Transfer Causal Learning: Insights from Analysis of Twins Dataset

Taoning Li

May 12, 2024

Contents

1 Introduction

1.1 Motivation

Estimation of causal effects is highly concerned in many fields since it enables informed decision-making by revealing the true impact of interventions or treatments on outcomes of interest. One of the most important goals is to integrate different data sets and leverage information from auxiliary studies for estimating better causal effects. It is therefore of interest to explore whether we can develop statistical

methods with theoretical guarantees and robust practical applications to enhance the estimation of causal effects by leveraging information of auxiliary data.

Randomized Controlled Trial (RCT) is a gold standard on estimating causal effects in the field of causal inference, which is because the treat assignment is completely controlled by researchers. However, RCTs suffer from two main challenges: (i) high costs leading to small sample sizes, and (ii) limited representativeness of the target population due to highly selected samples. On the other hand, observation studies (OSs) are available with massive data and low cost. However, unobserved confounders, which both affects treatment and outcome, is usually attached to OS, while this problem does not exist in RCT studies since the treatment assignment is completely decided by the researchers. It is of great interest to combine the advantage of RCT and OS therefore improve the estimation of causal effects. There are also various methods proposed to combine RCTs and OSs, see (Colnet et al., 2024, Yang and Ding, 2019).

Transfer learning (Olivas et al., 2009) aims to improve performance on another related domain or task by leveraging knowledge from other domains or tasks, particularly when data is limited or unavailable. We can naturally adapt the transfer learning methodology to leverage the auxiliary information. **The main purpose of this report** is to evaluate the effectiveness of Transfer Causal Learning (TCL) algorithm as proposed by Wei et al. (2024) using the Twins Data set, which documents infant deaths in the US from 1989 to 1991. Empirical results under various settings will be provided in Section 4.

1.2 Organization

The remaining parts of this report are organized as follows. We will introduce the settings and formulate the TCL problem in Section 2. The specific methodology will be introduced in Section 3. Section 4 provides the main results of this report, where we will first provide both data description and exploratory data analysis of Twins data set, then analyze the results. Finally, some conclusions and discussions will be given in Section 5.

2 Problem Formulation

2.1 Preliminaries

Throughout this report, we consider potential outcome framework proposed by Neyman-Rubin (Rubin, 1974). We will use the notation parallel to Wei et al. (2024). This section is a short version of Section 2 in Wei et al. (2024). Consider the tuple $(\boldsymbol{X}, Z, Y)$ in the target study, where random vector $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ represents covariates measured prior to receipt of treatment, r.v. $Z \in \{0, 1\}$ is treatment

indicator ($Z = 1$ if treated and 0 otherwise) and the random variable Y is the observed outcome under

$$Y = Y(1)Z + (1 - Z)Y(0),$$

where $Y(1), Y(0)$ represent the potential outcomes of the subject under treatment and control, respectively. The parameter of interest is average treatment effect(ATE) $\tau = \mathbb{E}(Y(1) - Y(0))$, which reflects the average difference in potential outcomes between individuals who received the treatment and control across the entire population. In most observational studies, treatment Z is not independent of covariate $\boldsymbol{X}$, leading to which is commonly referred to as confounded treatment. Throughout this report, we will assume ignorability, which is generally considered as

$$(Y(1), Y(0)) \perp Z|\boldsymbol{X}.$$

That is, potential outcome is independent of treatment conditioning on the covariates.

In the remaining part of this section, we shall go through Outcome Regression (OR) estimator, Inverse Probability Weighted (IPW) estimator and Doubly Robust (DR) estimator.

OR estimator. OR estimator is unbiased under the potential outcome framework. We assume the following model

$$m_z(\boldsymbol{X}) = \mathbb{E}\left[Y(z)|\boldsymbol{X}\right], \quad z \in \{0, 1\}.$$

Here, $m_z(\boldsymbol{X})$ is the mean function of $Y(z)$ conditional on $\boldsymbol{X}$. For binary Y, we assume a logistic regression model on m_z, and we assume m_z is linear for continuous Y. For $z \in \{0, 1\}$, let $n_z = \#\{i : z_i = z\}$ (# represents the cardinality of a set) and $\widehat{m}_z(\boldsymbol{x}_i)$ be the fitted potential outcome for i-th subject, the OR estimator for ATE is given by:

$$\widehat{\tau}_{\mathrm{OR}} = \frac{1}{n_1} \sum_{z_i=1} \widehat{m}_1(\boldsymbol{x}_i) - \frac{1}{n_0} \sum_{z_i=0} \widehat{m}_0(\boldsymbol{x}_i).$$

