Evaluating transfer learning for treatment-effect estimation in small samples

Anonymous authors
Paper under double-blind review

Abstract

Estimating how treatment effects vary across individuals typically requires large samples, yet in many applied settings, such as small clinical cohorts, regional surveys, and field experiments, the available data are limited and treatment is not randomized. We ask whether transfer learning can relax this constraint by borrowing strength from a larger, related source dataset. Building on the Treatment-Agnostic Representation Network (TARNet; Shalit et al., 2017) and its transfer-learning extension (Aloui et al., 2023), we provide the first systematic evaluation of this framework in small-sample, non-randomized settings, through two simulation studies and an empirical application. The first simulation constructs the target by subsampling the source, which isolates target sample size and treatment-selection bias. Transfer reduces both error and bias whenever the target is small or violates ignorability, and the gains from source data. The second simulation introduces genuine source-target mismatch corresponding to the two mismatch terms in the transfer bound: covariate shift and changes in the causal mechanism. Transfer is robust to covariate shift and purely prognostic changes, but it degrades steadily once the treatment effect itself changes, the one difference that alters the quantity being estimated. Moreover, the Causal Inference Task Affinity (CITA) score is far more sensitive to covariate shift than to the effect changes that most threaten transfer. In an application to the India Human Development Survey, transfer stabilizes otherwise implausible small-sample estimates of the effect of household firewood collection on children's study time, aligning them in sign and magnitude with the larger source. We close with practical guidance for transfer learning in small-sample causal estimation.

1 Introduction

Investigating treatment effects at both the population and individual levels is crucial across many fields, from healthcare and economics to the social sciences. The individual treatment effect (ITE) quantifies the difference in an individual's outcome when receiving a treatment versus not. In contrast to the average treatment effect (ATE), the ITE takes treatment heterogeneity into account and provides person-specific effects (Jin et al., 2023). Because the two potential outcomes are never both observed for the same unit, recovering the ITE from data is a challenging estimation problem, yet accurate individual-level effects support more targeted decisions than the ATE alone (Hoogland et al., 2021). This heterogeneity matters across applied fields where the same intervention can help some individuals while leaving others unaffected or worse off(Bloome & Schrage, 2021).

Shalit et al. (2017) introduced a theoretical framework and a family of algorithms for estimating the ITE from observational data under the strong ignorability assumption. The approach learns a representation in which the treated and control distributions are encouraged to be similar, and derives a generalization bound on the treatment-effect error that depends on both the standard estimation error and the distance between the induced treated and control distributions. Using an integral probability metric (IPM), they obtain explicit bounds for distances such as the Wasserstein distance and the Maximum Mean Discrepancy (MMD). On real and semi-synthetic data, the method is competitive with or improves upon established treatment-effect estimators. The balancing penalty defines the counterfactual-regression (CFR) objective,

and the Treatment-Agnostic Representation Network (TARNet) is its unpenalized special case; we build on TARNet and its transfer-learning extension below.

A central obstacle is sample size. Data collection is often costly, so the available datasets are small (Shaikhina & Khovanova, 2017), while flexible machine-learning estimators require substantial data; smaller datasets therefore reduce both the power and the precision of the resulting estimates (Kokol et al., 2022).

What counts as a "small" dataset is task- and model-dependent. Estimating heterogeneous effects is far more data-hungry than estimating average effects, so even a few hundred observations can be too few for reliable individual-level estimates (Zhou & Xie, 2020; Brand et al., 2021; Yang et al., 2020). Such settings are common in practice such as small clinical cohorts, rare subgroups, costly field experiments, and surveys subject to attrition (Bruhn & McKenzie, 2009; Lugtig, 2014) and they leave flexible models like TARNet prone to overfitting and to unstable individual-level predictions (Koch et al., 2025). More generally, a dataset is "small" when the available observations are insufficient to recover the underlying data-generating process with adequate reliability, so that the success of a flexible estimator depends not only on its capacity but on whether the data are informative enough for the complexity of the problem (Baćak & Kennedy, 2019).

1.1 Transfer learning

Causal conclusions rarely remain confined to the dataset in which they were estimated, and extending an effect to a new hospital, region, population, or time period requires addressing external validity: the joint distribution of covariates and outcomes in the target setting often differs from the source (Rojas-Carulla et al., 2018). Classical fixes such as post-hoc reweighting or stratification help only when all relevant differences are measured and correctly modeled which is an assumption that is hard to guarantee (Huang & Parikh, 2024). Simply pooling datasets can also bias estimates if background covariates differ substantially or if treatment effects vary across contexts; when one dataset is much larger, it can dominate estimation and overwhelm signal from the smaller one (Vo et al., 2022; Wei et al., 2023). More broadly, an effect that is internally valid in a randomized trial may fail to generalize if the trial's sample is narrow, while observational data in the target domain may be more representative yet introduce confounding. Ignoring effect modifiers can produce substantial external-validity bias; in some cases, a transferred randomized controlled experiment (RCT) estimate can be more biased for the target population than an adjusted observational estimate (Degtiar & Rose, 2023). These pitfalls underscore the risk of naively applying causal conclusions across settings without explicit adjustment.

One promising way to address these challenges posed by small datasets is through transfer learning, which entails applying insights gained in one setting to another (Wankmüller, 2024; Pan & Yang, 2009; Ruder et al., 2019). The idea is to start with a model trained on a rich dataset which is called as source dataset and adapt it to a small dataset which is called as target dataset, even when the two domains are not identical (Torrey & Shavlik, 2010) . Transferring causal knowledge is far more complex than standard predictive transfer learning. In causal inference, the goal is to predict how an intervention affects a specific individual's outcome. By searching for representations that capture causal regularities shared across environments, a transfer-learning approach can absorb stable structures from the source dataset while flexibly adjusting to local idiosyncrasies in the target dataset (Rojas-Carulla et al., 2018). The benefits extend further when multiple datasets record different interventions on the same underlying system. In such cases, transfer methods can exploit cross-study invariances, sometimes without needing a complete causal graph, to deliver more reliable predictions in new domains (Magliacane et al., 2018).[1]

Addressing the open question of how to reuse causal models across studies, Aloui et al. (2023) developed both theory and practice for transfer learning of ITEs with TARNet. They first establish a lower bound showing that any transferred model's performance on a new task is fundamentally limited by the (unobservable) counterfactual error, warning against naive reuse of source models. A complementary upper bound then shows that effective transfer is possible when the discrepancies between source and target distributions are

[1] Pearl & Bareinboim (2022) formalize conditions under which whole causal Directed Acyclic Graphs (DAGs) can be transferred from one environment to another. This concept of transportability is related to transfer learning but focuses on whole conceptualizations, and not only ITE estimation as we will concentrate on. One of the major differences is that for transfer learning, not the whole causal mechanism needs to be known (Aloui et al., 2023). See more in the discussion section.

small which are the discrepancies we manipulate directly in our simulations. They also introduce the CITA, a score that captures the similarity between source and target, and provide initial evidence that the method can improve efficiency in ITE estimation and that candidate sources can be judged for their suitability before transfer.

1.2 Scope and Outline

Transfer learning for causal inference is still an emerging area, and there is limited empirical evidence on how it behaves under the small-sample, non-randomized conditions where it would be most useful.

This paper studies the transfer-learning framework of Aloui et al. (2023) under exactly these conditions. To our knowledge, this is the first systematic evaluation of the framework across controlled simulation settings that vary the source and target sample sizes, the treatment-assignment mechanism in the target (randomized and non-randomized), and the degree of source–target mismatch. This design lets us identify when transfer learning is most beneficial for causal estimation in small samples. Although our empirical illustration uses household firewood collection and children's study time in India, the methodological problem is general: in many applications one wants heterogeneous effects in a small focal sample while richer data exist for a related region, cohort, institution, or time period.

The paper makes three contributions. First, we give a structured overview of the emerging literature on transfer learning for causal inference, situating Aloui et al. (2023) among related approaches. Second, we provide the first systematic evaluation of this framework across two controlled simulation studies: one isolates target sample size and treatment-selection bias under a shared source population, and the other isolates genuine source–target mismatch—covariate shift, a prognostic change, and a change in the treatment effect—tracing how each affects transfer. Third, we characterize when transfer helps and when it does not: it reduces error and bias for small or non-randomized targets and absorbs covariate shift and a prognostic change, but degrades once the treatment effect itself differs; and we show that the CITA affinity score tracks covariate shift far more closely than the effect changes that most threaten transfer. An application to the India Human Development Survey (IHDS) illustrates the methodological problem in a real, data-constrained setting, and should be read as an illustration rather than a definitive substantive analysis.

The remainder of the paper is organized as follows. We first introduce TARNet and its transfer-learning extension (TL-TARNet) and discuss the underlying assumptions, limitations, and related machine-learning approaches. We then report two simulation studies and an empirical application demonstrating the framework. Finally, we give practical guidelines for implementation and outline directions for future research.

2 Method

In this section, we describe methods for estimating ITEs and how transfer learning can improve ITE estimation in smaller samples using information from larger ones. We begin by introducing TARNet, then review the approach of Aloui et al. (2023) and related variants.

2.1 TARNet

We refer to $\tau(x)$ as the ITE, consistent with the machine-learning literature on which our approach builds. As in Shalit et al. (2017) and Aloui et al. (2023), we define the ITE for a unit with covariates x as

$$\tau(x) = \mathbb{E}[Y^1 - Y^0 \mid X = x], \tag{1}$$

the expected effect of treatment relative to control given covariates x. Following Imbens & Rubin (2015), $\tau(x)$ is equivalently referred to as the conditional average treatment effect (CATE); we retain the term ITE for continuity with this literature. Since $\tau(x)$ is not identified without additional assumptions, alternative ways to at least approximate this quantity are necessary. Particularly in situations where randomization to the treatment groups cannot be ensured, additional assumptions are needed.

TARNet and its balanced extension, the counterfactual regression (CFR), estimate ITEs from observational data by learning a representation in which the treated and control groups are distributionally similar. Because

the same unit can never be observed under both treatment and control, causal inference hinges on making the treated and control groups as comparable as they would be in a randomized trial. Simply fitting two separate outcome models, one for treatment group and one for control group, ignores this problem: if certain covariate profiles appear almost exclusively in the treated group, the control model has no information about them and counterfactual predictions become unreliable. TARNet mitigates this issue by jointly learning a representation function Φ and outcome predictions h. The representation function Φ maps the original covariates to a new space in which the distributions of treated and control units are more alike.

To ground their architecture theoretically, Shalit et al. (2017) first formalize the target quantity which is the expected precision in estimating heterogeneous effects ($\varepsilon_{\text{PEHE}}$) as squared difference between true ITE and estimated ITE:

$$\varepsilon_{\text{PEHE}}(h, \Phi) = \int_{\mathcal{X}} \big(\hat{\tau}(x) - \tau(x)\big)^2 \, p_F(x) \, dx, \tag{2}$$

with

$$\hat{\tau}(x) = h(\Phi(x), 1) - h(\Phi(x), 0). \tag{3}$$

where Φ is the shared representation function and h refers to the outcome predictions, which are applied to $\Phi(x)$ and indexed by treatment status.

