features between the source and target datasets and (3) using RadialGAN [22] to ‘translate’ the source dataset into the target domain. We compare these against our HTCE-learners that leverage the full source and target datasets for training. To ensure a fair comparison, we fixed equivalent hyperparameters in terms of number of layers and units in each layer for the CATE-learners trained on data from the different benchmarks and our HTCE-learners. We describe the hyperparameter used and provide implementation details in Appendix D. We evaluate the different methods using the Root Mean Squared Error of estimating  $\tau^T(x^T)$ , also known as the precision of estimating heterogeneous effects (PEHE) [5]. The results are averaged across 10 runs for which we report the mean PEHE and its standard error.

## 6.2 Results and discussion

**Benchmarks comparison and source of gain.** We first evaluate the HTCE-learners<sup>1</sup> against the benchmarks and perform a source of gain analysis to evaluate the importance of our different components and loss functions. We evaluate the impact of removing the shared and private layers that enable sharing information between the PO functions across domains (HTCE - PO sharing), removing the orthogonal regularization loss for the learnt shared and private feature representations in Equation 3 (HTCE -  $\mathcal{L}_{\text{orth}_z}$ ) and removing orthogonal regularization loss from the PO layers in Equation 5 (HTCE -  $\mathcal{L}_{\text{orthop}_0}$ ). Table 1 reports results on the Twins dataset. Note that our HTCE-learners achieve better performance compared to the baselines and that each component we propose brings performance improvements. When analyzing the performance across the different CATE learners we notice that the S-Learner achieves worse performance due to its strong inducting bias of fully-sharing information between PO functions within each domains, which becomes even more restrictive in the transfer setting.

**Varying the information sharing between domains.** To further gain insights into the differences between the benchmarks under different settings, we vary the parameter  $\alpha$ , which controls the amount of information shared between the PO in the source and target dataset. Figure 6 (top) reports the results for this analysis on the Twins dataset. Note that our flexible approach for information sharing between PO functions across domains achieves good performance both when the PO functions are significantly different between the source and target dataset ( $\alpha = 0.1$ ) and also when they have the same functional form and only depend on the shared features ( $\alpha = 1.0$ ). As a sanity check, we also notice that for ( $\alpha = 1.0$ ) the performance of the HTCE-learners matches the one of the corresponding CATE learners trained only on the shared features between the two domains. Moreover, note that for this setting of  $\alpha = 1.0$  when the PO only depend on the shared features between the two domains, the different CATE learners trained only on the target dataset have a decrease in performance as they also need to perform the task of implicit feature selection with limited amount of data.

**Varying the target dataset size.** We also evaluate performance when varying the size of the target dataset  $N_T$ . We report in Figure 6 (bottom) the results on the Twins datasets. Note that while our HTCE-learners still bring benefits for all values of  $N_T$ , the most performance improvements are when the target dataset size is small and we get diminishing returns as  $N_T$  increases. Moreover, note that TARNet trained only on the target dataset reaches the performance of the HTCE-TARNet learner with less amount of data compered to T-learner trained on the target dataset vs. the HTCE-T-learner.

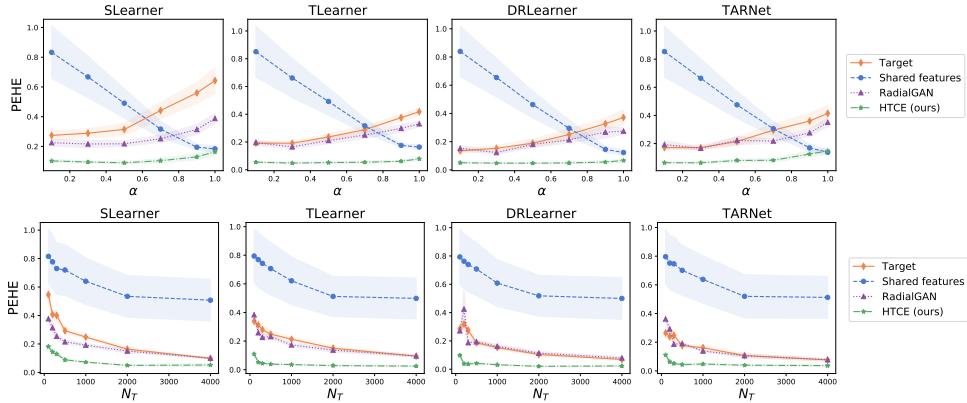


Figure 6: Performance comparison on Twins when varying the information sharing between PO functions across domains through  $\alpha$  (top) and the size of the target dataset through  $N_T$  (bottom).

<sup>1</sup>The code for the HTCE-learners can be found at <https://github.com/vanderschaarlab> and at <https://github.com/ioanabica/HTCE-learners>.

**Effect of selection bias.** Finally, we investigate how the selection bias in the source and target datasets impact transfer performance. We consider three setting for the selection bias in the source dataset  $\kappa_R = 0.0$  for random treatment assignment,  $\kappa_R = 2.0$  for moderate and  $\kappa_R = 10.0$  for strong selection bias. For each setting of  $\kappa_R$ , we vary the selection bias in the target dataset  $\kappa_T$  from  $\kappa_T = 0.0$  to  $\kappa_T = 10.0$  and report the results on Twins for the DR-learner in Figure 7. See Appendix E for results on the other learners and datasets.

Note that our HTCE-DR-learner has consistent performance when increasing the selection bias in both datasets and when there are significant discrepancies in the treatment assignment mechanism between the source and target datasets.

## 7 Discussion

We proposed heterogeneous transfer causal effect (HTCE) learners capable of improving the estimation of CATE on a target domain by leveraging data from a related source domain. This represents the first work that addresses the heterogeneous transfer learning problem for CATE estimation which is prevalent in many causal inference application areas where the personalized effects of interventions for a new population need to be estimated from small target datasets with different features than the available source datasets [3, 19]. We hope that the insights gained from this paper, in terms of the need for sharing information between PO functions to enable transfer and the differences between CATE learners in the transfer setting, will help guide further model development for this problem.

In terms of limitations and directions for future work, while we focused on the binary treatment setting which is the most common in the causal inference literature, addressing the problem of heterogeneous transfer learning is equally important for more complex treatment settings such as treatments with dosage [46, 47], treatment combinations [48, 49] and treatments in the temporal setting [50–54]. Moreover, while we demonstrate the benefits of our HTCE-learners experimentally, future work should also provide theoretical guarantees on their performance in terms of the similarities between the source and target datasets. We provide further discussion of possible extensions and limitations of our method in Appendix F. Finally, we acknowledge that in areas such as healthcare, acting based on incorrect estimates from such causal inference models can have life-threatening implications. Because of this, it becomes crucial that such models undergo extensive testing through clinical trials and expert validation and that they are only used for decision support alongside clinicians.

## Acknowledgments and Disclosure of Funding

We would like to thank the reviewers for their valuable feedback. Moreover, we would also like to thank Alicia Curth for insightful discussions. The research presented in this paper was supported by The Alan Turing Institute, under the EPSRC grant EP/N510129/1, by the US Office of Naval Research (ONR), and by the National Science Foundation (NSF) under grant number 1722516.

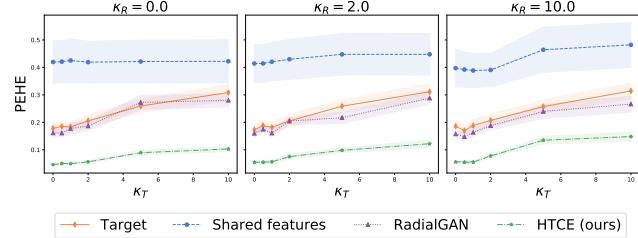


Figure 7: Performance comparison on Twins for DR-learner when varying the selection bias  $\kappa_R$  and  $\kappa_T$  in the source and target datasets.