Another commonly considered OR estimator is similar to $\widehat{\tau}_{\mathrm{OR}}$, but takes average under all the subjects on estimating both treatment and control group.

$$\widehat{\tau}'_{\mathrm{OR}} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{m}_1(\boldsymbol{x}_i) - \widehat{m}_0(\boldsymbol{x}_i)).$$

IPW estimator. Denote the propensity score $e(\boldsymbol{X})$ as $e(\boldsymbol{X}) = \mathbb{P}(Z = 1|\boldsymbol{X})$, which represents the probability of being treated conditional on the covariate being $\boldsymbol{X}$. The IPW estimator for ATE is:

$$\widehat{\tau}_{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i y_i}{\widehat{e}(\boldsymbol{x}_i)} - \frac{(1 - z_i) y_i}{1 - \widehat{e}(\boldsymbol{x}_i)}.$$

DR estimator. Doubly robust estimator combines the strengths of OR estimator and IPW estimator, which is given by

$$\widehat{\tau}_{\mathrm{DR}}=\frac{1}{n}\sum_{i=1}^{n}\frac{z_i y_i-\widehat{m}_1\left(\boldsymbol{x}_i\right)\left(z_i-\widehat{e}\left(\boldsymbol{x}_i\right)\right)}{\widehat{e}\left(\boldsymbol{x}_i\right)}-\frac{\left(1-z_i\right)y_i+\widehat{m}_0\left(\boldsymbol{x}_i\right)\left(z_i-\widehat{e}\left(\boldsymbol{x}_i\right)\right)}{1-\widehat{e}\left(\boldsymbol{x}_i\right)}.$$

Then $\widehat{\tau}_{\mathrm{DR}}$ is consistent, provided that either $m_z(\boldsymbol{x})$ or $e(\boldsymbol{x})$ is correctly specified under the assumption of ignorability.

2.2 Transfer Causal Learning Model

In this section, We are interested in estimating ATE under target population. Denote samples from the target domain as $\mathcal{D}=\{(\boldsymbol{x}_i,z_i,y_i)\}_{i=1}^{n}$, and we observe n_s samples $\mathcal{D}_s=\{(\boldsymbol{x}_{i,s},z_{i,s},y_{i,s})\}_{i=1}^{n_s}$ from source domain. It is generally supposed that $n\ll n_s$. Consider that the PS model takes the following form:

$$\mathbb{P}(Z=1\mid\boldsymbol{X})=e\left(\boldsymbol{X};\beta_{\mathrm{t}}\right),\quad \mathbb{P}\left(Z_s=1\mid\boldsymbol{X}_s\right)=e_s\left(\boldsymbol{X}_s;\beta_s\right),$$

where functions $e(\cdot),e_s(\cdot)$ have known form with unknown d_1-dimensional nuisance parameters, i.e., $\beta_t,\beta_s\in\mathbb{R}^{d_1}$. Similarly, the OR model has the following form: for $z\in\{0,1\}$,

$$\mathbb{E}\left[Y_z\mid\boldsymbol{X}\right]=m_z\left(\boldsymbol{X};\alpha_{z,t}\right),\quad \mathbb{E}\left[Y_{z,s}\mid\boldsymbol{X}_s\right]=m_{z,s}\left(\boldsymbol{X}_s;\alpha_{z,s}\right),$$

where functions $m_z(\cdot),m_{z,s}(\cdot)$ have known form with unknown nuisance parameters $\alpha_{z,t},\alpha_{z,s}\in\mathbb{R}^{d_2}$.

3 Methodology

This section provides methodology of ℓ_1-TCL. We introduce the main guarantee for knowledge transfer-ability by defining s-sparse vectors.

Definition 1 (s-sparse vector) A vector $v\in\mathbb{R}^d$ is said to be s-sparse (with $0\le s\le d$) if this vector has at most s non-zero elements, i.e., $\|v\|_0\le s$.

We assume

$$\Delta_\beta=\beta_t-\beta_s \text{ and } \Delta_{\alpha,z}=\alpha_{z,t}-\alpha_{z,s},\quad z\in\{0,1\}$$

are s-sparse. That is, the posterior information when $\boldsymbol{x}_i$ is given are similar in the sense that the regression coefficients vary only across a small fraction of components.