Since $\varepsilon_{\text{PEHE}}$ involves the counterfactual outcome, it cannot be computed using observed data. Instead, two successive bounds on ITE estimation are derived:

$$\varepsilon_{\text{PEHE}}(\Phi, h) \leq 2\Big(\underbrace{\varepsilon_{CF}(h, \Phi)}_{\text{counterfactual loss}} + \underbrace{\varepsilon_F(h, \Phi)}_{\text{factual loss}} - 2 \cdot \underbrace{\sigma_Y^2}_{\text{variance of } y} \Big) \tag{4}$$

$$\leq 2\Big(\underbrace{\varepsilon_F^{t=0}(h, \Phi)}_{\text{factual loss } t=0} + \underbrace{\varepsilon_F^{t=1}(h, \Phi)}_{\text{factual loss } t=1} + \underbrace{\text{IPM}_G\big(\{\Phi(x)\}_{t=0}, \{\Phi(x)\}_{t=1}\big)}_{\text{distance between groups}} \Big). \tag{5}$$

There are two components in the first bound in equation 4; the first one is the counterfactual loss $\varepsilon_{CF}(h, \Phi)$ that captures the error in predicting outcomes which is unobservable and the factual loss $\varepsilon_F(h, \Phi)$ that measures how well the model predicts the outcomes we actually observe in the data. The equation implies that $\varepsilon_{\text{PEHE}}$ cannot be small unless both factual and counterfactual losses are small.

Since we cannot compute the counterfactual loss directly, the second upper bound in equation 5 is derived. This bound includes two major components: First, $\varepsilon_F^{t=0}(h, \Phi)$ and $\varepsilon_F^{t=1}(h, \Phi)$ are factual losses for the control and treatment groups, respectively. Second, an imbalance between treatment groups due to non-randomization is quantified by an Integral Probability Metric (IPM) that measures the distance between the representation spaces $\{\Phi(x)\}_{t=1}$ (treatment group) and $\{\Phi(x)\}_{t=0}$ (control group). The subscript G indicates a family of functions. Intuitively, each function $g \in G$ is like a test that tries to detect a difference between the treated and control representations. The IPM compares all these tests and keeps the largest difference that any of them can find. Therefore, $\text{IPM}_G(p, q) = 0$ only when none of the functions in G can tell the two groups apart that means the treated and control representations look the same to every function in G. In practice, the authors choose G to be the set of all 1-Lipschitz functions, and a classical result (Kantorovich–Rubinstein duality) states that when you use this particular G, the IPM becomes exactly the 1-Wasserstein distance, which is a standard measure of how far two distributions are in optimal-transport terms (Villani et al., 2009).

If both factual losses within treatment and control groups (the model predicts the observed outcomes well) and the IPM distance go toward zero, the counterfactual error will be small, too. By minimizing the upper bound, the algorithm thus effectively reduces both the observable prediction error and the risk of extrapolation errors due to covariate imbalance.

2.2 Estimation of upper bound

The TARNet learns to estimate the potential outcomes by minimizing the following likelihood[2]:

$$\mathcal{L}(\Phi, h) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} w_i \cdot \ell_{(\Phi,h)}\left(x_i, t_i, y_i\right)}_{\text{empirical loss}} + \kappa \cdot \mathrm{IPM}_G\left(\left\{\Phi\left(x_i\right)\right\}_{i:t_i=0}, \left\{\Phi\left(x_i\right)\right\}_{i:t_i=1}\right) \tag{6}$$

with

$$w_i = \frac{t_i}{2\nu} + \frac{1-t_i}{2(1-v)}, \quad v = \frac{1}{N}\sum_{i=1}^{N} t_i$$

The first term in equation 6 is a sample-weighted empirical loss: $\ell_{(\Phi,h)}(x_i, a_i, y_i)$ measures how well the model predicts the factual outcome y_i, that is, it minimizes the factual losses in equation 5. The second term minimizes any residual covariate imbalance after the observations have been mapped through the representation function Φ. Intuitively, minimizing this term pushes the two distributions in the representation space closer together. v indicates how many treated units exist, and w_i uses this to weight treated and control examples so that both groups influence the model equally despite imbalance.

The hyperparameter $\kappa > 0$ is a user-chosen balancing weight, which can also be optimized by tuning for the loss against the similarity of the treated and control representations. Setting $\kappa = 0$ yields the original TARNet loss, while $\kappa > 0$ results in the CFR (counter-factual regression) objective of (Shalit et al., 2017).

Taken together, the generalization bound shows that the unobservable error of an ITE estimator is controlled by two measurable training terms: the factual loss and the imbalance term, which measures how similar the treated and control feature distributions become under the learned representation. When the representation makes these distributions nearly indistinguishable, imbalance shrinks; when predictions are accurate, factual loss shrinks. Thus, if both are small, the counterfactual error, and therefore the treatment effect error, is also small. Training TARNet minimizes exactly these observable terms, tightening the bound and reducing the true ITE error $\varepsilon_{\text{PEHE}}(h, \Phi)$.

2.3 Extension of the TARNet with Transfer Learning (TL-TARNet)

Aloui et al. (2023) extend the TARNet into the domain of transfer-learning. They show that, for a source model applied to a target task, the target $\varepsilon_{\text{PEHE}}(h, \Phi)$ can be upper-bounded by the source factual losses plus three IPM terms that measure mismatch between source and target datasets. The following equation is defining $\varepsilon_{\text{PEHE}}(h, \Phi)$, which measures how well the model that was trained on the source ($\hat{f}^S$) predicts individual-level treatment effects on a new target population:

$$\varepsilon^T_{\text{PEHE}}(\hat{f}^S) = \mathbb{E}_{x\sim p^T_F}\left[\left(\tau^T(x) - \left[\hat{f}^S(x,1) - \hat{f}^S(x,0)\right]\right)^2\right] \tag{7}$$

where the subscripts T and S indicate target and source datasets, respectively. $\tau^T(x) = \mathbb{E}[Y^{T,1} - Y^{T,0} \mid X^T = x]$ is the true ITE in the target task and $\left(\hat{f}^S(x,1) - \hat{f}^S(x,0)\right)$ is the model's ITE prediction obtained by subtracting its two outcome predictions of the source dataset. Because the counterfactual outcome is never observed, the $\varepsilon_{\text{PEHE}}$ cannot be computed directly. However, it can be bounded using quantities that are observable. $\varepsilon_{\text{PEHE}}$ can be bounded in terms of factual and counterfactual loss as follows:

$$\underbrace{\epsilon^T_F}_{\text{observed error}} + u \cdot \underbrace{\epsilon^{T,a=0}_{CF}}_{\text{unobserved error}} \leq \underbrace{\epsilon^T_{\text{PEHE}}}_{\text{ITE error}}, \tag{8}$$

where the term u is the fraction of treated individuals in the target task and is used to weight the counterfactual loss.

[2]The Derivation of the formula can be found in Shalit et al. (2017).

Earlier transfer-learning methods for causal inference, such as those of Vo et al. (2022) and Bica & van der Schaar (2022), addressed only the distributional shift between source and target domains while presuming the same causal mechanism holds in both. In social domains applications, however, this assumption is not always fulfilled; covariate distributions can change, treatment propensities can vary, and the way treatments affect outcomes may differ across contexts. Different from previous studies, Aloui et al. (2023) confront these challenges simultaneously with the following bound for the counterfactual error in equation 8:

$$\epsilon_{CF}^{T}(\hat{f}) \leq \underbrace{\epsilon_{F}^{S}(\hat{f})}_{\text{source fit}} + \underbrace{\mathrm{IPM}_{G}(p_{F}^{T}, p_{F}^{S})}_{\text{distribution difference}} + \underbrace{\mathrm{IPM}_{G}(p_{F}^{T}, p_{CF}^{T})}_{\text{treatment-selection bias}} + \underbrace{\mathbb{E}_{(x,a)\sim p_{F}^{S}}\left[\left|f^{S}(x,a) - f^{T}(x,a)\right|\right]}_{\text{causal-mechanism change}} \tag{9}$$

Their upper bound on the target counterfactual loss isolates three distinct contributions – source-to-target distributional shift, treatment-selection bias within the target dataset, and changes in the causal mechanism between source and target dataset which is making it clear how each source of mismatch inflates the error. As a consequence, $\varepsilon_{\text{PEHE}}$ can be written in the following form:

$$\epsilon_{PEHE}^{T}(\hat{f}) \leq 4 \cdot \epsilon_{F}^{S}(\hat{f}) + 4 \cdot \mathrm{IPM}_{G}(p_{F}^{T}, p_{F}^{S}) + 2 \cdot \mathrm{IPM}_{G}(p_{F}^{T}, p_{CF}^{T}) + 4 \cdot \mathbb{E}_{(x,a)\sim p_{F}^{S}}\left[\left|f^{S}(x,a) - f^{T}(x,a)\right|\right]. \tag{10}$$

The bounding equation of $\varepsilon_{\text{PEHE}}^{T}$ does not contain the counterfactual loss explicitly; instead, it is bounded by four observable terms. When the source factual loss, the source–target distributional mismatch, the treatment-selection bias in the target dataset, and the causal-mechanism discrepancy are all small, the counterfactual loss must also be small, and the overall error remains low. The following assumptions need to be satisfied to ensure that the causal transfer learning is valid:

- Assumption 1: The representation space Φ is injective (Φ^{-1} exists on $\mathrm{Im}(\Phi)$), which means that we are preventing any information loss.
- Assumption 2: There exists a real function space G on $\mathrm{Im}(\Phi)$ such that the function $r \mapsto \ell_{\Phi,h}^{T}(\Psi(r), a, y) \in G$. This assumption ensures differences in representation distributions (treatment vs. control, or source vs. target) are a good proxy for differences in the representation-level loss.
- Assumption 3: There exists a function class G' on $\mathcal{Y}$ such that $y \mapsto \ell_{\Phi,h}(x, a, y) \in G'$. It guarantees that differences between source and target outcome distributions translate into proportionally controlled changes in the loss.