## References

- [1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- [2] Jan-Eric Gustafsson. Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, 24(3):275–295, 2013.
- [3] Andrew Forney and Scott Mueller. Causal inference in ai education: A primer. 2021.
- [4] Paolo Frumento, Fabrizia Mealli, Barbara Pacini, and Donald B Rubin. Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498):450–466, 2012.
- [5] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.
- [6] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.
- [7] Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.
- [8] Ahmed M Alaa, Michael Weissz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *ICML 2017 Workshop on Principled Approaches to Deep Learning*, 2017.
- [9] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.
- [10] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.
- [11] Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.
- [12] Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in neural information processing systems*, 33:20037–20050, 2020.
- [13] Can Xu, Ahmed Alaa, Ioana Bica, Brent Ershoff, Maxime Cannesson, and Mihaela van der Schaar. Learning matching representations for individualized organ transplantation allocation. In *International Conference on Artificial Intelligence and Statistics*, pages 2134–2142. PMLR, 2021.
- [14] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34, 2021.
- [15] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.
- [16] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela van der Schaar. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 2020.
- [17] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. *International Conference on Learning Representations (ICLR)*, 2018.
- [18] Jinsung Yoon, Camelia Davtyan, and Mihaela van der Schaar. Discovery and clinical decision support for personalized healthcare. *IEEE journal of biomedical and health informatics*, 21(4):1133–1145, 2016.
- [19] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

- [20] Zhaozhi Qian, Ahmed M Alaa, Mihaela van der Schaar, and Ari Ercole. Between-centre differences for covid-19 icu mortality from early data in england. *Intensive care medicine*, 46(9):1779–1780, 2020.
- [21] Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed. Adapting surgical models to individual hospitals using transfer learning. In *2012 IEEE 12th international conference on data mining workshops*, pages 57–63. IEEE, 2012.
- [22] Jinsung Yoon, James Jordon, and Mihaela Schaar. Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. In *International Conference on Machine Learning*, pages 5699–5707. PMLR, 2018.
- [23] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [24] Jersey Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.
- [25] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
- [26] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.
- [27] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.
- [28] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.
- [29] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.
- [30] Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201, 2019.
- [31] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.
- [32] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.
- [33] Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.
- [34] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.
- [35] Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*, 2018.
- [36] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.
- [37] Claudia Shi, Victor Veitch, and David M Blei. Invariant representation learning for treatment effect estimation. In *Uncertainty in Artificial Intelligence*, pages 1546–1555. PMLR, 2021.
- [38] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- [39] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.
- [40] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016.

- [41] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829, 2019.
- [42] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. *arXiv preprint arXiv:1802.06403*, 80:5699–5707, 2018.
- [43] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.
- [44] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113, 2013.
- [45] Stuart J Pocock, Cono A Ariti, John JV McMurray, Aldo Maggioni, Lars Køber, Iain B Squire, Karl Swedberg, Joanna Dobson, Katrina K Poppe, Gillian A Whalley, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European heart journal*, 34(19):1404–1413, 2013.
- [46] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.
- [47] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981*, 2019.
- [48] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multi-cause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34, 2021.
- [49] Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. Ncore: Neural counterfactual representation learning for combinations of treatments. *arXiv preprint arXiv:2103.11175*, 2021.
- [50] Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, pages 7493–7503, 2018.
- [51] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *International Conference on Learning Representations*, 2020.
- [52] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series confounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.
- [53] Jeroen Berrevoets, Alicia Curth, Ioana Bica, Eoin McKinney, and Mihaela van der Schaar. Disentangled counterfactual recurrent networks for treatment effect inference over time. *arXiv preprint arXiv:2112.03811*, 2021.
- [54] Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. Syncitwin: Treatment effect estimation with longitudinal outcomes. *Advances in Neural Information Processing Systems*, 34, 2021.
- [55] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.
- [56] Maximilian Ilse, Patrick Forré, Max Welling, and Joris M Mooij. Combining interventional and observational data using causal reductions. *arXiv preprint arXiv:2103.04786*, 2021.
- [57] Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.
- [58] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. *Twenty-fifth AAAI conference on artificial intelligence*, 2011.
- [59] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

- [60] Seungwhan Moon and Jaime G Carbonell. Completely heterogeneous transfer learning with attention-what and what not to transfer. In *IJCAI*, volume 1, pages 1–2, 2017.
- [61] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.
- [62] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34, 2021.
- [63] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [64] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

## A Expanded related works

In this section, we further expand on the related works described in Section 2.

**Additional methods that leverage multiple datasets for causal inference.** While several other methods propose leveraging multiple datasets to improve CATE estimation, these focus on combining observational data with randomized data from clinical trials with the aim of handling the bias from hidden confounders [55–57]. Moreover, all of these methods learn potential outcome functions that are shared between the different domains. Conversely, we assume that there are no hidden confounders and propose an approach that can share information from a source domain to improve CATE estimation on a target domain of interest.

A different strand of work aims to address the problem of identifiability of causal effects for an observational dataset of interest by leveraging data from other datasets. In particular, [58] investigates how to transfer average treatment (causal) effects obtained from experimental data (such as randomized clinical trials) to an observed population that may have different distribution of covariates, treatments and outcomes and where the causal relationship of interest cannot be identified using only the observational data. The paper assesses under which conditions such average causal effects can be transported according to the differences between the randomized and observational data. The authors also provide a brief discussion of how to transfer these average causal estimates between observational datasets. Alternatively, [59] aims to help identify the average effects of interventions on a target population of interest by integrating multiple types of auxiliary data: data from a randomized study on the same population, data from an observational study on the same population, selection biased data from the same population and data from a randomized study from a different population. All of the auxiliary datasets have the same set of features as the target dataset. Both [58, 59] consider this transportability problem of average treatment effects in the context of causal diagrams. This is different from our set-up for CATE estimation where we assume that the potential outcomes are identifiable in both the source and target domains. Moreover, we also handle the case of heterogeneous feature spaces and only assume access to a source domain larger than a target domain, without putting any restrictions on whether this data is experimental or observational.

**Transfer learning and domains adaptation in the predictive setting.** Another method loosely related to ours is the one of [60], which considers the case of having both heterogeneous feature and label spaces in the context of natural language processing and proposes a method that learns a common embedding for the features and labels and then builds a mapping between them. In this paper, we consider a shared label space. In addition, [61] uses a causal approach to address the problem of domain adaptation for the predictive setting, where it considers labelled data in one or more source domains, unlabelled data in the target domain, same feature spaces between the source and target domains and aims to learn predictive functions that are invariant to the changes between domains to be able to reliably estimate the outcomes in the target domain. [61] proposes tackling this unsupervised domain adaptation by modelling the different distributions between the source and target domains as different contexts of a single underlying system (where the context represents interventions causing the distribution shifts between domains, such that the source and target domains come from different interventions).

Table 2 provides a comparison of our method with the most relevant related works. Note that our heterogeneous transfer learning approach is designed to share information between one source domain and a target domains. Moreover, we assume access to labels on the target domain and build target-specific outcome functions that share a common structure with the source domain.