In order to estimating treatment effects under the target domain, we only need to estimate β_t and $\alpha_{z,t}$ for $z\in\{0,1\}$ by leveraging the information given by samples $\mathcal{D}_s=\{(\boldsymbol{x}_{i,s},z_{i,s},y_{i,s})\}_{i=1}^{n_s}$ from the source domain. To this end, we first estimate β_s and $\alpha_{z,s}$ for $z\in\{0,1\}$ by using $\mathcal{D}_s$. We propose

a generic ℓ_1 transfer learning framework. Suppose a general nuisance vector $\theta \in \mathbb{R}^d$, which can be estimated by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}(\theta, \mathcal{D}), \tag{1}$$

where Θ is the parameter space, $\mathcal{L}(\theta, \mathcal{D})$ is a loss function and $\mathcal{D}$ is the set containing all samples that are used for estimation. For example, suppose $m_z(\boldsymbol{X}; \alpha_{z,s}) = \alpha_{z,s}^\top \boldsymbol{X}$. Then an estimator is given by solving (1) with parameter space $\Theta = \mathbb{R}^d$, sample set $\mathcal{D} = \mathcal{D}_s = \{(\boldsymbol{x}_{i,s}, z_{i,s}, y_{i,s})\}_{i=1}^{n_s}$ and sum of squared differences loss function $\mathcal{L}(\theta, \mathcal{D}) = \mathcal{L}(\theta, \mathcal{D}_s) = \sum_{i:z_{i,s}=z}(y_{i,s} - \theta^\top \boldsymbol{x}_{i,s})^2$. Therefore,

$$\hat{\alpha}_{z,s} = \operatorname{argmin} \sum_{i:z_{i,s}=z} (y_{i,s} - \alpha^\top \boldsymbol{x}_{i,s})^2 \text{ for } z \in \{0, 1\}.$$

under this regime. Since Z is binary, we assume a logistic regression model and use the log-likelihood function as loss function. Detailed discussions can be found in Wei et al. (2024), while we only provide a generic ℓ_1 transfer learning framework in the remaining part of this section.

A generic ℓ_1 transfer learning framework

Input: $\mathcal{D}_s, \mathcal{D}_t$.

Output: $\hat{\theta}_t$.

Step 1: Rough Estimation:

$$\hat{\theta}_s = \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}(\theta, \mathcal{D}_s) \tag{2}$$

Step 2: ℓ_1-Transfer:

$$\hat{\theta}_t = \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}(\theta, \mathcal{D}_t) + \lambda \|\theta - \hat{\theta}_s\|_1 \tag{3}$$

This framework is similar to the Trans-Lasso algorithm proposed in Li et al. (2022). We simply illustrate the rationale of this framework. In Step 1, we use samples from the source domain to obtain a rough estimation. In Step 2, we use the estimator in Step 1 as initial estimator and optimize θ_t by using samples from the target domain and adding a ℓ_1-penalty on the loss function, since we believe that $\Delta_\theta = \theta_s - \theta_t$ is sparse. λ is chosen by cross validation in practice.

Finally, the plug-in estimators for estimated $\hat{\alpha}_{z,s}$ and $\hat{\beta}_s$ is given by

$$\widehat{\tau}_{\text{TLIPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{z_i y_i}{g\left(\boldsymbol{x}_i^T\widehat{\beta}_t\right)} - \frac{(1-z_i)\,y_i}{1-g\left(\boldsymbol{x}_i^T\widehat{\beta}_t\right)},$$

$$\widehat{\tau}_{\text{TLOR}} = \frac{1}{n_1}\sum_{z_i=1}\boldsymbol{x}_i^T\widehat{\alpha}_{1,t} - \frac{1}{n_0}\sum_{z_i=0}\boldsymbol{x}_i^T\widehat{\alpha}_{0,t},$$

$$\widehat{\tau}_{\text{TLDR}} = \frac{1}{n}\sum_{i=1}^{n}\frac{z_i y_i - \boldsymbol{x}_i^T\widehat{\alpha}_{1,t}\left(z_i - g\left(\boldsymbol{x}_i^T\widehat{\beta}_t\right)\right)}{g\left(\boldsymbol{x}_i^T\widehat{\beta}_t\right)} - \frac{(1-z_i)\,y_i + \boldsymbol{x}_i^T\widehat{\alpha}_{0,t}\left(z_i - g\left(\boldsymbol{x}_i^T\widehat{\beta}_t\right)\right)}{1-g\left(\boldsymbol{x}_i^T\widehat{\beta}_t\right)}.$$

4 Real Data Analysis

4.1 Data Description: Twins

The Twins dataset utilizes data from twin births in the USA between 1989-1991, which is a benchmark task for evaluating the estimation of causal treatment effect and was first introduced by Louizos et al. (2017). The original idea of the Twins dates back to Almond et al. (2005). The unprocessed data source can be obtained at https://www.nber.org/research/data/linked-birthinfant-death-cohort-data, which contains the NBER (National Bureau of Economic Research) collection of Birth Cohort Linked Birth and Infant Death Data of the National Vital Statistics System of the National Center for Health Statistics. The user guide for the raw data can be obtained at https://data.nber.org/lbid/docs/LinkCO89Guide.pdf. Demographic data include variables such as date of birth, age, and educational attainment of parents, marital status, live-birth order, race, sex, and geographic area. Health data include items such as birth weight, gestation, prenatal care, attendant at birth, and Apgar score. Geographic data includes state, county, and city of mother's residence and state and county of the place of birth.