Supposing that Assumptions 1, 2 and 3 hold, performance of the model which is transferring causal knowledge from the source to the target dataset, $\varepsilon_{\text{PEHE}}(h, \Phi)$ is upper bounded as follows:[3]

$$\begin{aligned}\varepsilon_{\text{PEHE}}^{T}(\Phi, h) \;\leq\; 2\Big(& \underbrace{\varepsilon_{F}^{S,t=1}(\Phi, h) + \varepsilon_{F}^{S,t=0}(\Phi, h)}_{\text{factual loss source}} + \underbrace{\mathrm{IPM}_{G}\left(\left\{\Phi\left(x_{i}^{T}\right)\right\}_{t=1}, \left\{\Phi\left(x^{S}\right)\right\}_{t=1}\right)}_{\text{distance target-source } t=1} \\ & + \underbrace{\mathrm{IPM}_{G}\left(\left\{\Phi\left(x^{T}\right)\right\}_{t=0}, \left\{\Phi\left(x^{S}\right)\right\}_{t=0}\right)}_{\text{distance target-source } t=0} + \underbrace{\mathrm{IPM}_{G}\left(\left\{\Phi\left(x^{T}\right)\right\}_{t=0}, \left\{\Phi\left(x^{T}\right)\right\}_{t=1}\right)}_{\text{distance between groups in target}} \\ & + 2\gamma^{*}\Big).\end{aligned} \tag{11}$$

When the model trained on a source domain is applied to a target domain $\varepsilon_{\text{PEHE}}^{T}(\Phi, h)$ from the target dataset is controlled by five observable quantities. The first two are the factual losses from the source dataset both for the treated and control units ($\varepsilon_{F}^{S,t=1}(\Phi, h), \varepsilon_{F}^{S,t=0}(\Phi, h)$). The next three terms are IPM distances: two compare the treated-treated and control-control distributions across the source and target domains, while the third is identical to the treated-control distance within the target domain in TARNet. Finally, the residual γ^{*} is defined as the expected IPM between the source and target outcome distributions conditional on X. It captures any irreducible mismatch in causal mechanisms:

$$\gamma^{*} = \mathbb{E}_{x\sim p_{F}^{S}}\left[\mathrm{IPM}_{G'}\left(P(Y_{t}^{S} \mid x),\, P(Y_{t}^{T} \mid x)\right)\right] \tag{12}$$

[3]Derivation of the formula can be found in Aloui et al. (2023)

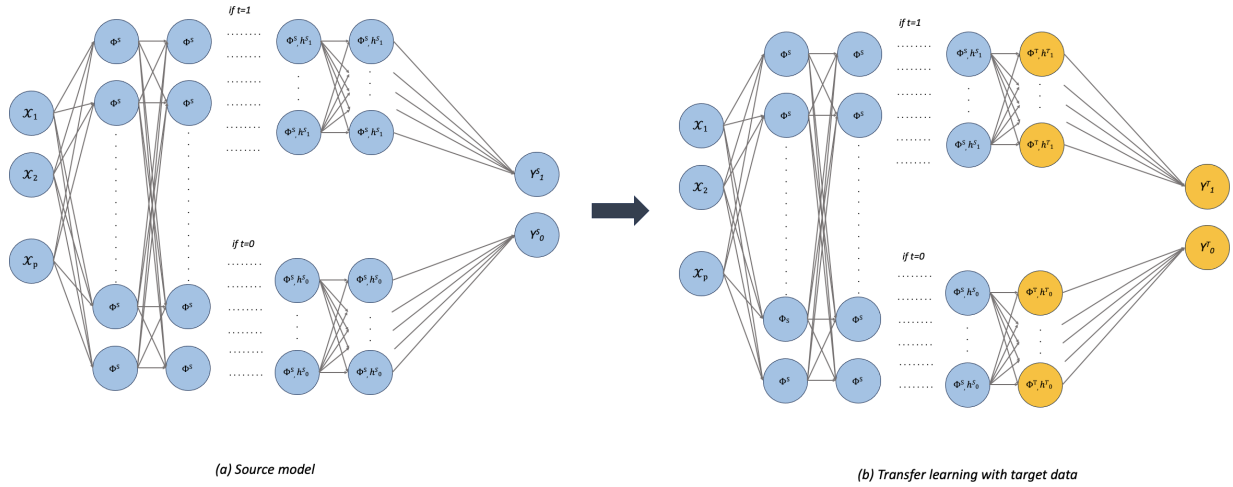

Figure 1: Representation of transfer learning.

Minimizing all five terms will lead to a successful transfer of the causal mechanisms. The resulting bound naturally reduces to the original TARNet bound when the source and target coincide, but it shows explicitly which divergences must be small for a transfer to succeed.

2.4 Implementation of h and Φ in TL-TARNet

Training begins by optimizing the entire TARNet as one shared encoder Φ and followed by two outcome heads h_0, h_1. A shared multi-layer encoder first transforms every covariate vector into a latent embedding; on top of that embedding sit two separate heads, one dedicated to predicting the control outcome and the other the treated outcome, so each training example only updates the head that matches its observed treatment arm. During learning the optimizer repeatedly takes mini-batches, computes the prediction loss on the factual outcomes, re-weights that loss to correct for any class-size imbalance, and simultaneously adds a penalty that measures how different the treated and control embeddings are using an IPM such as Wasserstein distance. During each iteration, the factual loss is back-propagated through both Φ and h_0, h_1 and the process repeats until convergence. In this way Φ is encouraged to produce a representation where the two treatment groups overlap, while h_0, h_1 specialize on outcome prediction.

Panel (a) of Figure 1 shows the source model once it has been fully trained; a shared encoder Φ^S (blue circles) transforms the covariates, and two specialized heads h_0^S and h_1^S map that representation to the potential outcomes $Y^{S,0}$ and $Y^{S,1}$.

Panel (b) illustrates the transfer-learning stage. All parameters up to a chosen depth-including the entire encoder and, if desired, part of each head are copied unchanged into the new model and kept fixed (still shown in blue). Only the remaining block (colored in yellow in our example) is re-initialized and left trainable; this lightweight section, denoted Φ^T, h_0^T, h_1^T, is then fine-tuned on the smaller target dataset. Freezing the early layers preserves the rich features learned from the source task, while updating the last one or two layers (depending on the target dataset size) lets the model adjust to the specific distribution and label characteristics of the target domain.

A detailed version of transfer learning training with TARNet in two stages is illustrated in Figure 2.

In the transfer learning stage, the network undergoes fine-tuning using a smaller target dataset. At this part, weights learned from the source dataset are transferred, and training continues exclusively with the

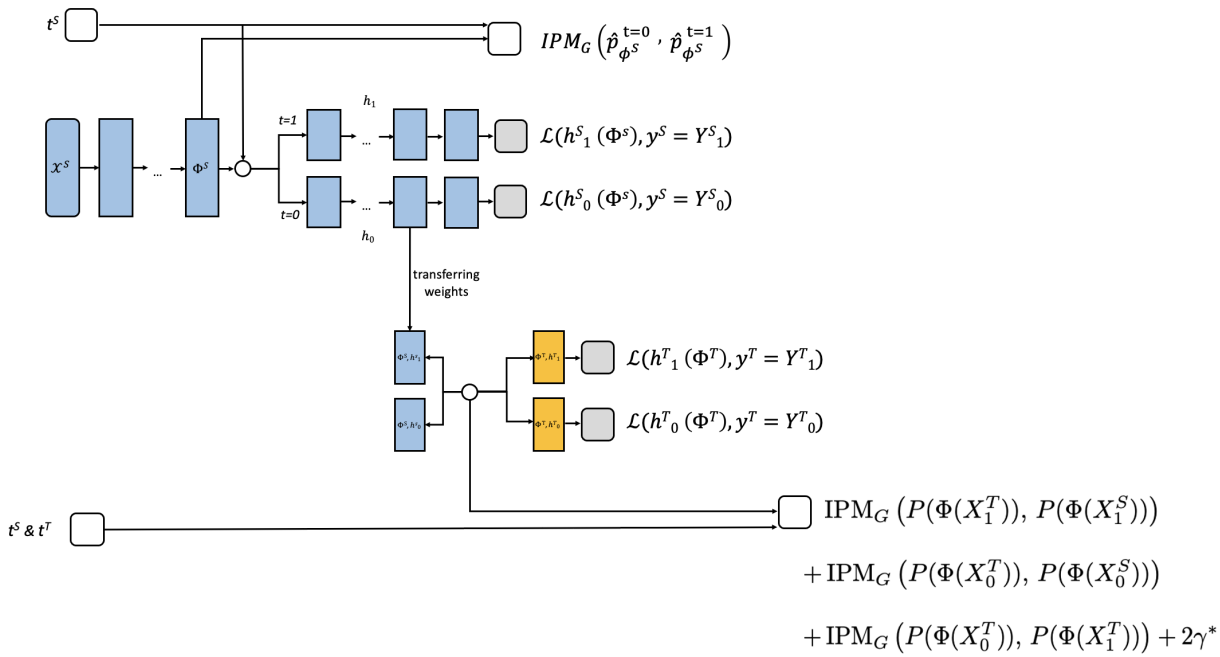

Figure 2: Representation of transfer learning with IPMs.

smaller target dataset, with an emphasis on reducing discrepancies in representation spaces and minimizing the target dataset factual loss. Specifically, there are three IPM components: (i) aligning treated versus control representations within the target dataset, (ii) aligning treated group representations between source and target datasets, and (iii) aligning control group representations between source and target datasets. Gradients update the encoder Φ, progressively shrinking discrepancies both within and across domains. Simultaneously, outcome heads h_0, h_1 adapt to align with target-specific representations. This targeted fine-tuning process typically lasts for only a few epochs and affects selected network layers (usually a few layers), allowing the model to retain generalizable features learned from the source while adapting to the new target task.

2.5 Causal Inference Task Affinity (CITA)

When several source datasets are available, the first challenge is to determine which source is worth transferring to the new task. CITA addresses this need by quantifying task affinity between source and target datasets. As shown by Aloui et al. (2023), CITA is strongly correlated with counterfactual loss; source tasks that are closer to the target in terms of CITA tend to exhibit smaller counterfactual loss on the target data. Hence, low CITA scores indicate more suitable source datasets for transfer, whereas high CITA scores point to poorer task alignment. Reliable use of CITA, however, presumes that the source model has learned a sufficiently good representation of its own task (Aloui et al., 2023).

After training a TARNet on the source dataset, we keep the learned weights fixed and calculate two Fisher information matrices one using the same source data $F^{S,S}$ and one using the target dataset $F^{S,T}$. For a network N_θ trained on source data D^S and a loss $L(\theta, D)$, the Fisher matrices are then

$$F^{S,S} = \mathbb{E}_{D\sim D^S}\left[\nabla_\theta L(\theta^S, D)\, \nabla_\theta L(\theta^S, D)^\top\right] \tag{13}$$

$$F^{S,T} = \mathbb{E}_{D\sim D^T}\left[\nabla_\theta L(\theta^S, D)\, \nabla_\theta L(\theta^S, D)^\top\right]. \tag{14}$$

Each matrix is the expected outer product of the loss gradients with respect to the network's parameters, providing a weight-by-weight measure of sensitivity.

The task affinity score is then the Frobenius norm of the difference. The Frobenius norm, which is the Euclidean length of a matrix when viewed as a vector, measures the overall magnitude of the discrepancy across all parameters.

$$d^{S,T} = \frac{1}{\sqrt{2}} \left\| (F^{S,S})^{1/2} - (F^{S,T})^{1/2} \right\|_F, \tag{15}$$

The task affinity is the difference between the principal square-roots of these matrices and values of $d^{S,T}$ near 0 indicate high similarity between the causal tasks, whereas values near 1 signal marked dissimilarity (additional details on this measure when treatment labels are switched are provided in Appendix B).

2.6 Advantages and limitations of the approach

The transfer framework of Aloui et al. (2023) provides a principled way to reuse causal models trained on larger datasets for smaller but similar datasets. It begins with a well-trained network on a source dataset and fine-tunes only on the target sample, which helps control overfitting and variance which are the common pitfalls in small clinical or social science studies. The target error is decomposed into three observable gaps which are covariate shift, treatment-selection bias, and changes in the causal mechanism and the method is designed to shrink all three simultaneously. Prior transfer-learning approaches typically addressed only one of these (distribution shift, treatment imbalance, or mechanism drift) in isolation; Aloui et al. (2023)'s unified treatment is thus especially useful in social sciences and health settings where policy, prescribing, and population mix often change together. Importantly, the upper-bound guarantees rely only on factual data and a frozen source network, so no unobservable counterfactual labels are required which is making the method practical in real-world deployments.