Method	Number of domains	Heterogeneous feature spaces	Heterogeneous label spaces	Access to labels on target domain	Domain-specific outcome functions
Causal inference	CATE learners [10, 14, 15]	1	No	No	No
	Johansson et al. [36]	2	No	No	No
	Shi et al. [37]	$N$	No	No	No
	Kunzel et al. [35]	2	No	No	Yes
Learning across domains	Ganin et al. [38]	2	No	No	No
	Magliacane et al. [61]	$N$	No	No	No
	Rudner et al. [41]	$N$	No	No	Yes
	Yoon et al. [22]	$N$	Yes	No	Yes
	Moon et al. [60]	2	Yes	Yes	No
(Ours) HTCE-learners	2	Yes	No	Yes	Yes

Table 2: Comparison of our proposed method with related works.

## B Heterogeneous transfer propensity estimator

For our *HTCE-propensity estimator* that is used as part of the two-step meta-learners (e.g. HTCE-DR-learner) we build a model that shares information between the treatment assignment mechanisms in the source and target datasets as illustrated in Figure 8. In particular, our HTCE-propensity estimator consists of encoders  $\phi_{\pi}^{p_T}, \phi_{\pi}^{p_R}, \phi_{\pi}^s$  for the heterogeneous feature spaces, such that  $\Phi_{\pi}^R(x^R) = [\phi_{\pi}^s(x^s) || \phi_{\pi}^{p_R}(x^R)]$  and  $\Phi_{\pi}^T(x^T) = [\phi_{\pi}^s(x^s) || \phi_{\pi}^{p_T}(x^T)]$ . To enable transfer, we use an approach similar to the one described in Section 4.2, but this time to share information between the hypothesis functions for the propensity estimation ( $g_{\pi}^R(\Phi_{\pi}^R(x^R)), g_{\pi}^T(\Phi_{\pi}^T(x^T))$ ). Moreover, we use the following outcome loss:

$$\mathcal{L}_{\pi} = \sum_{i=1}^{N_R} l(w_i, g_{\pi}^R(\Phi^R(x_i^R))) + \sum_{i=1}^{N_T} l(w_i, g_{\pi}^T(\Phi^T(x_i^T))) \quad (9)$$

where  $l(\cdot, \cdot)$  represents the binary cross-entropy loss (for binary treatment).

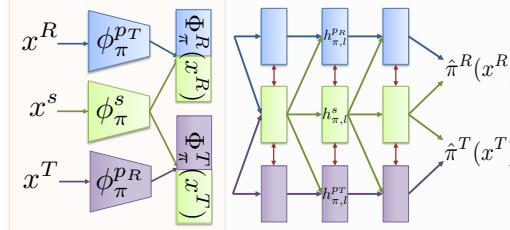


Figure 8: HTCE-propensity estimator.

## C Pseudo-code for the HTCE-learners

Algorithm 1 provides the pseudo-code for training our proposed HTCE-T-learner and Algorithm 2 provides the pseudo-code for training our proposed HTCE-TARNNet. Note that a similar training procedure can be used for the HTCE-S-Learner, HTCE-propensity estimator and in the second stage of the two-step learners (e.g. HTCE-DR-learner). The main differences are in the encoder used, in terms of sharing layers between PO functions within each domain and the outcome used for computing the corresponding loss function. Refer to Section 5.1 and Appendix B for architectural details. Similarly to [15], we do not use cross-fitting for the two-step learners (HTCE-DR-learner). Refer to [15] for more details. For simplicity, we set the mini-batch size to be the same in both the source and target datasets:  $B_T = B_R = B$ . Moreover, note that in practice, we multiply  $\mathcal{L}_{\text{orth}_z}$  and  $\mathcal{L}_{\text{orth}_{\text{PO}}}$  by 0.01 to ensure that all losses have a similar scale and we use early stopping on the validation dataset from the target domain to check for convergence.

---

**Algorithm 1** Pseudo-code for training the HTCE-T-learner.

---

1: Input: Dataset with expert demonstrations from source  $\mathcal{D}^R$  and target  $\mathcal{D}^T$  domains, learning rate  $\lambda$ , mini-batch size for source dataset  $B_R$ , mini-batch size for target dataset  $B_T$   
 2: Initialize:  $\theta$  (all parameters of the HTCE-T-Learner)  
 3: **while** not converged **do**  
 4:     Sample mini-batch of  $B_R$  demonstrations from the source dataset  $(x_i^R, w_i, y_i) \sim \mathcal{D}^R$  and a minibatch of  $B_T$  demonstrations from the target dataset  $(x_i^T, w_i, y_i) \sim \mathcal{D}^T$ .  
 5:     **Handle heterogeneous feature spaces:**  
 6:     **for**  $i = 1 \dots B_R$  **do** ▷ Process batch from source domain.  
 7:          $z_i^{s_R} = \phi_{w_i}^s(x_i^s)$ , where  $x_i^s$  is part of  $x_i^R$ ;  $z_i^{p_R} = \phi_{w_i}^R(x_i^R)$   
 8:         Set:  $\zeta^{s_R} = [z_1^{s_R} \dots z_{B_R}^{s_R}]^\top$ ,  $\zeta^{p_R} = [z_1^{p_R} \dots z_{B_R}^{p_R}]^\top$   
 9:         **for**  $i = 1 \dots B_T$  **do** ▷ Process batch from target domain.  
 10:              $z_i^{s_T} = \phi_{w_i}^s(x_i^s)$ , where  $x_i^s$  is part of  $x_i^T$ ;  $z_i^{p_T} = \phi_{w_i}^R(x_i^T)$   
 11:             Set:  $\zeta^{s_T} = [z_1^{s_T} \dots z_{B_T}^{s_T}]^\top$ ,  $\zeta^{p_T} = [z_1^{p_T} \dots z_{B_T}^{p_T}]^\top$   
 12:         Compute orthogonal regularization loss for the feature representations:  

$$\mathcal{L}_{\text{orth}_z} = \|\zeta^{s_R}^\top \zeta^{p_R}\|_F^2 + \|\zeta^{s_T}^\top \zeta^{p_T}\|_F^2$$
  
 13:     **Share information between PO functions across domains:**  
 14:     **for**  $i = 1 \dots B_R$  **do** ▷ Process batch from source domain.  
 15:         **for**  $l = 1 \dots L$  **do**  
 16:             **if**  $l == 1$  **then**  
 17:                  $\tilde{h}_{w_i,1}^{p_R} = [z_i^{s_R} || z_i^{p_R}]$ ;  $\tilde{h}_{w_i,1}^s = \tilde{h}_{w_i,1}^{p_R}$   
 18:             **else**  
 19:                  $\tilde{h}_{w_i,l}^{p_R} = [h_{w_i,l-1}^s || h_{w_i,l-1}^{p_R}]$ ;  $\tilde{h}_{w_i,l}^s = [h_{w_i,l-1}^s]$ .  
 20:              $h_{w_i,l}^s = \text{Shared\_Layer}_{w_i,l}(\tilde{h}_{w_i,l}^s)$ ;  $h_{w_i,l}^{p_R} = \text{Private\_Layer}_{w_i,l}^R(\tilde{h}_{w_i,l}^{p_R})$   
 21:              $\hat{y}_i^R = \psi(h_{w_i,L}^{p_R} + h_{w_i,L}^s)$   
 22:         **for**  $i = 1 \dots B_T$  **do** ▷ Process batch from target domain.  
 23:             **for**  $l = 1 \dots L$  **do**  
 24:                 **if**  $l == 1$  **then**  
 25:                      $\tilde{h}_{w_i,1}^{p_T} = [z_i^{s_T} || z_i^{p_T}]$ ;  $\tilde{h}_{w_i,1}^s = \tilde{h}_{w_i,1}^{p_T}$   
 26:                 **else**  
 27:                      $\tilde{h}_{w_i,l}^{p_T} = [h_{w_i,l-1}^s || h_{w_i,l-1}^{p_T}]$ ;  $\tilde{h}_{w_i,l}^s = [h_{w_i,l-1}^s]$ .  
 28:                  $h_{w_i,l}^s = \text{Shared\_Layer}_{w_i,l}(\tilde{h}_{w_i,l}^s)$ ;  $h_{w_i,l}^{p_T} = \text{Private\_Layer}_{w_i,l}^T(\tilde{h}_{w_i,l}^{p_T})$   
 29:                  $\hat{y}_i^T = \psi(h_{w_i,L}^{p_T} + h_{w_i,L}^s)$   
 30:         Compute outcome loss:  

$$\mathcal{L}_y = \sum_{i=1}^{B_R} l(y_i, \hat{y}_i^R) + \sum_{i=1}^{B_T} l(y_i, \hat{y}_i^T)$$
  