The treatment here is whether the infant is the heavier one of the twins, and the outcome is the mortality of the infant within one year. Twins dataset also record a list of covariates relating to the parents, the pregnancy and birth: mother and father education, marital status, race and residence; number of previous births; pregnancy risk factors such as diabetes, renal disease, smoking and alcohol use; quality of care during pregnancy; whether the birth was at a hospital, clinic or home; and number of gestation weeks prior to birth. It is noticeable that in this special dataset, under the potential outcome structure of Rubin, the potential outcome of the treatment and the control (mortality of the heavier and the lighter infant) are all observable and hence bias can be calculated to evaluate the performance of candidate estimation methods.

In Louizos et al. (2017), only twins of the same sex are recorded, which is parallel to the processed dataset that we proposed. However, Louizos et al. (2017) considers twin pairs that both weigh less or equal than 2000 grams due to the outcome is quite rare (3.5% first-year mortality), where the proposed dataset includes weights of all twins with the same sex. We can treat all newborn twins less than 2000

g as the target domain and the remaining newborn twins as the source domain. This motivates us to use the information from the samples in source domain to better estimate the causal effect on the target domain. In addition, we are also strongly interested in leveraging the mortality information of twins that with different sex to better estimate the causal effect of the target domain. Since Twins dataset includes both treatment outcome and counterfactual outcome, we can also use Twins dataset to simulate observational dataset (Louizos et al., 2017).

4.2 Exploratory Data Analysis

Figure 1 shows the density and quantiles of the outcome on a random selected group with 1000 infants. It can be seen from the summary statistics that the average weight of the control group is 2261 grams, while the treatment group weighs 2545 grams on average. However, Louizos et al. (2017) considers twin pairs that both weigh less or equal than 2000 grams due to the outcome is quite rare, which is below than the average weight of both treatment groups and control groups. This motivates us to consider the CATE (conditional average treatment effect).

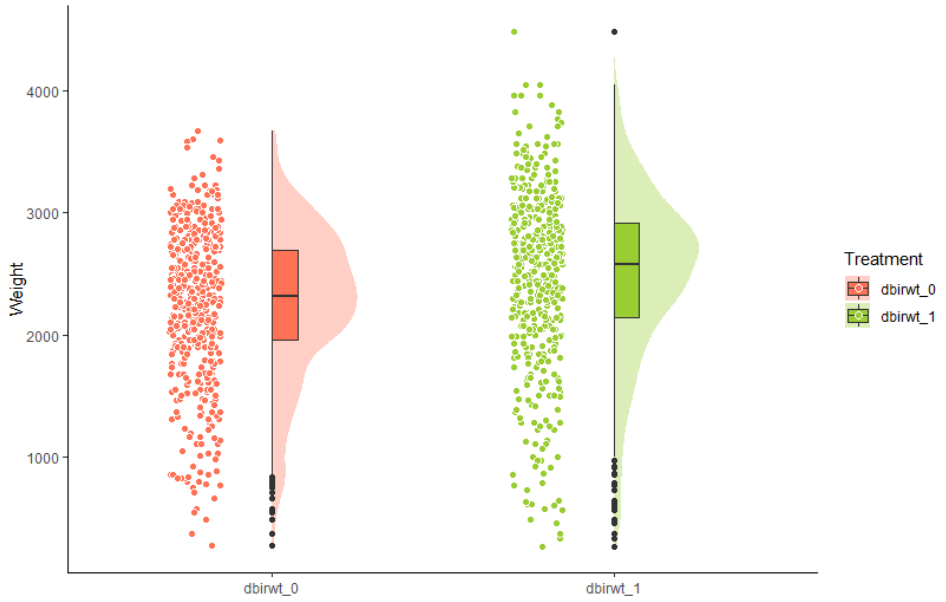

Figure 1: Weight distribution of lighter twin (left panel) and heavier twin (right panel) among 500 randomly selected twin pairs.

In the dataset spanning 1989 to 1991, out of 71,345 twin pairs, 3,590 pairs (5.03%) had at least one twin pass away within a year. Among these, 1,282 pairs (35.71%) saw only the lighter twin die, and 774 pairs (21.55%) saw only the heavier twin die. Initial analysis indicates that treatment (whether the twin is the heavier one) has a negative treatment effect on the outcome (whether the twin dies within one year).