CITA complements this framework by offering a fast, pre-transfer measure of task closeness. Its score is label-invariant: by minimizing distance over all permutations of treatment labels, tasks that differ only by a relabeling receive distance 0. Because the score correlates closely with the counterfactual error observed after fine-tuning, practitioners can select the most compatible source dataset in advance and avoid inefficient transfer.

There are also some limitations in the Aloui et al. (2023)'s work. From a practical perspective, estimating diagonal Fisher matrices on both source and target dataset (and on each label permutation) means extra computation on top of transfer learning training and storage of large parameter-sized tensors, which can be costly in terms of time and computation. Secondly; it should be highlighted that the Fisher matrices that drive both the theoretical bounds and CITA must come from source models that are already error accurate on their own tasks. Poorly fitted sources produce meaningless distances and can mislead the selection process.

2.7 Extensions and alternative approaches

During the last about five years, several alternative approaches were developed with a similar aim to transfer information from a source to a target dataset.

Bica & van der Schaar (2022) address the challenge of improving ITE estimation in a target domain by leveraging information from a source domain with a different feature space. This heterogeneous transfer learning problem is particularly relevant in fields like healthcare, where treatment effects need to be estimated for new patient populations with different clinical covariates and limited data availability. To tackle this, the authors propose a framework that combines representation learning to align heterogeneous feature spaces with a flexible multi-task architecture. They extend standard ITE learners to their transfer learning counterparts and evaluate their approach on a new semi-synthetic benchmark designed for heterogeneous transfer learning. Their results demonstrate significant performance improvements over traditional methods while offering insights into how different transfer strategies affect causal effect estimation across domains. Compared to Aloui et al. (2023), their approach only allows for variations in the covariate distribution, but not for changes in the causal mechanisms.

Similarly, Aglietti et al. (2020) construct a joint model that integrates interventional source data with observational target data. They employ the concept of transportability, which lets researchers use observational

data from the source population for the target population, even in the absence of randomized trials in the target group.

Along these lines, Vo et al. (2022) investigate ways to enhance the estimation of treatment effects within a specific target population. They achieve this by using diverse yet related data sources, while considering the potential differences in the distributions between the source and target populations. Also, they present three levels of knowledge transfer by modeling the outcomes, treatments, and confounders ensuring a consistent transfer. For their approach, they assume that DAGs have to be identical, which is crucial because they focus on observational data only.

Wei et al. (2023) introduce a procedure based on generalized linear models (GLMs) with penalization for situations with many covariates. They design an approach for scenarios where source and target domains share most of the same covariate space. The proposed method, incorporates regularization with transfer learning for estimating nuisance parameters, such as propensity scores, and integrates these estimates into standard causal effect estimators like inverse probability weighting.

Chauhan et al. (2024) propose a HyperITE via soft weight sharing that allows dynamic information exchange between treatment groups. This approach is not a classic transfer learning approach because source and target dataset stem from the same dataset. They use a mirror approach were the experimental group provides the source dataset for the target control group and vice versa. Experimental results show that HyperITE reduces ITE estimation error, with its benefits becoming more pronounced as dataset size decreases.

Lastly, it is important to distinguish between transfer learning and transportability, a concept developed in Pearl (2011). While both concepts aim at reusing causal knowledge across settings, they differ in their assumptions. Transportability relates to the questions when and how causal conclusions (mechanisms or effects) can be validly moved from one population to another. Here statistical alignment alone is insufficient; external validity is central (Pearl & Bareinboim, 2022). Sound transport requires explicit assumptions about which mechanisms are invariant and which change across environments which are often formalized with selection diagrams. Violations of these assumptions can yield biased or invalid conclusions (Bareinboim & Pearl, 2012; 2013). Causal transfer learning, in contrast, seeks to transfer estimated causal effects to a new setting and does not require the entire causal graph to match, as long as the effect of interest remains invariant (Aloui et al., 2023).

3 Simulation Study

We evaluate the transfer-learning framework in two simulations and one empirical application, each designed to isolate a different term of the transfer bound in Section 2.3 (Eq. 11). That bound decomposes the target error ϵ_{PEHE}^{T} into the source factual loss, two source–target IPM distances that measure covariate shift (treated–treated and control–control), a within-target IPM that measures treatment-selection bias, and a residual γ^* that measures how far the causal mechanism itself has moved between source and target. The first simulation draws the target by subsampling the source, so that the covariate-shift and mechanism terms are zero by construction and only the target sample size and the within-target selection bias remain active; it asks how small the target can be and whether transfer corrects a violation of ignorability. The second simulation switches the covariate-shift and mechanism terms on, one at a time, drawing the target from a population whose covariates, or whose causal mechanism, differ from the source, and traces how the error and bias of transfer respond as each mismatch grows. The empirical application then supplies a genuinely distinct target for which no ground truth is available, where several terms are active at once. Each component of the study thus isolates a different part of the same bound, and together they move from the most controlled setting to the least.

The datasets were generated in R, and the transfer learning models were implemented in Python using the Torch library. All data and analysis code is available in the following link: `https://anonymous.4open.science/r/TL_TARNet-256A/README.md` .

3.1 Simulation 1: source-target alignment

We first evaluate how transfer learning of ITEs behaves in settings where target samples are small and treatment assignment in the target is often non-randomized. To our knowledge, TL-TARNet has not previously been evaluated under these conditions. The simulation is designed to answer two questions: (i) what source and target sample sizes are needed for transfer to help, and (ii) whether transfer can reduce bias when the target sample violates the ignorability assumption. Because we generate the target by subsampling the source, the source and target share the same covariate distribution and the same outcome model; the simulation therefore isolates the effects of small target sample size and of treatment-selection bias in the target, and it does not vary source–target covariate shift or changes in the causal mechanism. We return to those dimensions in Simulation 2.

3.2 Data Generation

Source Data Source data were generated for the following model for the outcome:

$$y_i = \alpha + \beta T_i + \boldsymbol{\gamma}\mathbf{X}_i + \boldsymbol{\omega} T_i \mathbf{X}_i + \epsilon_i \tag{16}$$

where $\alpha = 0$ is an intercept, $T_i \sim \text{Bern}(0.5)$ is a binary treatment, $X_i \sim (0, \Sigma)$ collects five covariates whose covariance Σ is an equicorrelation matrix with unit variances and pairwise correlation 0.5 (i.e. 1 on the diagonal and 0.5 off it), and $\epsilon_i \sim (0, \sigma^2)$ is a residual. The coefficient $\beta = 1$ sets the average treatment effect (ATE), γ collects the prognostic effects of the covariates on the outcome, and ω induces treatment heterogeneity, so that the individual treatment effect depends on the covariates, $\tau(X_i) = \beta + \omega X_i$. We set $\gamma = (0.3, \ldots, 0.3)$ and $\omega = (0.05, \ldots, 0.05)$, and the residual variance to $\sigma^2 = 1$. With X_i mean-zero the population ATE is $\mathbb{E}[\tau(X)] = \beta = 1$. In the source the ignorability assumption $(Y_i^1, Y_i^0) \perp T_i \mid X_i$ holds by construction, because treatment is assigned at random.

Target Data Based on the source data, we created smaller target data sets by subsampling. Under the condition of randomized sampling, we randomly drew N^T persons from the source, thus reflecting the same population. Under the condition of non-randomized sampling, we actively violated the ignorability assumption: Persons with a high potential outcome Y^0 under the control condition were more likely to appear in the treatment arm. Formally, each person received a probability[4] $p_i^T := P(T_i = 1|Y_i^0) = \text{expit}(Y_i^0)$. Because p^T increased with Y^0, the resulting datasets were skewed toward high-Y^0 treated subjects and low-Y_0 controls. This mimics self-selection into treatment by individuals who would have scored higher even without treatment (for example, heavier smokers being more likely to enroll in a smoking-cessation programme).

Data conditions We generated large source data sets $N^S = 1000, 5000, 10000, 20000$ and 30000 persons which is a range reflected smaller data sets as well as large-scale data sets. For the target data sets we included sample size of $N^T = 50, 100, 250$ and 500 persons. This included typical sample sizes for smaller experiments, as well as large (quasi)-experimental settings. For each condition $R = 100$ data sets were generated.

3.3 Data analysis

For the data analysis, we used two approaches: The original TARNet that only used the target data and the extended TL-TARNet that first trained a model on the source data and then analyzed the target data sets with the pre-trained models which were explained above. We employed the same general neural network architecture across all simulations, but we adjusted its complexity depending on the dataset size. Because the number of trainable parameters increases with network depth and width, larger datasets can support more complex architectures, whereas smaller datasets require reduced capacity to avoid over-parameterization.

The TARNet architecture consists of a shared encoder and two separate heads for predicting potential outcomes under control and treatment. The encoder maps the input through three fully connected layers, with rectified linear unit (ReLU) activation functions applied after each layer. This shared representation

[4] $\text{expit}(z) = 1/(1 + e^{-z})$ is the logistic link.

is then passed to two parallel networks (control and treatment), which model the outcome for the control and treatment groups, respectively. Each head consists of hidden layers followed by a final linear layer that outputs the prediction. We use ReLU in all hidden layers to introduce nonlinearity while keeping gradient-based optimization stable and computationally efficient, which is standard practice for feedforward networks in this setting.

In this design, it is important to calculate the number of trainable parameters. Over-parameterization, where the number of parameters exceeds the number of observations, can lead to overfitting or unstable training. Therefore, we constrained all architectures so that the number of trainable parameters was on the order of one tenth of the dataset size. For example, for a source dataset with $10,000$ observations and 5 input features, we used an encoder with 3 layers of 16 neurons each, and 2 hidden layers with 8 neurons in both the treatment and control heads. In this case, the total number of estimated parameters is $1,074$, which is suitable for training with $10,000$ observations (see, e.g., Kavzoglu & Mather, 2003).

We selected the learning rate and number of epochs based on the observed convergence behavior of the training loss. For the source dataset with $10,000$ observations, a learning rate of 0.005, 700 epochs, and batch size of 32 resulted in stable convergence in the training. Given the relatively small dataset size, we used a small batch size, which is often preferable for optimization and effective use of limited data.

After training on the source data, transfer learning is performed in two phases. In the first phase, we train the model to minimize the distance between the source and target distributions. In the second phase, we minimize the prediction loss on the target data. In both phases, it is important to choose the learning rates carefully. During the minimization of the distance, we use a small learning rate and a larger number of epochs. For instance, for a non-randomized intervention target dataset with 250 observations and a source dataset with 10,000 observations, a learning rate of 10^{-5} with 1,500 epochs led to convergence in both phases.

3.4 Simulation 1: results

We assessed whether transferring a representation learned on the source improves ITE estimation on the target, and how any gain depends on the source sample size and on the presence of selection bias in the target. For each condition we summarize the $R = 100$ replications by two quantities: the mean estimated ITE (averaged over individuals and replications) and the mean $\varepsilon_{\text{PEHE}}$. We also report dispersion of the predicted ITEs in the Appendix. We omit the setting with a source of 1,000 and a target of 500, because there the target would constitute a large fraction of the source. Across the conditions reported below, the randomized target data sets have an average CITA of 0.18 and the non-randomized (biased) target data sets an average CITA of 0.24.