 31:     Compute orthogonal regularization loss for the PO shared and private layers:  

$$\mathcal{L}_{\text{orth}_{\text{PO}}} = \sum_{w \in \{0,1\}} \sum_{l=1}^L \|\theta_{w,l}^s{}^\top \theta_{w,l,1:m_{l-1}^s}^{p_R}\|_F^2 + \|\theta_{w,l}^s{}^\top \theta_{w,l,1:m_{w,l-1}^s}^{p_T}\|_F^2$$
  
 32:     Parameters update:  
 33:      $\theta \leftarrow \theta - \lambda \nabla_\theta (\mathcal{L}_y + \mathcal{L}_{\text{orth}_z} + \mathcal{L}_{\text{orth}_{\text{PO}}})$   
 34: **Output:** Learnt parameters  $\theta$ 


---

---

**Algorithm 2** Pseudo-code for training the HTCE-TARNet.

---

1: Input: Dataset with expert demonstrations from source  $\mathcal{D}^R$  and target  $\mathcal{D}^T$  domains, learning rate  $\lambda$ , mini-batch size for source dataset  $B_R$ , mini-batch size for target dataset  $B_T$   
2: Initialize:  $\theta$  (all parameters of the HTCE-TARNet)  
3: **while** not converged **do**  
4:     Sample mini-batch of  $B_R$  demonstrations from the source dataset  $(x_i^R, w_i, y_i) \sim \mathcal{D}^R$  and a minibatch of  $B_T$  demonstrations from the target dataset  $(x_i^T, w_i, y_i) \sim \mathcal{D}^T$ .  
5:     **Handle heterogeneous feature spaces:**  
6:     **for**  $i = 1 \dots B_R$  **do** ▷ Process batch from source domain.  
7:          $z_i^{s_R} = \phi^s(x_i^s)$ , where  $x_i^s$  is part of  $x_i^R$ ;  $z_i^{p_R} = \phi^R(x_i^R)$   
8:         Set:  $\zeta^{s_R} = [z_1^{s_R} \dots z_{B_R}^{s_R}]^\top$ ,  $\zeta^{p_R} = [z_1^{p_R} \dots z_{B_R}^{p_R}]^\top$   
9:         **for**  $i = 1 \dots B_T$  **do** ▷ Process batch from target domain.  
10:           $z_i^{s_T} = \phi^s(x_i^s)$ , where  $x_i^s$  is part of  $x_i^T$ ;  $z_i^{p_T} = \phi^R(x_i^T)$   
11:          Set:  $\zeta^{s_T} = [z_1^{s_T} \dots z_{B_T}^{s_T}]^\top$ ,  $\zeta^{p_T} = [z_1^{p_T} \dots z_{B_T}^{p_T}]^\top$   
12:         Compute orthogonal regularization loss for the feature representation:  

$$\mathcal{L}_{\text{orth}_z} = \|\zeta^{s_R}^\top \zeta^{p_R}\|_F^2 + \|\zeta^{s_T}^\top \zeta^{p_T}\|_F^2$$

13:     **Share information between PO functions within each domain:**  
14:     **for**  $i = 1 \dots B_R$  **do** ▷ Process batch from source domain.  
15:          $\tilde{z}_i^R = \Omega^{p_R}([z_i^{s_R} || z_i^{p_R}])$   
16:         **for**  $i = 1 \dots B_T$  **do** ▷ Process batch from target domain.  
17:          $\tilde{z}_i^T = \Omega^{p_T}([z_i^{s_T} || z_i^{p_T}])$

18:     **Share information between PO functions across domains:**  
19:     **for**  $i = 1 \dots B_R$  **do** ▷ Process batch from source domain.  
20:         **for**  $l = 1 \dots L$  **do**  
21:             **if**  $l == 1$  **then**  
22:                  $\tilde{h}_{w_i,1}^{p_R} = \tilde{z}_i^R$ ;  $\tilde{h}_{w_i,1}^s = \tilde{h}_{w_i,1}^{p_R}$   
23:             **else**  
24:                  $\tilde{h}_{w_i,l}^{p_R} = [h_{w_i,l-1}^s || h_{w_i,l-1}^{p_R}]$ ;  $\tilde{h}_{w_i,l}^s = [h_{w_i,l-1}^s]$ .  
25:                  $h_{w_i,l}^s = \text{Shared\_Layer}_{w_i,l}(\tilde{h}_{w_i,l}^s)$ ;  $h_{w_i,l}^{p_R} = \text{Private\_Layer}_{w_i,l}^R(\tilde{h}_{w_i,l}^{p_R})$   
26:                  $\hat{y}_i^R = \psi(h_{w_i,L}^{p_R} + h_{w_i,L}^s)$   
27:         **for**  $i = 1 \dots B_T$  **do** ▷ Process batch from target domain.  
28:             **for**  $l = 1 \dots L$  **do**  
29:                 **if**  $l == 1$  **then**  
30:                      $\tilde{h}_{w_i,1}^{p_T} = \tilde{z}_i^T$ ;  $\tilde{h}_{w_i,1}^s = \tilde{h}_{w_i,1}^{p_T}$   
31:                 **else**  
32:                      $\tilde{h}_{w_i,l}^{p_T} = [h_{w_i,l-1}^s || h_{w_i,l-1}^{p_T}]$ ;  $\tilde{h}_{w_i,l}^s = [h_{w_i,l-1}^s]$ .  
33:                      $h_{w_i,l}^s = \text{Shared\_Layer}_{w_i,l}(\tilde{h}_{w_i,l}^s)$ ;  $h_{w_i,l}^{p_T} = \text{Private\_Layer}_{w_i,l}^T(\tilde{h}_{w_i,l}^{p_T})$   
34:                      $\hat{y}_i^T = \psi(h_{w_i,L}^{p_T} + h_{w_i,L}^s)$   
35:         Compute outcome loss:  

$$\mathcal{L}_y = \sum_{i=1}^{B_R} l(y_i, \hat{y}_i^R) + \sum_{i=1}^{B_T} l(y_i, \hat{y}_i^T)$$

36:     Compute orthogonal regularization loss for the PO shared and private layers:  

$$\mathcal{L}_{\text{orth}_{\text{PO}}} = \sum_{w \in \{0,1\}} \sum_{l=1}^L \|\theta_{w,l}^s{}^\top \theta_{w,l,1:m_{l-1}^s}^{p_R}\|_F^2 + \|\theta_{w,l}^s{}^\top \theta_{w,l,1:m_{w,l-1}^s}^{p_T}\|_F^2$$

37:     Parameters update:  
38:      $\theta \leftarrow \theta - \lambda \nabla_\theta (\mathcal{L}_y + \mathcal{L}_{\text{orth}_z} + \mathcal{L}_{\text{orth}_{\text{PO}}})$   
39:     **Output:** Learnt parameters  $\theta$

---

## D Experimental details

### D.1 Dataset description

In this section, we provide a description of the three datasets used for evaluation: MAGGIC [45], Twins [43] and TCGA [44]. Note that these datasets have diverse characteristics in terms of the number of features and the proportion of categorical/continuous features. To the best of our knowledge, none of the medical datasets used contain personally identifiable information nor offensive content. Given that Twins and TCGA are datasets from a single domain, we then explain how we used them to obtain source and target datasets with heterogeneous feature spaces.