Next we evaluate the correlation of risk factors. In the following analysis, we only use the cleaned samples that have fully observed covariates. We study the correlations of the covariates. Since there are approximate 50 covariates, we only choose to study correlations between risk factors.

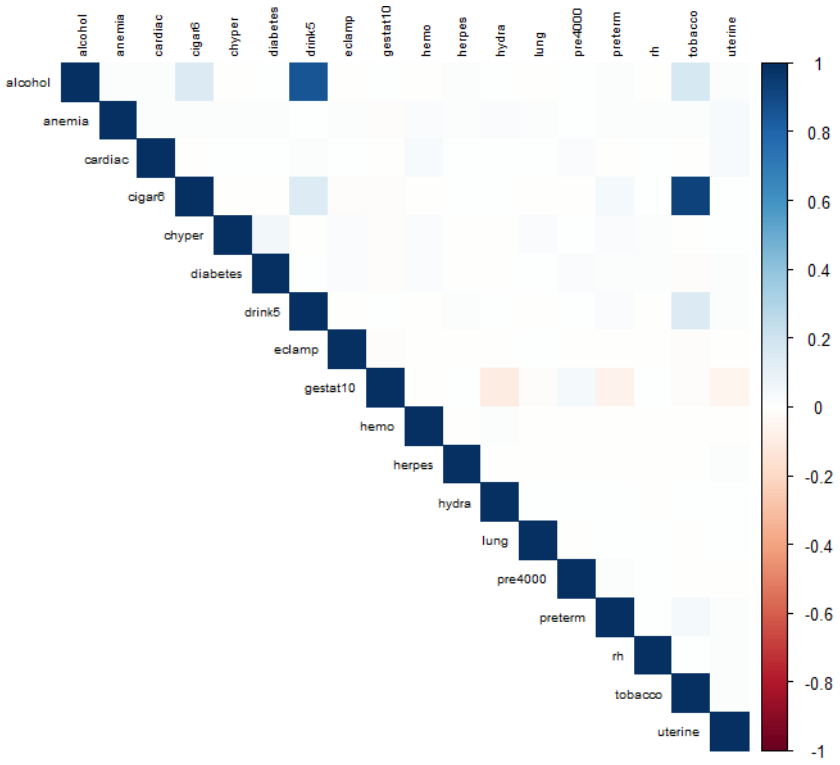

Figure 2: Correlation plot of the risk factors.

The examination of the correlation plot (Figure 2) is imperative for subsequent data processing and analysis. Notably, there exists a significant correlation between 'alcohol' and 'drink5', both of which gauge maternal alcohol consumption. Similarly, a comparable correlation is observed between 'cigar5' and 'tobacco'. Therefore, we shall preserve only one covariate between 'alcohol' and 'drink5', and similarly between 'cigar5' and 'tobacco'. Additionally, a special covariate, 'gestat10', describing the weeks of gestation in pregnant women, warrants introduction. A lower 'gestat10' value signifies a shorter pregnancy duration and consequently, an elevated mortality risk. 'gestat10' exhibits slight correlations with 'hydra' (risk factor: Hvdramnios/Oliqohvdramnios), 'pre4000' (risk factor: Previous infant 4000+ grams), 'preterm' (risk factor: Previous pre-term or small), and 'uterine' (risk factor: Uterine bleeding).

4.3 Results

In this real data application, we are interested in whether the heavier twin (treatment) has a higher mortality rate within one year (outcome) in comparison with the lighter twin (control). We divide the first ten states of birth in the United States alphabetically into the source domain, and the remaining infants form the samples of the target domain. The underground true ATE on the source domain and target domain are

$$\tau_s = -0.007363437, \quad \tau_t = -0.006115548.$$

We are intersted in the empirical results on the following settings of transfer causal learning:

(i) both samples of source domain and target domain come from RCT;

(ii) samples of source domain come from OS, samples of target domain come from RCT;

(iii) both samples of source domain and target domain come from OS.

For observational studies, we simulate propensity score as

$$z_i|\boldsymbol{x}_i \sim Ber(gestat10/10 - 1),$$

where 'gestat10' is the index of gestation weeks prior to the birth. Penalty parameter λ is selected through 5-fold cross validation.