Average ITE Figure 3 illustrates the results for the mean ITE estimates. The left panel with randomized intervention shows no bias in the mean ITE, irrespective of source or target sample size. The non-randomized target (right panel) violates ignorability. Without transfer (TARNet), the mean ITE is heavily biased, at roughly 2, and this bias does not diminish as the target sample grows. With transfer (TL-TARNet), the estimate is shrunk toward the source model, in which ignorability holds by construction. Also the bias is smaller when the source sample is larger.

Precision of ITE estimates Figure 4 shows the expected precision in estimating ITEs ($\varepsilon_{\text{PEHE}}$).

In the left panel (randomized target), transfer learning substantially improves precision over target-only TARNet, with the largest gains at the smallest target sizes, where TARNet's error is high. As the target grows, TARNet's PEHE falls steeply, and by $N_T = 500$ the two are comparable. In the right panel, where ignorability is violated, target-only TARNet is far more imprecise throughout, as expected. TL-TARNet is markedly better. Here, unlike the randomized case, its precision degrades somewhat as the target grows — with more biased target data the model leans further toward the target and re-imports part of its bias and this degradation is milder for the larger sources.

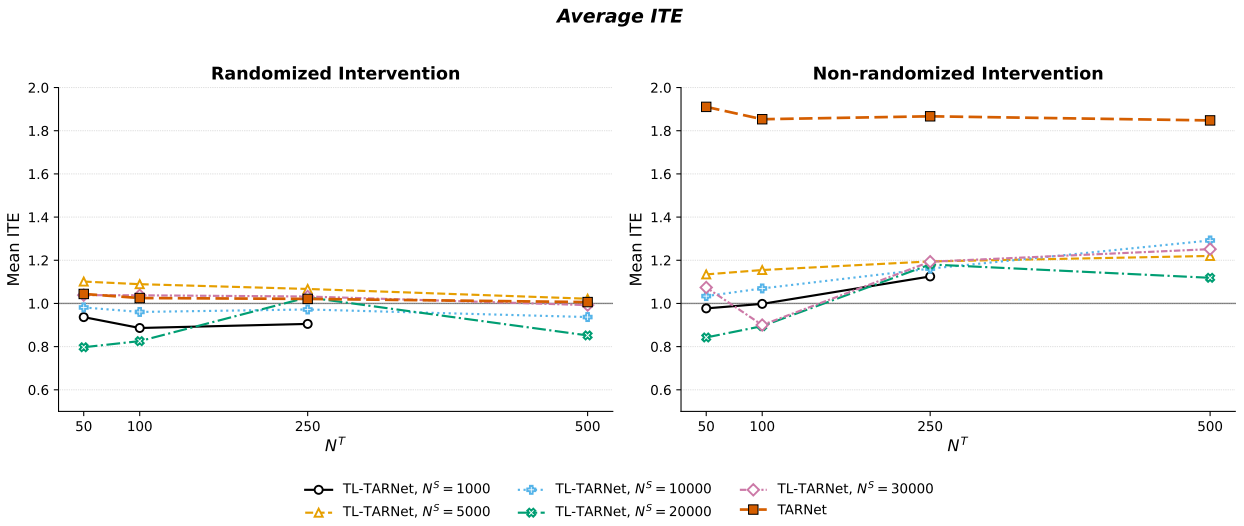

Figure 3: Mean ITE for TARNet and TL-TARNet models (i.e. with and without transfer learning) across different sizes of source dataset and target data sets. The population value is indicated with horizontal line at ($\tau = 1$).

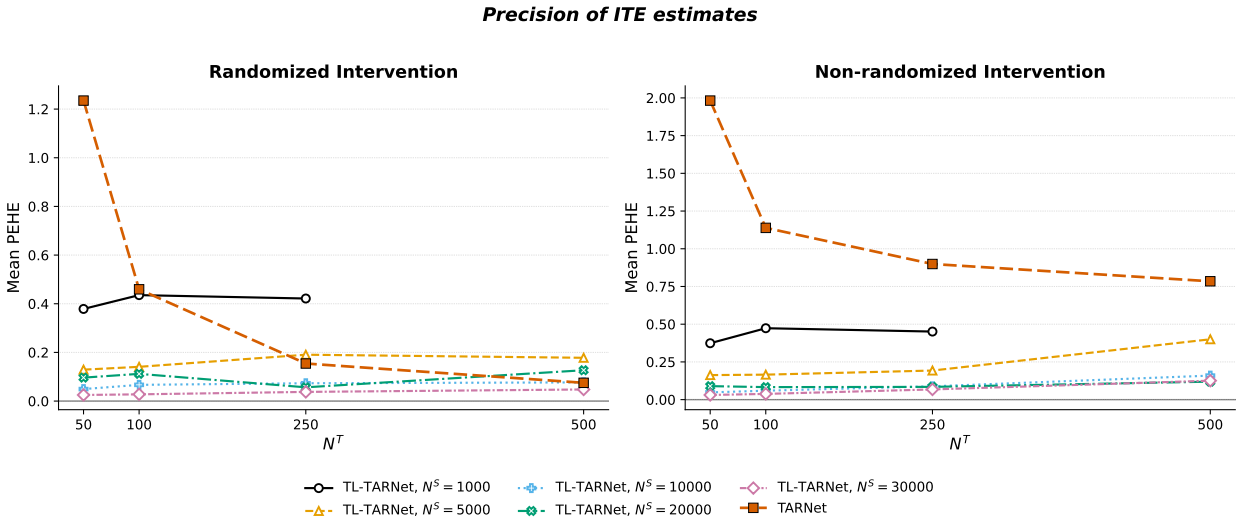

Figure 4: Mean $\varepsilon_{\text{PEHE}}$ for TARNet and TL-TARNet models (i.e. with and without transfer learning) across different sizes of source dataset and target data sets. No error is indicated by the horizontal line at zero.

Summary Two insights stand out. When the target is randomized, transfer learning and TARNet are comparable in bias, but TL-TARNet offers better precision for the ITEs. When the target is non-randomized, transfer learning substantially improves both bias and precision.

3.5 Simulation 2: source–target mismatch

The first simulation generated the target by subsampling the source, so that the source and target shared the same covariate distribution and the same outcome model. That design isolated the effects of small target sample size and of treatment-selection bias. In this simulation we will check cases where a target whose covariates are distributed differently from the source, and a target whose causal mechanism differs from the source. These are exactly the two quantities that enter the upper bound on the target error in Aloui et al. (2023): an IPM term that measures source–target covariate shift, and a residual term γ^* that measures the difference between the two causal mechanisms. The simulation reported here switches them on, one at a time, and traces how the error and bias of transfer learning change as each kind of mismatch grows. It also adds a third estimator, a frozen source-only model (introduced below), which is informative here, unlike in Simulation 1, precisely because the target now differs from the source.

3.5.1 Data Generation

Outcome model The first simulation used a linear outcome model (Equation 16). Because TL-TARNet relies on a neural network whose advantage is its ability to represent nonlinear relationships, we generated the outcomes from a model that adds nonlinear terms to Equation 16:

$$y_i = \alpha + \beta T_i + \gamma X_i + \omega T_i X_i + \lambda\big[g(X_i) + T_i\, q(X_i)\big] + \varepsilon_i, \tag{17}$$

where $g(X_i) = \sin X_{i1} + X_{i2}X_{i3} + X_{i4}^2$ described a nonlinear prognostic surface (how the covariates shaped the outcome regardless of treatment), and $q(X_i) = X_{i1}X_{i2} + \cos X_{i3}$ described nonlinear treatment heterogeneity (how the covariates shaped the size of the effect). The scalar λ controlled the degree of nonlinearity; setting $\lambda = 0$ recovered the linear model of Equation 16, and we used $\lambda = 1$ throughout. As before, $\alpha = 0$ and $\varepsilon_i \sim (0, \sigma^2)$. The coefficient $\beta = 1$ sets the constant component of the treatment effect, while ω and $q(.)$ introduce heterogeneity so the individual treatment effect is $\tau(X_i) = \beta + \omega X_i + \lambda\, q(X_i)$. Note that the prognostic part of the model (the terms with γ and g) entered both potential outcomes equally and so cancelled in the ITE; only β, ω, and q determined the effect itself. This distinction matters below.

Source Data. The source played the role of the clean, well-understood reference data set. For every source we drew the covariates from $X_i \sim (0, I)$ with five covariates, assigned treatment at random, $T_i \sim (0.5)$, and used the reference coefficients $\gamma_S = \omega_S = (0.5, \ldots, 0.5)$. The ignorability assumption therefore held in the source by construction. As in the first simulation we generated one source data set for each source size and reused it across all conditions.

Target Data: covariate shift. In the first kind of mismatch the kinds of people in the target differed from the source, but the outcome rule stayed the same which is "different people, same rule." We drew the target covariates from a shifted and correlated normal distribution,

$$X_i^T \sim \big(\delta\mathbf{1},\ (1-\rho)I + \rho\,\mathbf{1}\mathbf{1}\big), \tag{18}$$

where δ shifted the mean of every covariate and ρ introduced correlation between covariates, while the marginal variances stayed at one. The outcome coefficients were left at their source values, so the rule connecting covariates to outcomes was unchanged and only the population moved. A single mismatch level $c \in \{0, 0.25, 0.5, 0.75, 1\}$ set $\delta = c$ and $\rho = 0.6\,c$, so that $c = 0$ reproduced the source population and $c = 1$ was the most shifted. The size of the shift was summarized by the Kullback–Leibler divergence between the target and source covariate distributions,

$$\mathrm{KL}\big(P_X^T \,\|\, P_X^S\big) = \tfrac{1}{2}\Big[k\delta^2 - (k-1)\log(1-\rho) - \log\big(1 + (k-1)\rho\big)\Big], \tag{19}$$

with $k = 5$. This condition activated only the covariate-shift (IPM) term of the bound; because the outcome rule was unchanged, the mechanism term γ^* remained zero.

Equation 19 is the Kullback–Leibler divergence between two Gaussians, specialized to the present design. The general expression reduces to three contributions. The trace term vanishes because every covariate keeps unit marginal variance, so $\mathrm{tr}(\Sigma) = k$; only the mean and the correlation change, not the scale of any covariate. The term $k\delta^2$ is the mean shift—the squared distance of the target mean $\delta\mathbf{1}$ from the source origin, δ^2 per covariate and k of them—which grows quadratically in δ. The two logarithmic terms come from the determinant of the equicorrelation matrix, $\det\Sigma = (1-\rho)^{k-1}\big(1+(k-1)\rho\big)$, and isolate the effect of the correlation ρ; they vanish when $\rho = 0$. The divergence is therefore zero at $c = 0$, where $\delta = \rho = 0$ and the target coincides with the source, and increases monotonically with c, giving a single scalar that orders the covariate-shift conditions by distance from the source.

Target Data: mechanism change. In the second kind of mismatch the people were drawn from the same distribution as the source ($\delta = \rho = 0$) but the outcome rule itself changed which can be thought as "same people, different rule." We moved the target coefficients by an amount $m \in \{0, 0.25, 0.5, 0.75, 1\}$ along a fixed unit direction $v = (1, -1, 1, -1, 1)/\sqrt{5}$. Because γ and ω play different roles, we moved them separately, giving two conditions. In the effect-change condition we moved the effect-modification coefficients, $\omega^T = \omega^S + mv$ with $\gamma^T = \gamma^S$, which genuinely changes the individual treatment effect—the quantity being estimated. In the prognostic-change condition we instead moved the prognostic coefficients, $\gamma^T = \gamma^S + mv$ with $\omega^T = \omega^S$, which changes the baseline outcome surface but, because γ cancels in the ITE, leaves the treatment effect unchanged. The prognostic-change condition is a deliberate control: the target's effect is identical to the source's, yet the source model is still miscalibrated on the target's outcomes, so it isolates whether transfer is disturbed by a change that does not alter what we are estimating. Both conditions leave the covariate distribution untouched and so act through the mechanism part of the bound rather than the covariate-shift term.

