Journal of Machine Learning Research 27 (2026) 1-61 Submitted 5/25; Revised 6/26; Published 7/26

Incorporating external data for analyzing randomized clinical trials: A transfer learning approach

Yujia Gu GUYJ@MAIL.TSINGHUA.EDU.CN
Department of Statistics and Data Science
Tsinghua University
Beijing, 100084, China

Hanzhong Liu LHZ2016@TSINGHUA.EDU.CN
Department of Statistics and Data Science
Tsinghua University
Beijing, 100084, China

Wei Ma MAWEI@RUC.EDU.CN
Institute of Statistics and Big Data
Renmin University of China
Beijing, 100872, China

Editor: Kun Zhang

Abstract

Randomized clinical trials are the gold standard for analyzing treatment effects. However, increasing costs and ethical concerns may limit trial recruitment, resulting in insufficient sample sizes and potentially invalid inference. Incorporating external trial data with similar characteristics (treatments, diseases, biomarkers, etc.) into the analysis appears promising for addressing these issues. Transfer learning, which in our context utilizes external trials as the source domain and current trials as the target domain, may offer a viable approach. In this paper, we present a formal framework for applying transfer learning to the analysis of clinical trials, considering three key perspectives: transfer algorithm, theoretical foundation, and inference method. To cover broad types of randomized trials, we study this problem under stratified randomization, or more generally, covariate-adaptive randomization. For the algorithm, we adopt a parameter-based transfer learning approach to enhance the lasso-adjusted stratum-specific estimator developed for estimating the treatment effect. A key component in constructing the transfer learning estimator is deriving the regression coefficient estimation within each stratum, accounting for the bias between source and target data. To provide a theoretical foundation, we derive the convergence rate for the estimated regression coefficients and subsequently establish the asymptotic normality for the transfer learning estimator. Our results show that when external trial data resemble current trial data, the sample size requirements can be reduced compared to using only current trial data. Finally, we propose a consistent nonparametric variance estimator to facilitate inference that is robust to model misspecifications and applicable to various commonly used randomization procedures. Numerical studies demonstrate the effectiveness and robustness of our proposed estimator across various scenarios. Our study highlights the potential of transfer learning in analyzing randomized clinical trials.

Keywords: Covariate-adaptive randomization; External trial data; Lasso; Robust inference; Transfer learning.

1. Introduction

Randomization is the gold standard in clinical trials and other intervention studies, such as online A/B tests and experiments in development economics. After randomization, a treatment effect estimator that can provide valid inference should be constructed. However, randomized clinical trials may suffer from rising costs, ethical concerns, and recruitment difficulties, especially when the trial involves rare or life-threatening diseases. Consequently, the inference may become invalid due to the lack of sufficient experimental units, i.e., patients enrolled in the trial. To solve this problem, data from external trials are encouraged to assist in analyzing the current trial (Food and Administration, 2022). Including these external data, such as historical control data or supplementary trials, may improve the precision of estimates in the current trial (Pocock, 1976; Boatman et al., 2021; Liu et al., 2022). The use of external trial data is most promising when the current trial and the external trial share similar drugs, biomarkers, or treatment procedures. For example, when analyzing the Covidicus trial, researchers leveraged external Recovery trial data to evaluate the efficacy of dexamethasone, a common treatment in both trials, for critically ill COVID-19 patients, although the Bayesian methods used may be sensitive to parametric model settings (Chevret et al., 2022). The transferability induced by the similarities in study characteristics provides a natural foundation for using transfer learning in these settings.

The recent development of transfer learning, especially in the field of statistics, offers insights into incorporating external trial data into the analysis of current trial data. Transfer learning (Torrey and Shavlik, 2010) uses information from similar source data to support the analysis of target data. To date, transfer learning has been successfully applied in many fields, such as natural language processing, computer vision, and epidemiology. Due to its empirical success, transfer learning has received considerable attention in statistical disciplines. For example, it has been studied in a variety of problems, such as classification (Cai and Wei, 2021), nonparametric regression (Cai and Pu, 2024), and contextual multi-armed bandits (Cai et al., 2024). Most relevant to our work, due to the availability of a large number of baseline covariates in modern clinical trials, transfer learning, when applied to high-dimensional problems, has been shown to improve the convergence rate of regression coefficients compared to using only the target data (Li et al., 2022; Tian and Feng, 2023). Moreover, transfer learning can also improve estimation and inference accuracy in the Gaussian graphical model (Li et al., 2023). These works inspire us to apply transfer learning to estimating treatment effects in clinical trials, especially in a high-dimensional setting.