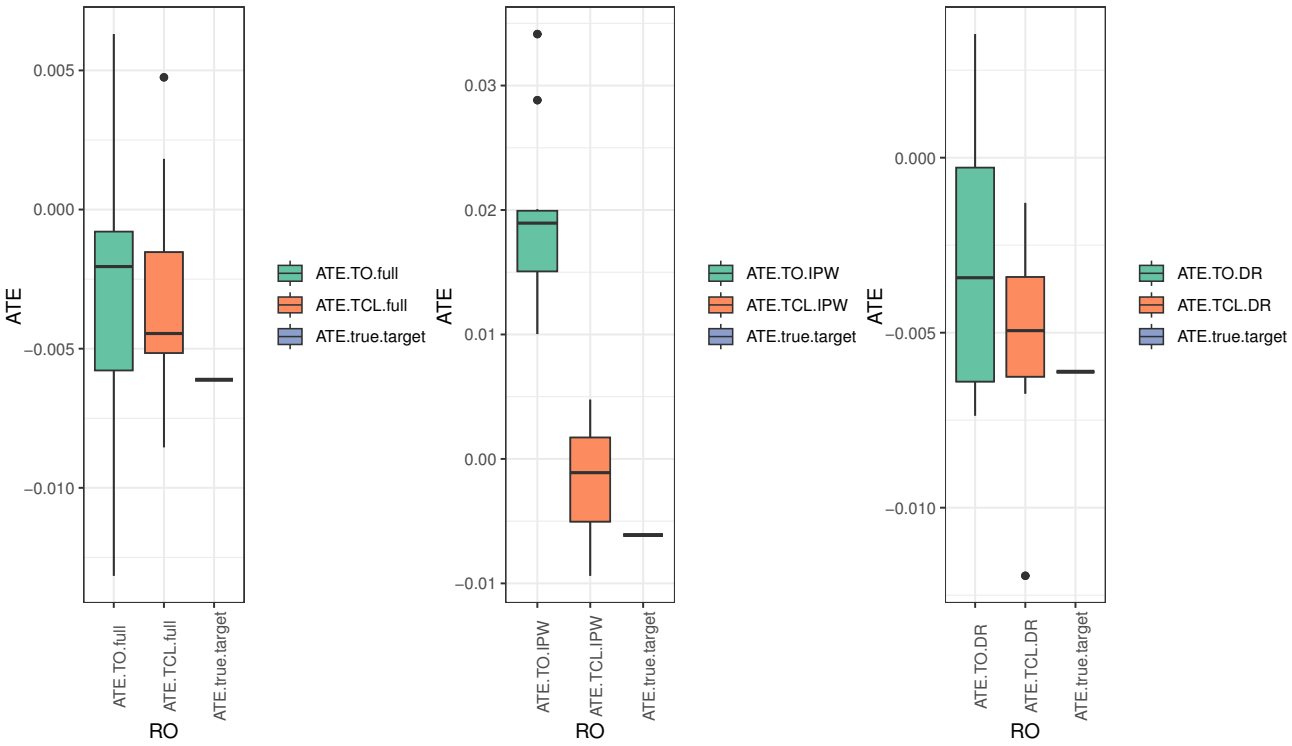

Figure 3: Estimation of Target Only (TO) and Transfer Causal Learning (TCL) under Observational Study (source) + RCT (target).

Figure 3 shows the performance of TCL algorithm in comparison (middle bar in each panel) with baseline estimators (left bar in each panel) and underground true ATE (right bar in each panel) under Observational Study (source) + RCT (target). Similarly, figure 4 shows the performance of TCL algorithm under Observational Study (source) + Observational Study (target). Three panels in figures 3 and 4 represent OR estimator (left panel), IPW estimator (middle panel) and DR estimator (right panel), repsectively.

Under the settings above, it can be seen that Transfer Causal Learning algorithm leverages the information in the presence of source data in a robust pattern, in comparison with using samples from target domain only. The reasons are considered as: **(i)** The distribution of source and target domain are quite similar, since we divide the source and target domain by the states in the US. This division