Treatment assignment. As in the first simulation, the target intervention was either randomized or non-randomized. Under randomization we set $T_i \sim (0.5)$. Under non-randomization we again let persons with a high control outcome be more likely to be treated, $P(T_i = 1 \mid Y_i^0) = (z_i)$,[5] but with z_i the standardized control outcome rather than Y_i^0 itself. Standardizing kept the strength of the selection the same across all conditions, so that changing the nonlinearity or the mechanism did not also, and unintentionally, change how much selection bias was introduced.

Data conditions. We generated source data sets of $N^S = 1000$ and 5000 persons and target data sets of $N^T = 50, 100$ and 500 persons. Each of the three mismatch conditions (covariate shift, effect change, prognostic change) was run at the five mismatch levels above, under both randomized and non-randomized assignment. For each condition we generated $R = 100$ data sets. These numbers are chosen because of our conclusion from Simulation 1.

3.5.2 Data analysis

We compared three estimators on every target data set: target-only TARNet, trained on the target alone; TL-TARNet, which fine-tunes the pre-trained source model on the target; and a source-only model, which applies the pre-trained source model to the target without any fine-tuning. The source-only baseline isolates the contribution of the fine-tuning step: when the target is small or biased, part of TL-TARNet's advantage could be shrinkage toward the large unbiased source rather than adaptation, and comparing against the frozen source separates the two. This baseline is informative here precisely because the target now differs from the source; where the target is drawn from the source, as in Simulation 1, a frozen-source model would merely re-evaluate the source on its own population.

For each data set we recorded ϵ_{PEHE} and the mean bias of the estimated ITE, always against the target's own true ITE rather than the source's. Because both kinds of mismatch also change how spread out the true treatment effects are, and ϵ_{PEHE} is measured on that scale, we report a normalized error, $\epsilon_{PEHE}/\mathrm{Var}(\tau)$, where $\mathrm{Var}(\tau)$ is the variance of the true ITE in the target. A value near one corresponds to a method

[5] $(z) = 1/(1 + e^{-z})$ is the logistic link.

no better than assigning the average effect to everyone; smaller is better, and zero is exact. Raw error is discussed only where it changes the interpretation.

Finally, we use CITA as a common axis. The two mismatch conditions are measured in different units—a covariate-shift level c and a mechanism change m are not directly comparable—but both produce a CITA value between source and target, so the error of each method can be plotted against CITA regardless of which mismatch produced it.

The decomposition suggests a clear prediction, though one that must be read carefully. Fine-tuning can absorb a covariate shift, because the relationship being transferred is still correct and only has to be realigned to a new covariate region. A change in the causal mechanism is more subtle. A prognostic change perturbs both potential-outcome surfaces equally and so cancels in the contrast τ, leaving the estimand intact; an effect change perturbs only the treated surface, alters τ itself, and cannot be undone on a small target. We therefore expected an effect change to degrade transfer while a covariate shift and a prognostic change would not. We read off the value of CITA at which TL-TARNet stops outperforming target-only TARNet as a practical threshold for when borrowing from a source is worthwhile.

3.6 Simulation 2: results

Transfer learning and sample size. Table 1 reports the normalized error by sample size. The size of the source mattered more than the size of the target. Increasing the source from 1000 to 5000 observations more than halved the error of transfer learning (0.32 to 0.13; lower panel of Table 1), whereas the target-only error was, as it must be, unchanged, since that model never uses the source. A larger source trains a better representation to hand over: the source-only model itself achieved a normalized error of about 0.13 at $N_S = 5000$ against about 0.37 at $N_S = 1000$.

Table 1: Normalized error ($\varepsilon_{\text{PEHE}}$ divided by the variance of the true ITE) by sample size, averaged over the remaining factors, as mean (SE) over the $R = 100$ replications per condition. Lower is better; a value near one is no better than predicting the average effect.

	target-only	source-only	TL-TARNet
Target size N^T			
50	2.420 (0.031)	0.250 (0.003)	0.241 (0.003)
100	1.015 (0.012)	0.246 (0.003)	0.232 (0.003)
500	0.427 (0.005)	0.236 (0.003)	0.203 (0.003)
Source size N^S			
1000	1.288 (0.016)	0.366 (0.003)	0.318 (0.003)
5000	1.288 (0.016)	0.134 (0.001)	0.133 (0.001)

Covariate shift, prognostic change, and effect change. The central result concerns which kind of source–target difference transfer can tolerate. Table 2 and Figure 5 trace the normalized error of transfer learning as each mismatch grows from none to full. All three conditions start from the same level-0 error of about 0.220 and behave very differently from there. Under covariate shift the error did not deteriorate (0.14 at full shift): the shifted population still obeyed the same outcome rule, so the transferred relationship remained valid. The slight fall, like the steep rise in raw error, reflects the larger and more dispersed effects of the shifted population (Table 2) rather than a change in estimation quality.

This ordering separates "the target is different" into a harmless form which is a different population or a different background, and a damaging form, a different effect, and only the last degrades transfer. We stress that it is the change in the estimand, not the size of any single term in the bound, that does the separating. A prognostic change and an effect change both contribute to the mechanism residual γ^*, which is built from the discrepancy in outcome distributions; because a prognostic change shifts both potential-outcome surfaces while an effect change shifts only the treated one, γ^* is if anything larger in the harmless prognostic

Table 2: Effect of each mismatch as it grows from none (level 0) to full (level 1), as mean (SE) over $R = 100$ replications. SD(τ) is the spread of the true ITE in the target; $\varepsilon_{\text{PEHE}}/\text{Var}(\tau)$ is the normalized error of TL-TARNet.

	SD(τ)		TL-TARNet $\varepsilon_{\text{PEHE}}/\text{Var}(\tau)$		CITA	
Mismatch	none	full	none	full	none	full
covariate shift	1.55	3.53	0.225 (0.004)	0.137 (0.003)	0.240 (0.002)	0.530 (0.001)
prognostic change	1.54	1.55	0.223 (0.004)	0.268 (0.004)	0.239 (0.002)	0.246 (0.002)
effect change	1.56	1.94	0.225 (0.004)	0.353 (0.004)	0.241 (0.001)	0.247 (0.002)

case. What distinguishes the two is that a prognostic shift cancels in the contrast τ, whereas an effect change alters τ itself and enters irreducibly. The result is therefore best read as showing that the bound is loose for changes that cancel in the estimand, rather than as a direct confirmation of its mechanism term. The same contrast also shows that the gains are genuine information transfer rather than mere capacity control: a smaller, frozen network would regularize the estimate regardless of the source's content, yet a source carrying the wrong effect actively degrades it so it is what is transferred, not simply the reduced number of free parameters, that matters.

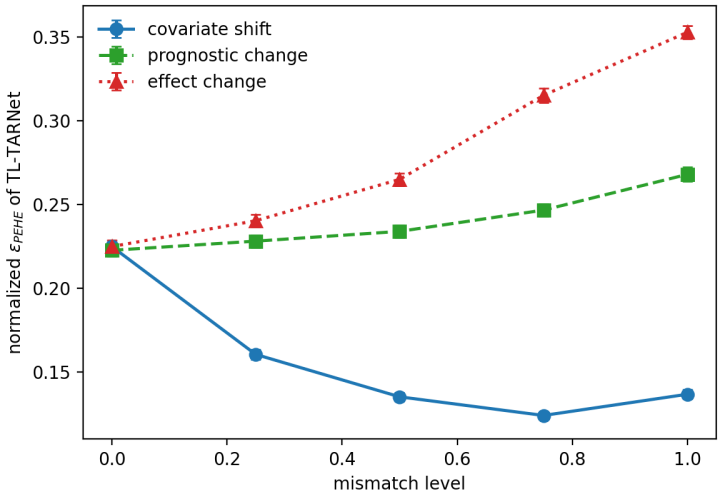

Figure 5: Normalized error of TL-TARNet as each kind of mismatch grows. Covariate shift (blue) does not degrade transfer; a prognostic change (green) costs little, since the treatment effect is unchanged; a change in the effect (red) degrades transfer steadily, because the source supplied the wrong effect rule and the target is too small to fully correct it.

Selection bias. Whether the target intervention was randomized or confounded acted largely independently of the mismatch: confounding added a roughly constant penalty within each mismatch condition (for the target-only model under an effect change, a normalized error of 1.00 randomized versus 1.32 biased, averaged over the remaining factors), and the ordering across mismatch types was the same under both kinds of assignment. Transfer learning and the source-only model were far less disturbed than the target-only model, because both inherit a representation estimated on the randomized source.

Causal inference task affinity. CITA behaved as a measure of covariate similarity more than of causal similarity. It rose clearly with covariate shift (from 0.24 to 0.53; Table 2) but barely moved under either a prognostic or an effect change (remaining near 0.24). This is a consequential limitation: the one kind of mismatch that actually degraded transfer, a change in the effect, is the kind CITA was least able to

detect, whereas the kind it tracked most strongly, covariate shift, was the kind transfer tolerated. Across the whole range of affinity our conditions produced (CITA roughly 0.1 to 0.6), transfer learning never lost to the target-only model; the difference between them favoured transfer at every level, so within this design we did not observe a CITA threshold beyond which borrowing becomes harmful. The mismatches we imposed, and in particular the changes in mechanism, did not drive CITA high enough to locate such a crossover; finding it would require either larger changes in the effect or sources more distant than those studied here.

Summary. Transfer learning improved small-sample ITE estimation substantially, with the largest gains for the smallest targets and the largest (sufficiently large) sources; it withstood covariate shift and a prognostic change but degraded under a genuine change in the treatment effect—the one mismatch that moves the estimand; and CITA, though a useful affinity score, proved far more sensitive to covariate shift than to the effect changes that most threaten transfer.

4 Empirical study

This application illustrates the methodological problem the paper addresses: estimating heterogeneous effects in a small target population while borrowing strength from a larger, related dataset, here through a substantive question about how household constraints shape children's study time. We evaluate the method in the regime it is designed for, namely small targets paired with a larger, sufficiently similar source, and characterize its behavior there. Because the data are observational and no ground-truth ITE is available, we assess transfer by its consistency with the larger source rather than by estimation error.

We use data from IHDS, Wave II (2011–12), a nationally representative dataset collected by the National Council of Applied Economic Research (NCAER) and the University of Maryland. The survey covers over 42,152 households across 33 states and union territories of India. The IHDS provides detailed information on health, education, jobs, and gender roles to build an overall picture of household welfare across India.

In this empirical illustration, we build on the framework of Choudhuri & Desai (2021) and specifically examine their hypothesis, which posits that households that rely on collected firewood (vs. purchased fuel or clean fuel) have children who study less. Following the authors, we define the treatment as a binary indicator for households that cook with firewood and gather it themselves rather than buy it and the outcome as children's total weekly minutes spent studying (school time plus homework). We control for household income, maternal education, caste–religion group, household electricity access, and distance to school. The sample is restricted to children aged 6–18 who are enrolled in school. Unlike the original paper which additionally includes village development indicators, and state fixed effects, we adopt a more parsimonious specification and exclude these variables.