**MAGGIC.** MAGGIC [45] consists of datasets with heterogeneous features from 30 domains representing medical studies of patients who have experienced heart failure. Note that this medical dataset is private. Refer to [45] for details on how the dataset was collected and curated. Each medical study consists of a number of patients ranging from 190 to 13279. From these, we sample a dataset with  $< 500$  patients to represent our target domain and a dataset with  $> 500$  patients to represent our source domain. The different datasets in MAGGIC consist of a mixture of categorical and continuous features representing clinical covariates for patients who have experienced heart failure such as age, gender, serum creatinine, diabetes, beta-blocker prescription, lower systolic blood pressure, lower body mass, time since diagnosis, current smoker, chronic obstructive pulmonary disease, etc. The total number of features across the different medical studies in MAGGIC is 216 and the average number of features in a single study is 73. After sampling the source and target dataset from the different medical studies in MAGGIC, we simulate treatment assignments and outcomes as described in Section 6. Note that different source and target datasets are sampled for each of the different 10 random seeds used for all experimental results.

**Twins.** Twins [43] represents one of the standard benchmark datasets in the causal inference literature for estimating the effects of binary treatments. The dataset consists of data from 11400 pairs of twin births in the USA recorded between 1989-1991 [43]. Refer to [43] for details of how the dataset was collected and curated. We use the publicly available version of the dataset from: <https://github.com/AliciaCurth/CATENets>. For each twin pair, 39 relevant covariates were recorded related to the parents, pregnancy and birth such as marital status, mother's age, number of previous births, pregnancy risk factors, quality of care during pregnancy, number of gestation weeks, etc. These represent a mixture of continuous and categorical features. We obtain our full dataset with  $N_F = 114000$  examples and  $D_F = 39$  features by randomly sampling one of the twins to observe. We then sub-sample our source and target datasets as described below and simulate treatments and outcomes as described in Section 6.

**TCGA.** TCGA is a dataset consisting of cancer patients for which gene expression measurements were recorded [44]. The dataset is publicly available and has been used by previous methods in causal inference [46, 47, 62]. In particular, we use the publicly available version of the TCGA dataset as used by DRNet [47] <https://github.com/d909b/drnet>. Refer to [44] for details of how the dataset has been collected and curated. The dataset consists of  $N_F = 9659$  patients (samples), for which we use the measurements from the  $D_F = 100$  most variable genes. Note that these are all continuous features. Moreover, we log-normalized the gene expression data and scale each feature in the  $[0, 1]$  interval. We then sub-sample our source and target datasets as described below. The treatment assignment and outcomes are simulated as described in Section 6.

**Obtaining source and target datasets for Twins and TCGA.** Unless the size of the target dataset is fixed, from the full datasets, we sample a target dataset of size  $N_T \sim \mathcal{U}(100, 500)$  and a source dataset of size  $N_R \sim \mathcal{U}(1000, N_F - N_T)$ , where  $N_T$  is the size of the target dataset,  $N_R$  is the size of the source dataset and  $N_F$  is the size of the full dataset. For the experiments where we vary the size of the target dataset, we set  $N_T \in \{100, 200, 300, 500, 1000, 2000, 4000\}$  and we again sample the size of the source dataset from  $N_R \sim \mathcal{U}(1000, N_F - N_T)$ . Moreover, to create heterogeneous feature spaces for the source and target domains, let  $D_F$  be the number of features in the full dataset. From these, we randomly sample  $D_{pr} \in \mathcal{U}(5, D_F/3)$  features that are private for the source dataset,  $D_{pt} \in \mathcal{U}(5, D_F/3)$  features that are private for the target dataset and  $D_s \in \mathcal{U}(5, D_F/3)$  features that are shared between the two. Note that different sizes for the source and target datasets and different sets of shared and private features are sampled for each of the different 10 random seeds used for all experimental results.

We use the full source datasets for training the benchmarks making use of the source dataset. Each target dataset undergoes a split into 56/24/20% for training, validation and testing respectively. The validation dataset is used for early stopping.

**Summary statistics for the datasets** We provide in Table 3 summary statistics of the different characteristics of the full source and target datasets used across our 10 experimental runs. Note that for the experiments where we vary the size of the target dataset,  $N_T$  is fixed as described above. To highlight the diversity of the datasets used for evaluating the benchmarks, we highlight in Table 3 the mean and standard deviations for the size of the source dataset  $N_R$ , size of the target dataset  $N_T$ , number of features shared between the source and target datasets  $D_s$ , number of features private to the source dataset  $D_{p_R}$ , number of features private to the target dataset  $D_{p_T}$ , proportion of shared features in the source dataset  $D_s/(D_s + D_{p_R})$ , proportion of shared features in the target dataset  $D_s/(D_s + D_{p_T})$  and the Maximum Mean Discrepancy (MMD) (computed using RBF kernel) between the shared features of the source and target datasets  $MMD(X^{SR}, X^{ST})$ . Note that the MMD between the shared features of the source and target datasets for MAGGIC is much higher than for TCGA and Twins, as the different datasets in MAGGIC represent medical studies with different patient populations.

Table 3: Summary statistics of the datasets sampled for training the benchmarks.

	TCGA	Twins	MAGGIC
$N_R$	$4219 \pm 2569$	$5106 \pm 2988$	$2533 \pm 3711$
$N_T$	$360 \pm 120$	$360 \pm 120$	$267 \pm 75$
$D_s$	$20 \pm 8$	$9 \pm 2$	$27 \pm 7$
$D_{p_R}$	$15 \pm 7$	$9 \pm 2$	$25 \pm 9$
$D_{p_T}$	$21 \pm 9$	$8 \pm 3$	$21 \pm 15$
$D_s/(D_s + D_{p_R})$	$0.57 \pm 0.17$	$0.49 \pm 0.10$	$0.54 \pm 0.13$
$D_s/(D_s + D_{p_T})$	$0.50 \pm 0.14$	$0.52 \pm 0.08$	$0.60 \pm 0.19$
$MMD(X^{SR}, X^{ST})$	$0.0009 \pm 0.0005$	$0.0011 \pm 0.0008$	$0.056 \pm 0.057$

## D.2 Implementation details and hyperparameter setting

As mentioned in Section 6, we evaluate the following benchmarks (1) training the CATE-learners only on the target dataset, (2) using only the shared features between the source and target datasets and (3) using RadialGAN [22] to ‘translate’ the source dataset into the target domain (4) training the HTCE-learners on the full source and target datasets. We describe here the implementation details for RadialGAN [22], for the different CATE-learners used (S-Learner, T-Learner, DR-Learner and TARNet) and the corresponding HTCE-learners (HTCE-S-Learner, HTCE-T-Learner, HTCE-DR-Learner and HTCE-TARNet).