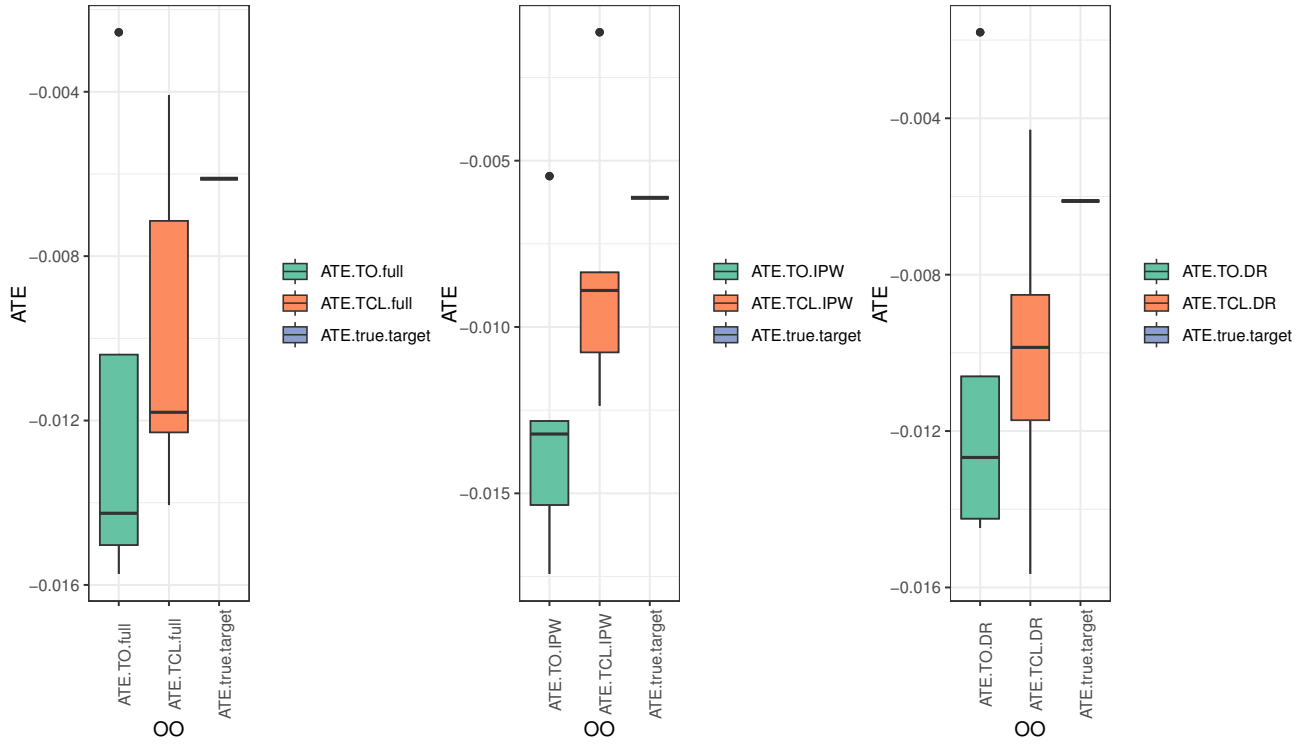

Figure 4: Estimation of Target Only (TO) and Transfer Causal Learning (TCL) under Observational Study (source) + Observational Study (target).

would not cause too many differences between source and target domain. **(ii)** The propensity score generated under above setting can be easily transferred. Under the Observational Study (source) + Observational Study (target) setting, the propensity score is considered only related to the gestation variable. This naturally satisfies the s-sparse condition of Δ_β as we assumed in Section 3.

To consider (i) stated before, we initially examine the behavior of TCL when the distribution of the source domain differs from that of the target domain. We partition the samples from the source domain into twin pairs both weighing less than 2000 grams, with the remaining samples forming the target domain. We consider RCT (source) + RCT (target) to control the effect of estimation error of nuisance parameters.

Figure 5 shows the performance of TCL under different domains. In the left panel, domains are categorized based on states, while in the right panel, they are classified according to weights. Five bars of each panel represent OR estimator uses samples from target only; full-OR estimators (τ'_{OR} introduced in Section 2) uses samples from target only; OR estimator uses both samples from the source and the target; full-OR estimators uses both samples from the source and the target; underground true ATE.

In the right panel, the distribution of source domain differs from the target domain a lot, therefore TCL hardly leverages the information from the source domain and tend to be robust. However, results from the left panel shows that the results of TCL algorithm may also cause negative transfer. In this regime, TCL leverages information from the source domain, but behaves less robust.