We use Uttar Pradesh as the source because, after cleaning, it had the largest usable sample ($N^S = 1{,}247$). We consider three targets: (i) a random subsample of Uttar Pradesh ($N^T = 350$), (ii) a biased subsample of Uttar Pradesh constructed with the same selection function used in the simulation ($N^T = 350$), and (iii) the Punjab region ($N^T = 332$), a genuinely distinct population that, unlike the subsamples, introduces real source–target covariate shift.

For the target data selection, Punjab was chosen because it combined a relatively low CITA distance to Uttar Pradesh (0.32, versus, e.g., 0.37 for Jammu & Kashmir and 0.33 for Kerala) with more stable training than other regions. For each of the random and biased subsamples we generated 10 candidate data sets and retained the one with stable convergence and the lowest CITA.

The source model, trained on the full Uttar Pradesh data, gives a mean ITE of -110 minutes per week (standard deviation 7.79), i.e., children in firewood-collecting households study roughly two fewer hours per week than children in purchasing/clean-fuel households. We treat this source estimate as a reference for the target analyses, not as ground truth, and ask whether transfer yields target estimates that are consistent with it in sign and magnitude.

Table 3 reports the results. As expected from their shared population, the Uttar Pradesh subsamples have lower CITA than Punjab (0.09 and 0.11 versus 0.32), and within this application it is the ordering that is informative: smaller values mark the targets more closely related to the source. We caution against reading

the magnitudes absolutely. Although CITA is bounded between 0 and 1, it is computed from the Fisher matrices of a particular trained source network (Eqs. 13–15), so its scale depends on that network and on how the value was obtained; the numbers here are accordingly not comparable to the simulation values of Section 3, where the same kind of near-identical source–target pair produced CITA around 0.18–0.24.

Name	N^T	CITA	Mean with TL	Mean without TL	STD with TL	STD without TL
Random-subsample	350	0.09	-104.31	-31.86	10.48	2.03
Biased-subsample	350	0.11	-94.28	25.98	16.23	4.22
Punjab	332	0.32	-99.68	63.14	41.37	4.15

(N^T = target data sample size; CITA = Causal Inference Task Affinity; TL = transfer learning; STD = standard deviation)

Table 3: Estimation results of transfer learning using IHDS-II data.

Without transfer, the target-only estimates are implausible for two of the three targets: the mean ITE is positive for the biased subsample (25.98) and for Punjab (63.14), contradicting the hypothesized negative effect, and is attenuated toward zero for the random subsample (-31.86). With transfer, all three target means move toward the source reference and recover the hypothesized sign: -104.31 (random subsample), -94.28 (biased subsample), and -99.68 (Punjab). In this setting, transfer therefore stabilizes otherwise implausible small-sample estimates and brings them into agreement with the larger source. Absent ground truth, we describe this as consistency with the source reference rather than as a reduction in estimation error.

The standard deviations report the dispersion of predicted individual effects within a single fit; how much heterogeneity each model assigns across people. Target-only TARNet collapses to near-constant predictions (2.03, 4.22, 4.15), recovering essentially no heterogeneity, whereas the transfer models spread effects across individuals (10.48, 16.23, 41.37), consistent with the vertical scatter in Figure 6.

Figure 6 plots individual-level target predictions (vertical axis) against source predictions (horizontal axis), with transfer (circles) and without (crosses, target-only TARNet), for the random subsample (left) and the biased subsample (right). Because treatment is observational throughout, the panel labels refer to how each subsample was drawn, not to randomized treatment assignment. The TARNet predictions form a tight horizontal band almost regardless of the source value, indicating near-constant predicted effects and substantial disagreement with the source. The transfer predictions spread vertically, reflecting heterogeneity that the small target sample alone does not recover. The dashed diagonal marks exact source–target agreement; the systematic departures from it, especially for the biased subsample, show how far the target-only estimates sit from the source reference and how transfer moves them toward it.

5 Discussion

Building on the transfer-learning framework of Aloui et al. (2023), we provided a systematic evaluation of how transfer learning with the TARNet architecture performs when target samples are small and treatment assignment is often non-randomized which are conditions common in social science, healthcare, and policy evaluation. We brought the relevant generalization bounds and the CITA affinity measure into a single account and assessed the method in two complementary simulations and an empirical application, with the goal of using a larger source dataset to improve ITE estimation in a smaller target.

The results identify the conditions under which transfer is useful. When the source and target share a population, transfer reduced both bias and imprecision whenever the target was small or its assignment violated ignorability, with the clearest gains when a small target was paired with a sufficiently large source; beyond a moderate source size the additional benefit was small, so the binding requirement is less the sheer size of the source than that it be large enough to learn a stable representation. When the source and target differ, the kind of difference proved decisive. Transfer was robust to a covariate shift and a purely prognostic change in the outcome surface—the first because the underlying causal function was unchanged and only had to be realigned to a new covariate region, the second because moving the baseline left the treatment effect itself, the quantity being estimated, intact—but it degraded steadily once the treatment effect itself changed,

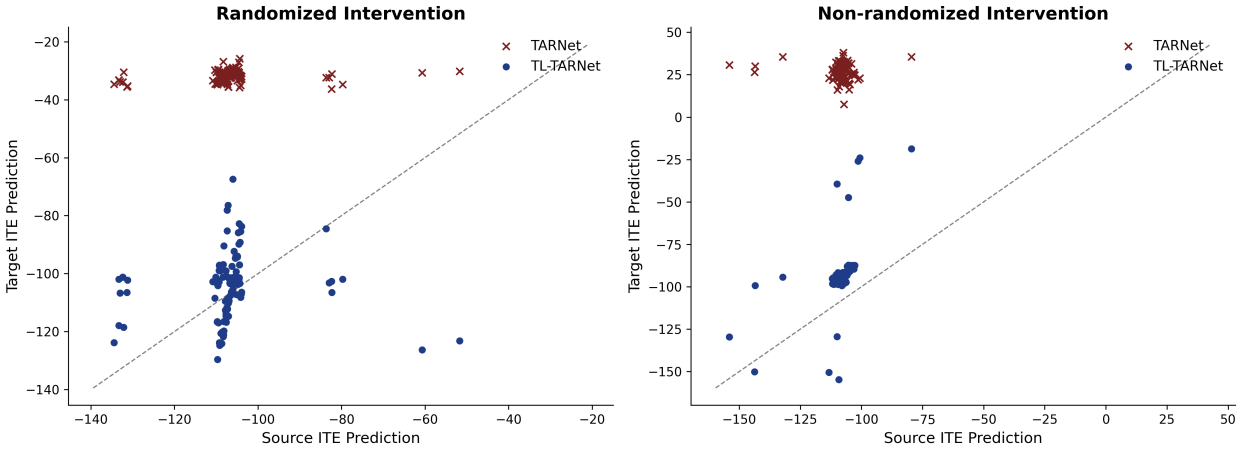

Figure 6: Source versus target individual-level ITE predictions, with transfer (circles) and without (crosses, target-only TARNet). Panels correspond to the random subsample (left, labeled "Randomized Intervention") and the biased subsample (right, labeled "Non-randomized Intervention").

the one difference that moves the estimand, which a small target cannot correct. We are deliberately careful about the mechanism: a prognostic and an effect change both contribute to the bound's residual γ^* (which is, if anything, larger for the harmless prognostic change), so it is not γ^* that separates them but whether the change cancels in the contrast τ. The ordering is therefore better read as showing that the bound is loose for changes that cancel in the estimand than as a direct confirmation of it. The practical lesson is sharp nonetheless: "the target is different" is harmless when it means a different population or a different background, and damaging only when it means a different effect.

A consequential qualification concerns how this difference is detected in practice. The CITA affinity score, which the framework uses to judge a source's suitability before transfer, responded clearly to covariate shift but barely moved under a change in the effect. The one kind of mismatch that actually degraded transfer is therefore the kind CITA is least able to see, whereas the kind it tracks most strongly is the kind transfer tolerates. Across the whole range of affinity our conditions produced, transfer never lost to the target-only model, so we did not observe a threshold beyond which borrowing becomes harmful; locating such a threshold would require larger changes in the effect, or sources more distant than those studied here.

In the empirical application, where the true effect is unknown, transfer produced target estimates that agreed with the larger source in sign and magnitude and stabilized small-sample estimates that were otherwise implausible. Because no ground truth is available there, we read this as evidence of consistency with the source rather than as a measured reduction in error, and we note that the result is strongest for the targets drawn from the source population and most tentative for the genuinely distinct target, where agreement with the source cannot by itself be taken to mean the estimate is closer to the truth.

Our design also defines the scope within which these conclusions hold. The first simulation draws the target from the source and so speaks to sample size and selection bias under a shared population; the second isolates covariate shift and mechanism change one at a time rather than in combination; and the empirical targets were selected for their affinity to the source and for stable training, so they illustrate the method where it is expected to apply rather than provide an unbiased estimate of its expected performance. Within this scope, the evidence supports a clear practical conclusion: in data-constrained settings, drawing on a larger and sufficiently similar source can stabilize ITE estimation and, where ground truth allows the comparison and the source supplies the right effect, reduce bias.

5.1 Practical considerations for the application of transfer learning

Successful transfer learning depends critically on the source model learning a stable, informative representation. Achieving such a learning requires careful monitoring of the optimization process. One major impact of the representation of the source data set is the target-dataset selection. The CITA measure is evaluated based on the trained source model; accordingly, the suitability of a target dataset is assessed in light of this representation. If this initial training was suboptimal, the distance will not represent a meaningful tool that can be used to improve estimation in the target data set.

In standard deep learning, a single loss (e.g., MSE) is minimized. With TARNet under transfer learning, optimization becomes multi-objective: the training loss combines prediction error with a distributional distance between treatment and control (or source and target) representations. These terms often differ in scale and evolve at different rates during training. If the hyperparameter κ that balances loss and distributional similarities is too small, distribution matching is neglected; if it is too large, the model underfits outcomes to satisfy the distance term. Even with a reasonable κ, using one learning rate to reduce both components can be difficult because their gradients may conflict in direction or magnitude. Consequently, convergence depends critically on the choice of κ and the learning rate. In practice, training may begin with a small learning rate to stabilize optimization (however with higher computational cost), with schedules adjusted after inspecting the loss values. In case of using small learning rates, it is important to increase epoch size substantially. Alternatively, a stepwise procedure can be used: first learn a latent space that aligns the treatment and control groups, and then fit the outcome model by minimizing a predictive loss (Du et al., 2021).

Another consideration during training is the choice of batch size. For small datasets, smaller batches often yield more stable generalization, whereas very large batches can degrade learning dynamics. In TARNet, where the objective typically includes a distributional distance between treatment and control representations, this term is well defined only when both groups are present within the same mini-batch. With group imbalance and very small batches, some mini-batches may contain observations from a single group, leading to noisy distance estimates and instability. Furthermore, before training, the dataset is partitioned into training and test sets, reducing the number of observations available for model fitting. Mini-batch training further lowers the effective sample size per update (and per group in TARNet). Under these conditions, highly overparameterized TARNet architectures such as with many layers or large hidden widths are prone to overfitting and unstable optimization. A more compact architecture is therefore generally more appropriate in small-sample settings for both source and target datasets. It is important to note that, researchers can decide how many layers are needed to be frozen and trainable for the target data set. While we used one unfrozen layers in our simulation and empirical application, it is possible to leave more than one layer to be trainable. Increasing the numbers of unfrozen layers naturally will require larger sample sizes in the target data set.