**RadialGAN.** We use the publicly available implementation of RadialGAN [22]<sup>1</sup> which we augment to the treatment effects setting by considering the treatment separate from the rest of the features. While RadialGAN, was developed to work with any number of source datasets and uses multiple GAN architectures to learn to translate from any one of the source datasets to the target domain, we consider here a simplified case with one source dataset. We use the same hyperparameter setting described in [22] and found in the publicly available implementation for the generators, discriminators, encoders and decoders architectures and optimization (optimizer, batch size, learning rate).

**CATE-learners.** We use the publicly available implementation of the different CATE learners from [15]<sup>2</sup>. This allows us to have a consistent implementation for the S-Learner, T-Learner, DR-Learner and TARNet [10]. Moreover, we use components similar to those in [10, 15] for all networks which ensure that  $\hat{\mu}_w(x)$ ,  $\hat{\pi}(x)$  and  $\hat{\tau}(x)$  have access to similar number of layers and units in total, thus resulting in equally complex nuisance functions. In particular, for the S-learner,  $\hat{\mu}(x, w)$  consists of 5 layers of 200 units each, for the T-learner, each of  $\hat{\mu}_0(x), \hat{\mu}_1(x)$  consists of 5 layers with 200 units each. For the DR-Learner, we set each of  $\hat{\mu}_0(x), \hat{\mu}_1(x), \hat{\pi}(x), \hat{\tau}(x)$  to have 5 layers with 200 units each. Finally, TARNet consists of 3 representation layers of 200 units each, followed by two outcome heads, each with 2 layers of 100 units each. We use ReLU activation function for the hidden layers

<sup>1</sup><https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/RadialGAN>

<sup>2</sup><https://github.com/AliciaCurth/CATENets>

and linear/sigmoid activation for the last layer (for continuous/binary outcomes respectively). The different CATE learners are trained using the Adam optimizer [63], with learning rate set to 0.0001 and batch size 128. We perform early stopping on the validation dataset from the target domain.

**HTCE-learners.** Each  $\phi^{p_R}, \phi^{p_T}, \phi^s$  for handling the heterogeneous feature spaces consists of one hidden layer with 100 units each. For the HTCE-S-Learner, we consider a building block of 5 layers with 100 hidden units for each shared and private subspace to share information between  $\hat{\mu}^R(x^R, w), \hat{\mu}^T(x^T, w)$ . Note that this essentially gives us 200 hidden units for the hypothesis function in each domain, thus ensuring that the nuisance functions for the CATE vs. HTCE-learners in each domain have similar complexity. For the HTCE-T-learner, we use separate feature encoders for each treatment  $w \in \{0, 1\}$  ( $\phi_w^{p_R}, \phi_w^{p_T}, \phi_w^s$  each consisting of one hidden layer with 100 units), and we use two building blocks of 5 layers with 100 units for each shared and private subspace to share information between  $(\hat{\mu}_0^R(x^R), \hat{\mu}_0^T(x^T))$  and  $(\hat{\mu}_1^R(x^R), \hat{\mu}_1^T(x^T))$ . Similarly, for the HTCE-DR-learner, we use the same architecture as for the HTCE-T-learner to obtain the plug-in estimates of the potential outcomes, which consists of separate feature encoders for each treatment  $w$ , and two building blocks of 5 layers with 100 units for each shared and private subspace to share information between  $(\hat{\mu}_0^R(x^R), \hat{\mu}_0^T(x^T))$  and  $(\hat{\mu}_1^R(x^R), \hat{\mu}_1^T(x^T))$ . Moreover, for the propensity estimation, we use separate feature encoders ( $\phi_\pi^{p_R}, \phi_\pi^{p_T}, \phi_\pi^s$  each consisting of one hidden layer with 100 units) and a building blocks of 5 layers with 100 units for each shared and private subspace to share information between  $(\hat{\pi}^R(x^R), \hat{\pi}^T(x^T))$ . Finally, for the pseudo-outcome regressor we also use separate feature encoders ( $\phi_y^{p_R}, \phi_y^{p_T}, \phi_y^s$  each consisting of one hidden layer with 100 units) and a building blocks of 5 layers with 100 units for each shared and private subspace to share information between  $(\hat{\tau}^R(x^R), \hat{\tau}^T(x^T))$ . For HTCE-TARNet, we use feature encoders shared between the two treatments ( $\phi^{p_R}, \phi^{p_T}, \phi^s$  each consisting of one hidden layer with 100 units), followed by 3 representation layers (for each of  $\Omega^R, \Omega^T$ ) and two building blocks of 2 layers with 100 units for each shared and private subspace to share information between  $(\hat{\mu}_0^R(x^R), \hat{\mu}_0^T(x^T))$  and  $(\hat{\mu}_1^R(x^R), \hat{\mu}_1^T(x^T))$ . We use ReLU activation function for the hidden representation layers, SeLU activation for the shared and private subspace layers and linear/sigmoid activation for the last layer (for continuous/binary outcomes respectively). Note that for the building block of sharing information between outcome functions, we build upon the FlexTENet implementation [14]. The different HTCE-learners are trained using the Adam optimizer [63], with learning rate set to 0.0001 and batch size 128. Moreover, we also use the validation dataset from the target domain for early stopping.

All of the experiments were run on a virtual machine with 6CPUs, an Nvidia K80 Tesla GPU and 56GB of RAM.

## E Additional experimental results

This section provides results for the different experimental settings considered in Section 6 for the remaining datasets and for all learners. Note that we observe similar trends for the different experimental settings across all datasets.

### E.1 Benchmarks comparison and source of gain

Table 4 reports results on the TCGA and MAGGIC datasets for the comparison of the HTCE-learners against the benchmarks and the source of gain analysis. To reiterate, we compare the following benchmarks (1) training the CATE learners only on the target dataset, (2) using only the shared features between the source and target datasets to train the CATE learners (3) using RadialGAN [22] to ‘translate’ the source dataset into the target domain and training the CATE learners on the target dataset augmented in this way and (4) training our HTCE-learners on the full source and target datasets. Moreover, we evaluate the impact of removing the shared and private layers that enable sharing information between the PO functions across domains (HTCE - PO sharing), removing the orthogonal regularization loss for the learnt shared and private feature representations in Equation 3 (HTCE -  $\mathcal{L}_{\text{orth}_z}$ ) and removing orthogonal regularization loss from the PO layers in Equation 5 (HTCE -  $\mathcal{L}_{\text{ortho}_{\text{po}}}$ ). We note that on the TCGA and MAGGIC datasets as well, the HTCE-learners achieve better performance compared to the baselines and that each component we propose brings performance improvements.

Table 4: Benchmarks comparison and source of gain analysis in terms of PEHE on TCGA and MAGGIC datasets.