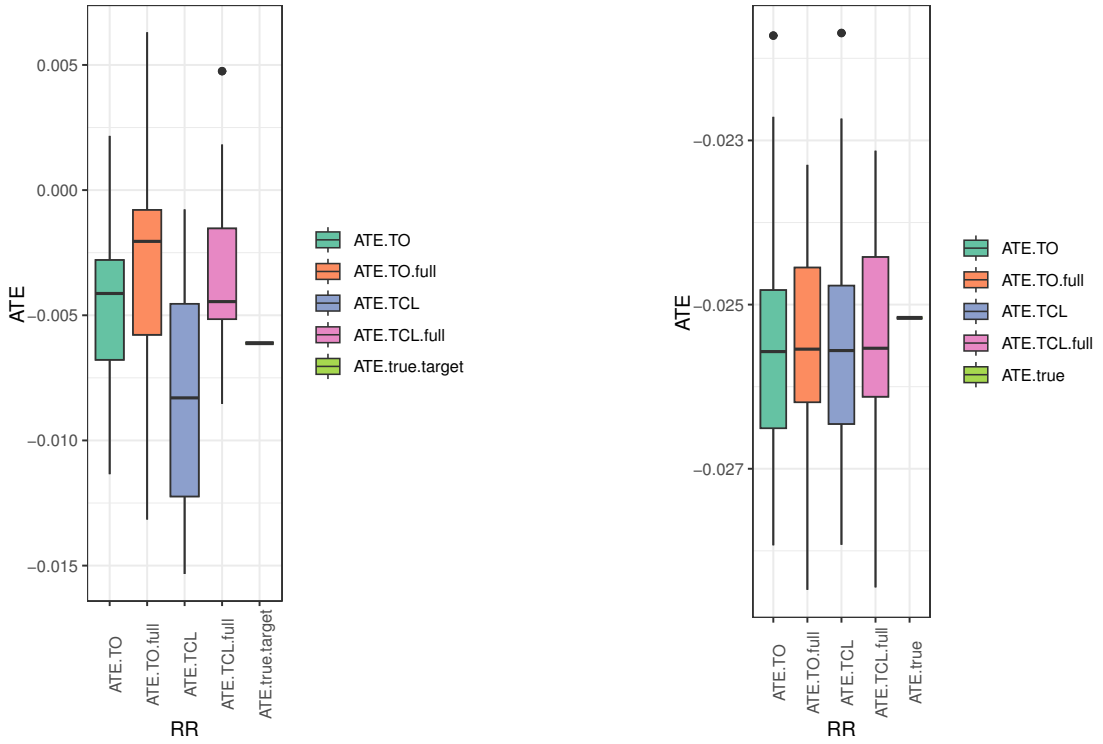

Figure 5: Outcome regression estimation under two different source domains and target domains (both RCT). Left panel: categorized by states, right panel: categorized by weights.

Finally, we shall consider (ii) as stated before. We shall use the same category as the right panel in Figure 5 as a standard baseline. To study how TCL algorithm behaves in more complex observational studies, we now generate the propensity score from the source domain as

$$z_i|\boldsymbol{x}_i \sim Ber(\sigma(w_1 \times gestat10 + w_2 \times drink5 + w_3 \times cigar6)),$$

where w_1, w_2, w_3 are determined based on a scaling of the discrete variables. We consider three variables that is highly correlated to the mortality rate of infants. To make the probability be in the interval $[0, 1]$, we impose $\sigma(\cdot)$ to be the sigmoid function. We use IPW estimator for the consideration of brevity, since we only focus on the transfer of propensity score. Figure 6 shows the results of TCL algorithm in this regime. TCL estimates slightly better than IPW that only uses target only. However, robustness and confidence should be further considered, as the TCL algorithm tends to produce higher estimations compared to estimators that solely utilize target samples, regardless of whether the estimator is higher or lower than the true underlying ATE.

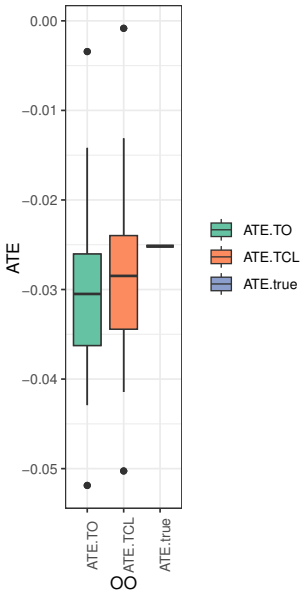

Figure 6: IPW estimators under more complex propensity settings.

5 Conclusion & Discussion

In conclusion, the TCL algorithm can enhance the estimation of ATE by leveraging auxiliary data. However, further discussion is needed to determine its robustness under real-world data and data with confounding factors. We can also consider to advance the transfer learning estimator by incorporating conditional average treatment effects (CATE) in future work.

References

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *The Quarterly Journal of Economics 120*(3), 1031–1083.

Colnet, B., I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang (2024). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science 39*(1), 165–191.

Li, S., T. T. Cai, and H. Li (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 149–173.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling (2017). Causal effect inference with deep latent-variable models. *Advances in neural information processing systems 30*.

Olivas, E. S., J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, L. Serrano, et al. (2009). *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI global.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Wei, S., H. Zhang, R. Moore, R. Kamaleswaran, and Y. Xie (2024). Transfer learning for causal effect estimation.

Yang, S. and P. Ding (2019). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*.