5.2 Future directions

Several extensions follow from these findings. The first concerns the diagnostic question our simulations left open. Because transfer never lost to the target-only model across the range of affinity we induced, we did not locate a threshold of source–target dissimilarity beyond which borrowing becomes harmful; the mismatches we imposed, and in particular the changes in the treatment effect, did not drive the affinity score high enough to reach a crossover. A natural next step is to push the effect change further, and to study sources more distant than those considered here, in order to map the region in which transfer ceases to help and to turn the qualitative ordering we observe into a concrete rule for practitioners. Closely related, and arguably more pressing, is a source-selection diagnostic sensitive to the difference that matters. CITA responds chiefly to covariate shift and is largely blind to a change in the effect which is the very mismatch that degrades transfer. The mechanism residual $\gamma*$ is not the remedy either: as our prognostic-change condition shows, it is an outcome-distribution discrepancy that responds to harmless changes as well, and more strongly if anything, so it does not isolate the damaging case. What a diagnostic must track is the discrepancy in the treatment-effect function itself but that is precisely the quantity it cannot observe before transfer, since it depends on the target effect transfer is meant to estimate. Approximating this effect-level discrepancy

without target labels, the same unobservable that limits the lower bound of Aloui et al. (2023), is the genuinely open problem, and the more reliable basis for deciding when a related source can be trusted.

Across studies and data sources, conceptually comparable variables are often operationalized differently, through different survey items, coding schemes, scales, or administrative definitions so that aligning them is non-trivial even when the substantive question is shared. Methods that recognize and accommodate such measurement differences would extend transfer to the many cases in which source and target do not share an identical feature space.

A third direction is the construction of larger source datasets by combining several smaller studies. Relevant information in applied research is frequently distributed across surveys, panels, administrative records, and regional studies, each covering different subpopulations, contexts, or covariates. Developing principled ways to pool such sources, preserving shared structure while accounting for differences in composition, measurement, and causal mechanism, would make it possible to assemble sources large enough for reliable transfer even when no single study suffices on its own, which is especially valuable when some groups, treatment conditions, or contexts are sparse or absent in any one dataset. Together, these directions would strengthen the robustness and practical reach of transfer-learning approaches across the heterogeneous data environments typical of social science.

References

Virginia Aglietti, Theodoros Damoulas, Mauricio Álvarez, and Javier González. Multi-task causal learning with gaussian processes. *Advances in neural information processing systems*, 33:6293–6304, 2020.

Ahmed Aloui, Juncheng Dong, Cat P Le, and Vahid Tarokh. Transfer learning for individual treatment effect estimation. In *Uncertainty in Artificial Intelligence*, pp. 56–66. PMLR, 2023.

Valerio Baćak and Edward H Kennedy. Principled machine learning using the super learner: an application to predicting prison violence. *Sociological Methods & Research*, 48(3):698–721, 2019.

Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 698–704, 2012.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134, 2013.

Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. *Advances in Neural Information Processing Systems*, 35:37184–37198, 2022.

Deirdre Bloome and Daniel Schrage. Covariance regression models for studying treatment effect heterogeneity across one or more outcomes: understanding how treatments shape inequality. *Sociological Methods & Research*, 50(3):1034–1072, 2021.

Jennie E Brand, Jiahui Xu, Bernard Koch, and Pablo Geraldo. Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociological Methodology*, 51(2):189–223, 2021.

Miriam Bruhn and David McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232, 2009.

Vinod Kumar Chauhan, Jiandong Zhou, Ghadeer Ghosheh, Soheila Molaei, and David A Clifton. Dynamic inter-treatment information sharing for individualized treatment effects estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3529–3537. PMLR, 2024.

Pallavi Choudhuri and Sonalde Desai. Lack of access to clean fuel and piped water and children's educational outcomes in rural india. *World Development*, 145:105535, 2021.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

Jeroen Hoogland, Joanna IntHout, Michail Belias, Maroeska M Rovers, Richard D Riley, Frank E. Harrell Jr, Karel GM Moons, Thomas PA Debray, and Johannes B Reitsma. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in medicine*, 40(26):5961–5981, 2021.

Melody Y Huang and Harsh Parikh. Towards generalizing inferences from trials to target populations. *Harvard Data Science Review*, 6(4), 2024.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

Taskin Kavzoglu and Paul M Mather. The use of backpropagating artificial neural networks in land cover classification. *International journal of remote sensing*, 24(23):4907–4938, 2003.

Bernard J Koch, Tim Sainburg, Pablo Geraldo Bastias, Song Jiang, Yizhou Sun, and Jacob G Foster. A primer on deep learning for causal inference. *Sociological Methods & Research*, 54(2):397–447, 2025.

Peter Kokol, Marko Kokol, and Sašo Zagoranski. Machine learning on small size samples: A synthetic knowledge synthesis. *Science Progress*, 105(1):00368504211029777, 2022.

Cat P Le, Juncheng Dong, Mohammadreza Soltani, and Vahid Tarokh. Task affinity with maximum bipartite matching in few-shot learning. *arXiv preprint arXiv:2110.02399*, 2021.

Peter Lugtig. Panel attrition: Separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research*, 43(4):699–723, 2014.

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. *Advances in neural information processing systems*, 31, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Judea Pearl. Transportability across studies: A formal approach. 2011.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 451–482. 2022.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pp. 15–18, 2019.

Torgyn Shaikhina and Natalia A Khovanova. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial intelligence in medicine*, 75:51–63, 2017.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI global, 2010.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Thanh Vinh Vo, Pengfei Wei, Trong Nghia Hoang, and Tze Yun Leong. Adaptive multi-source causal inference from observational data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1975–1985, 2022.

Sandra Wankmüller. Introduction to neural transfer learning with transformers for social science text analysis. *Sociological Methods & Research*, 53(4):1676–1752, 2024.

Song Wei, Hanyu Zhang, Ronald Moore, Rishikesan Kamaleswaran, and Yao Xie. Transfer learning for causal effect estimation. *arXiv preprint arXiv:2305.09126*, 2023.

Siyun Yang, Fan Li, Monique A Starks, Adrian F Hernandez, Robert J Mentz, and Kingshuk R Choudhury. Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. *Statistics in Medicine*, 39(28):4218–4237, 2020.

Xiang Zhou and Yu Xie. Heterogeneous treatment effects in the presence of self-selection: A propensity score perspective. *Sociological Methodology*, 50(1):350–385, 2020.

A Additional information on Simulation 1

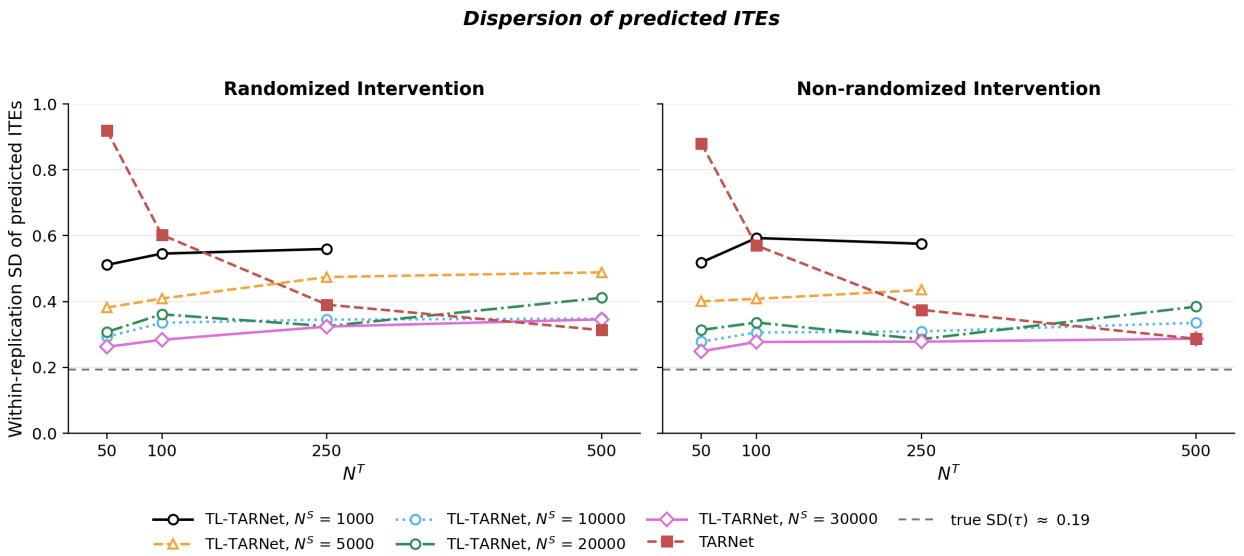

Figure 7: Dispersion of predicted ITEs in Simulation 1: within-replication standard deviation of the predicted effects, averaged over the $R = 100$ replications, for randomized (left) and non-randomized (right) targets. The dashed line marks the dispersion of the true effects, $\mathrm{SD}(\tau) \approx 0.19$. All methods over-disperse relative to the truth; target-only TARNet inflates heterogeneity most at small N^T, while transfer keeps dispersion closer to the truth. TARNet does not use the source, so its line averages over the source-size conditions.

B Additional information on the CITA measures

Causal inference tasks possess an inherent symmetry: if we swap the treated and control labels, the two outcome-prediction subproblems remain identical except for the sign of the treatment effect. An affinity measure should therefore assign a distance of zero to a task and its label-flipped counterpart. Following Le et al. (2021), we compute the one-sided distance $d[s, t]$ (source Fisher curvature on source versus target dataset) and then repeat this calculation after every permutation of the treatment labels in the target set.

We denote by S_{M+1} the set of all permutations of the treatment labels $\{0, \ldots, M\}$. For a source-trained network N_{θ_a}, let $F_{a,a}$ be its Fisher information matrix on the source data D_a, and let $F_{a,\sigma(t)}$ denote the Fisher matrix of the same network evaluated on the target dataset D_t after the labels have been permuted according to $\sigma \in S_{M+1}$. For each permutation define

$$d_\sigma \;=\; \frac{1}{\sqrt{2}} \left\| F_{a,a}^{1/2} - F_{a,\sigma(t)}^{1/2} \right\|_F,$$

where $F^{1/2}$ is the principal matrix square-root and $\|\cdot\|_F$ the Frobenius norm. The CITA is defined as the minimum of these distances, making the score strictly label-invariant while requiring no retraining: the two heads of the frozen network are simply read in the opposite order when labels are swapped. CITA distance between the source task T_s and the target task T_t is

$$d_{\mathrm{sym}}[s,t] \;=\; \min_{\sigma \in S_{M+1}} d_\sigma.$$

Aloui et al. (2023) show that this symmetrized score dominates the one-sided alternative; CITA drops to zero when tasks differ only by a relabeling and preserves the relative ordering of source tasks under any label flip, whereas the non-symmetrized distance does not. Consequently, CITA provides a more reliable gauge of true task similarity and a safer basis for selecting sources to transfer from.