Learner	TCGA				MAGGIC			
	S-Learner	T-Learner	DR-Learner	TARNet	S-Learner	T-Learner	DR-Learner	TARNet
Target	0.15 ± 0.02	0.12 ± 0.01	0.10 ± 0.01	0.11 ± 0.01	0.39 ± 0.04	0.21 ± 0.01	0.19 ± 0.01	0.19 ± 0.01
Shared features	0.30 ± 0.07	0.30 ± 0.07	0.29 ± 0.07	0.30 ± 0.07	0.32 ± 0.03	0.28 ± 0.03	0.27 ± 0.03	0.27 ± 0.03
RadialGAN	0.12 ± 0.01	0.11 ± 0.01	0.09 ± 0.01	0.13 ± 0.03	0.31 ± 0.01	0.19 ± 0.01	0.19 ± 0.02	0.19 ± 0.01
HTCE - PO sharing	0.15 ± 0.03	0.10 ± 0.01	0.09 ± 0.01	0.10 ± 0.01	0.29 ± 0.02	0.15 ± 0.01	0.14 ± 0.01	0.17 ± 0.01
HTCE - $\mathcal{L}_{\text{orth}_z}$	0.10 ± 0.01	0.07 ± 0.01	0.06 ± 0.01	0.07 ± 0.01	0.25 ± 0.02	0.10 ± 0.01	0.09 ± 0.01	0.12 ± 0.01
HTCE - $\mathcal{L}_{\text{ortho}_{\text{po}}}$	0.11 ± 0.01	0.08 ± 0.01	0.07 ± 0.01	0.05 ± 0.01	0.26 ± 0.02	0.10 ± 0.01	0.10 ± 0.01	0.12 ± 0.01
HTCE (ours)	<b>0.07 ± 0.01</b>	<b>0.06 ± 0.01</b>	<b>0.04 ± 0.01</b>	<b>0.06 ± 0.01</b>	<b>0.24 ± 0.02</b>	<b>0.08 ± 0.01</b>	<b>0.08 ± 0.01</b>	<b>0.10 ± 0.01</b>

When looking at the performance of the benchmarks, using only the shared features between the two datasets generally has poor performance. This is due to the fact that this introduces hidden confounders in the CATE estimation since all of the patient features in both domains are used to obtain the potential outcomes and the treatment assignment in our data simulation (see Section 6). Moreover, note that RadialGAN only slightly improves performance over only training the CATE learners on the target dataset. We hypothesise that this may also be due to the small size of the target dataset which hinders the generator networks in RadialGAN to learn how to map examples from the source domain to the target one.

### E.2 Varying the information sharing between domains

Figure 9 reports the results for varying the parameter  $\alpha$ , which controls the amount of information shared between the PO in the source and target domains, on the TCGA (top) and MAGGIC (bottom) datasets. We again notice that our HTCE-learners, which use a flexible approach for information sharing between PO functions across domains achieve good performance both when the PO functions are significantly different between the source and target dataset ( $\alpha = 0.1$ ) and also when they have the same functional form and only depend on the shared features ( $\alpha = 1.0$ ). In addition, for high values of  $\alpha$  and especially for  $\alpha = 1.0$  we again notice the degradation in performance due to the fact that the learners need to perform the implicit task of feature selection.

### E.3 Varying the target dataset size.

Figure 10 reports the results on when varying the size of the target dataset  $N_T$  on the TCGA dataset. Note that for the MAGGIC dataset, the size of the target domain is not simulated and it represents the size of the various studies selected as target domains. Thus, we only report here results on TCGA where we directly simulate the size of the target dataset. We notice that the benefits of doing transfer learning degrade as we increase the size of the target dataset.

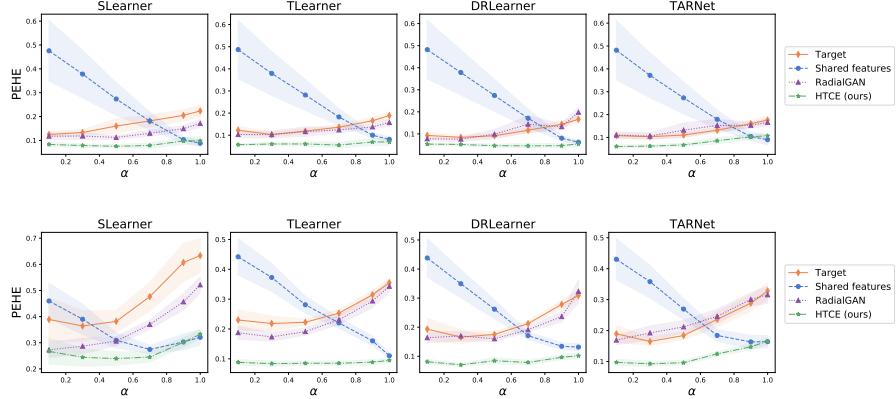


Figure 9: Performance comparison on TCGA (top) and MAGGIC (bottom) when varying the information sharing between PO functions across domains through  $\alpha$ .

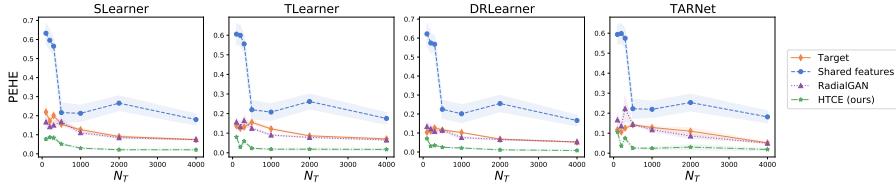


Figure 10: Performance comparison on TCGA when varying the size of the target dataset ( $N_T$ ).

#### E.4 Effect of selection bias

Finally, we report results when varying the selection bias in the source and target domains for all learners on all three datasets: Twins (Figure 11), TCGA (Figure 12) and MAGGIC (Figure 13). We consider three setting for the selection bias in the source dataset  $\kappa_R = 0.0$  for random treatment assignment,  $\kappa_R = 2.0$  for moderate and  $\kappa_R = 10.0$  for strong selection bias. For each setting of  $\kappa_R$ , we vary the selection bias in the target dataset  $\kappa_T$  from  $\kappa_T = 0.0$  to  $\kappa_T = 10.0$ . Our HTCE-learners have consistent performance when increasing the selection bias in both datasets and when there are significant discrepancies in the treatment assignment mechanism between the source and target datasets. While the impact of increasing the selection bias differs among the learners (e.g. TARNet vs. DRLearnert), note that this is due to the intrinsic characteristics of each learner on a single domain, as the DRLearnert uses propensity weighting to adjust for the selection bias, while TARNet does not.

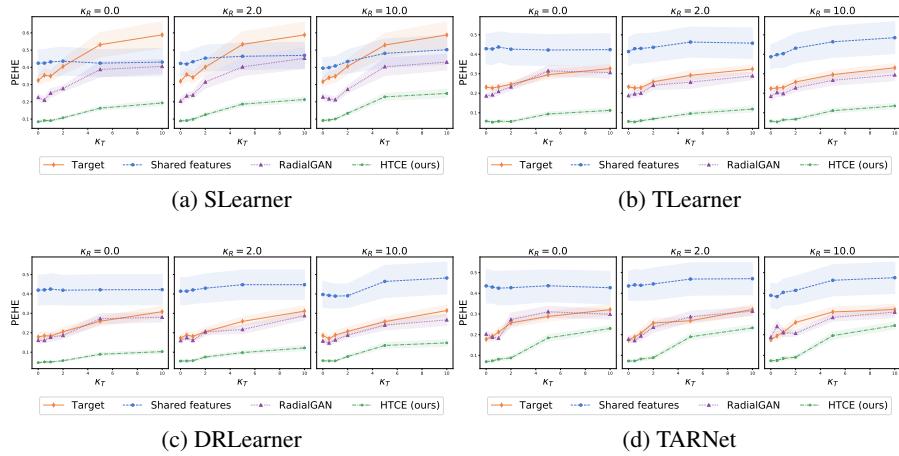


Figure 11: Performance comparison on Twins for all learners when varying the selection bias  $\kappa_R$  and  $\kappa_T$  in the source and target datasets.