While incorporating external trial data through transfer learning is conceptually intuitive, ensuring robust inference is pivotal for its acceptance in clinical trial practice, especially under covariate-adaptive randomization. Covariate-adaptive randomization is commonly used in clinical trials to produce a more balanced treatment allocation within strata formed by covariates, such as stratified block randomization (Zelen, 1974) and minimization method (Pocock and Simon, 1975). According to recent surveys, covariate-adaptive randomization is implemented in about 80% of trials (Ciolino et al., 2019). The inference under covariate-adaptive randomization has been challenging due to the dependence between treatment assignments. Various valid tests have been proposed for different allocation methods under model assumptions of the true data generation process (e.g., Ma et al.,

2015, 2022b; Wang et al., 2023; Shao et al., 2010). In practice, the data-generating model in clinical trials is usually unknown, which makes robust treatment effect estimation particularly important. Here, robustness means that the estimator can provide valid inference even when the working regression models are misspecified. Recently, Bugni et al. (2018) proposed robust treatment effect estimators based on several regression models adjusting for strata variables, where the regression models can be misspecified. Subsequently, many researchers have adjusted additional baseline covariates in the regression models to further improve efficiency (Ma et al., 2022a; Liu et al., 2023; Ye et al., 2023, 2022; Gu et al., 2023). Notably, under a high-dimensional setting, Liu et al. (2023) applied lasso regression to obtain a lasso-adjusted stratum-specific treatment effect estimator, which is optimal among the class of all regression-adjusted estimators. However, the validity of the stratum-specific estimator requires sufficiently many units in each stratum to ensure the convergence of the estimated projection coefficient, which may fail in high-dimensional settings, especially when there are many small strata. One possible solution is to incorporate external data through transfer learning to improve the estimation of the projection coefficient. For more discussion on covariate-adaptive randomization and its inference, please refer to Section 1.1.

To fulfill our mission, challenges from three perspectives need to be tackled: *transfer algorithm*, *theoretical foundation*, and *inference method*. Firstly, a transfer learning approach must be tailored to the context of covariate-adaptive randomization to ensure proper knowledge or feature transfer from the source (external trial) to the target (current trial) data, without compromising the performance on the target task in terms of the robustness of treatment effect estimation. Secondly, providing theoretical guarantees that the treatment effect estimation will benefit from transfer learning is crucial, given the dependence in treatment assignments across experimental units resulting from covariate-adaptive randomization. Finally, valid inference is critical for accurate treatment effect estimation. In previous transfer learning studies, only the inference for individual components of regression coefficients was considered. We expect our transfer learning approach to produce valid inference that is applicable to commonly used randomization methods and robust against model misspecification.

In this paper, we propose a robust treatment effect estimator via transfer learning to address the aforementioned issues. For the first challenge, we apply a parameter-based transfer learning approach (Li et al., 2022) to enhance the lasso-adjusted stratum-specific treatment effect estimator developed for covariate-adaptive randomization (Liu et al., 2023). Importantly, this method also strengthens data privacy and security, which is crucial in clinical trials due to the sensitivity of patient information. Specifically, we begin by using lasso regression to obtain regression coefficients from the source data within each stratum. We then correct the bias in these regression coefficients between the source and target data by applying lasso regression to the target data. The target regression coefficient estimator is obtained by adding the bias estimator to the source regression coefficient estimator. Finally, the transfer learning-enhanced treatment effect estimator (referred to hereafter as the transfer learning estimator for simplicity) is constructed by substituting the original regression coefficients in the stratum-specific estimator with these bias-corrected target regression coefficients.

In previous transfer learning frameworks, the units were typically assumed to be independent and identically distributed (i.i.d.). However, covariate-adaptive randomization

imposes a complex dependence structure between covariates and treatment assignments, leading to different strategies for obtaining l_1 convergence rate of the target regression coefficient estimator. Specifically, we incorporate the coupling framework proposed by Bugni et al. (2018, 2019) and prove the restricted strong convexity condition under covariate-adaptive randomization to derive the l_1 convergence rate. Under mild conditions, we show that, when the source and target data are similar, the l_1 convergence rate of the regression estimators is faster than the convergence rate in Liu et al. (2023). This result leads to an improvement in inference based on our proposed estimator, which requires fewer units in the target data. Notably, our proposed estimator achieves optimal efficiency for a regression-adjusted estimator under covariate-adaptive randomization with multiple treatments (Gu et al., 2023).