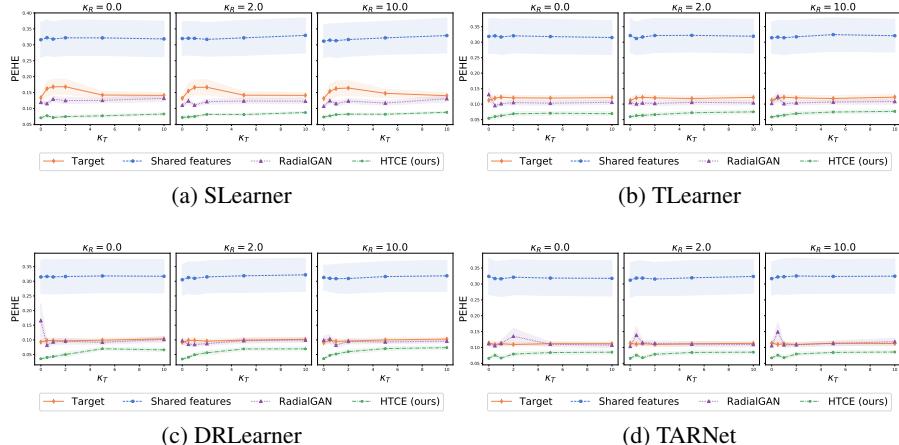


Figure 12: Performance comparison on TCGA for all learners when varying the selection bias  $\kappa_R$  and  $\kappa_T$  in the source and target datasets.

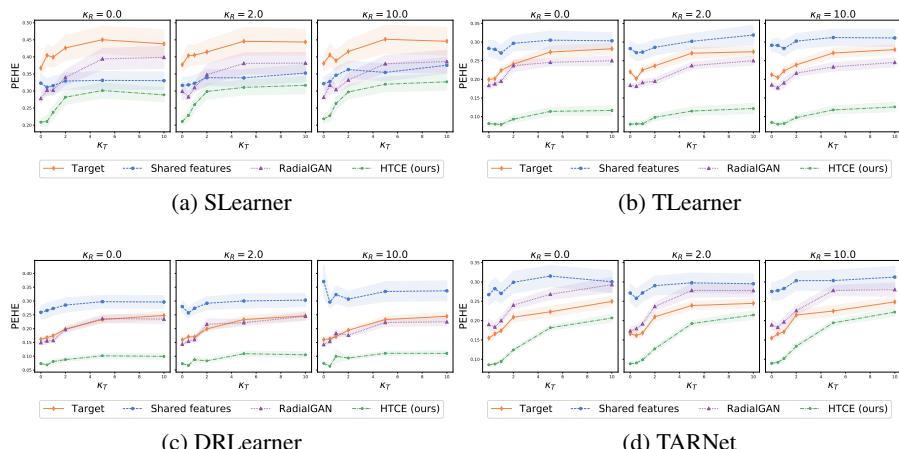


Figure 13: Performance comparison on MAGGIC for all learners when varying the selection bias  $\kappa_R$  and  $\kappa_T$  in the source and target datasets.

## F Additional discussion about extensions and limitations

In this section, we describe how our method could be extended to additional scenarios than the ones considered in the paper and further discuss limitations.

**Multiple source datasets.** While in this paper we mainly consider the standard and most common transfer learning setting [64] of leveraging a source dataset to improve the estimation of outcomes on a target dataset of interest, our proposed approach could be easily extended to having multiple source domains, and in fact, would scale linearly to incorporating additional domains.

More precisely, consider access to  $M$  source domains,  $\{\mathcal{D}^{R_m}\}_{m=1}^M$ , where each source domain consists of  $\mathcal{D}^{R_m} = \{(X_i^{R_m}, W_i, Y_i)\}_{i=1}^{N_{R_m}}$  and a target domain as described in Section 3. The objective would still be the one from Section 3 of improving the estimation of CATE for the target domain, but in this case, we need to leverage data from all  $M$  source domains. Our proposed building blocks could be extended as follows to handle the additional source domains. The building block for handling heterogeneous feature spaces between source and target domains could be extended by having encoders  $\phi^{R_m}$  for the private features  $X^{p_{R_m}}$  of each source domain (part of  $X^{R_m}$ ). Moreover, we can build an approach that shares information between the potential outcomes of each source domain with the target domain by having  $L$  layers shared between all source domains and the target domains and  $L$  private layers for each source domain. For source domain  $m$ , the input to the layer  $(l+1)$ -th can be obtained using  $\tilde{h}_{w,l+1}^{p_{R_m}} = [h_{w,l}^s | h_{w,l}^{p_{R_m}}]$  where  $h_{w,l}^{p_{R_m}}$  is the output of the  $l$ -th private layer for source domain  $m$  and  $h_{w,l}^s$  is the output of the  $l$ -th shared layer across all domains.

While from a model development perspective this extension can be easily done, one also needs to consider whether the multiple source domains satisfy the underlying implicit assumptions for which such an architecture would be appropriate. In particular, it would be important to consider whether the PO functions across all source domains share information with the PO functions in the target domain, as using source domains that are significantly different from the target domain could harm performance.

**Streaming datasets.** Another important setting to consider is the one of having streaming datasets. In this scenario, one option could be the case where we already have a source domain and a target domain and we have incoming data streaming from the target domain. One such example in healthcare would be the case where in a single hospital we start collecting more or different clinical covariates for the patients which we now want to use for CATE estimation. The source domain would be the patients with only the original set of features, while the target domain would be the patients with the new set of features. However, as we start collecting these additional features, the initial target dataset will be small but it can start increasing with time as we observe more patients. In this setting, we can train an HTCE-learner with the initial data available from the source and target domains. However, instead of retraining the full HTCE-learner as we obtain more patients from the target domain, one option would be to fine-tune the weights using the incoming examples. While this is outside the scope of our paper, we believe that it would be important to investigate appropriate ways for performing such fine-tuning.

Another option could be the case of having full source datasets streaming, while the target dataset remains fixed. This would happen in the setting where for instance we gradually get access to data from multiple locations and we want to use these datasets as source datasets. In this scenario, one possibility would be to use the approach described above for having multiple source datasets and retraining a new model that incorporates all of the available source datasets as we get access to them. Another possibility would be to, instead of retraining a full model as we go from  $M$  to  $M+1$  source domains, we can add the needed private encoder and layers for the  $(M+1)$ -th domain and train only the new parameters and fine-tune the shared ones with the data from the new domain. While this is again outside our scope, it can provide interesting avenues for future work.

**Unknown domains.** In this paper, we assume that the source and target domains are known. However, given that we handle the setting of transfer learning for heterogeneous feature spaces, if the domains are unknown, one way to split available data into different domains, in this case, would be by using the same features to denote a single domain. For instance, if different patients in the dataset have recorded different sets of features, then the patients can be grouped according to having the same information collected for them and these can denote the different domains. Then, one needs to decide which represents the target domain, while keeping in mind that doing transfer learning is most

beneficial when the target dataset is small (as highlighted by our experimental results in Section 6.2 and Figure 6 (bottom)).

**Personalized feature spaces.** In the case where different features are available for different patients and the source and target domains are unknown, then the patients can be grouped according to their feature spaces as described in the paragraph above. On the other hand, if the source and target datasets are pre-defined and the patients within each dataset have different features available for them, one possible option would be to consider the super-set of their features as the different feature spaces and consider the features that are not available for each patient as missing. However, as our HTCE-learners has not been designed for this particular setting of having missing data, one would also need to investigate if additional assumptions (in addition to ensuring that the no hidden confounders and overlap assumptions are still satisfied) are needed to be able to obtain valid estimates of causal effects.