Finally, we facilitate valid inference by developing a consistent nonparametric variance estimator for the transfer learning estimator. In general, inference under transfer learning for high-dimensional data is non-trivial. However, thanks to our framework, we can preserve robust inference for the treatment effect, avoiding negative transfer when incorporating source data. Moreover, the proposed inference procedure is widely applicable to a broad class of covariate-adaptive randomization procedures. Additionally, the robustness of our transfer learning approach can adapt to scenarios where the covariate distributions or functional forms of data generation models may differ between the source and target data.

1.1 Related work

Covariate-adaptive randomization and its inference. Covariate-adaptive randomization aims to provide more balanced treatment allocation with respect to baseline covariates. For example, stratified block randomization (Zelen, 1974) defines sets of strata based on covariates and allocates experimental units within each stratum using block randomization. Minimization methods (Taves, 1974; Pocock and Simon, 1975) have been proposed to balance covariates over their margins. This approach has been generalized to control various types of imbalance measures, including overall and within-stratum imbalance measures (Hu and Hu, 2012; Hu et al., 2023). Other commonly used covariate-adaptive randomization approaches include stratified biased coin design (Efron, 1971; Shao et al., 2010) and model-based approaches (Begg and Iglewicz, 1980; Atkinson, 1982). We refer readers to Rosenberger and Lachin (2015) for a detailed discussion of these methods.

Our theoretical analysis of the inference for the regression-adjusted treatment effect estimators is related to recent studies that have shown that under covariate-adaptive randomization, regression-adjusted treatment effect estimators are valid for inference. Bugni et al. (2018) proposed robust treatment effect estimators that are resistant to model misspecification with regression adjustment for stratification, which can produce valid inference. With the inclusion of additional baseline covariates, stratum-common and stratum-specific treatment effect estimators have been proposed to improve the efficiency of the estimators (Ye et al., 2023, 2022; Ma et al., 2022a). Gu et al. (2023) showed that the stratum-specific estimator could gain efficiency compared to the stratum-common estimator under multiple treatments, regardless of whether the allocation ratios are the same across strata. In high-dimensional settings, Liu et al. (2023) obtained the regression coefficient estimator by lasso and replaced the original coefficient estimator in stratum-common and stratum-

specific estimators to get lasso-adjusted estimators. Under regular conditions on sparsity, they showed that the lasso-adjusted estimators could achieve the same efficiency as the regression-adjusted estimators under two treatments. When the working model is correctly specified to capture the true underlying relationship instead of a linear form, researchers reach a theoretical guarantee for the efficiency gain if the function form is properly estimated (Rafi, 2023; Bannick et al., 2025; Tu et al., 2024).

Transfer learning. Transfer learning is widely applied in many fields, such as computer vision (Saenko et al., 2010; Donahue et al., 2014), natural language processing (Ruder et al., 2019; Wolf et al., 2020), and biomedical analysis (Schweikert et al., 2008; Petegrosso et al., 2017). We refer readers to Pan and Yang (2009); Weiss et al. (2016); Zhuang et al. (2020) for comprehensive surveys on transfer learning. Based on the type of knowledge being transferred, transfer learning problems can be categorized into several subproblems: instance-based, feature-based, parameter-based and relation-based problems (Pan and Yang, 2009). Our paper focuses on the parameter-based problem as we transfer the regression coefficients from the source data to the target data. Due to its successful empirical performance, transfer learning draws extensive attention in the statistics field. In addition to the work we have mentioned, transfer learning has also been applied to statistical problems such as high-dimensional quantile regression (Huang et al., 2022; Jin et al., 2024) and precision medicine (Wu and Yang, 2023).

Bayesian historical borrowing. In many cases, historical trial data with settings similar to the current trial data are available. Incorporating these historical data may improve the precision of estimations in the current trial (Pocock, 1976). As the treatments usually differ between trials, researchers mainly focused on using historical control data. However, differences in population distribution may lead to heterogeneity between the historical and current data. Under the Bayesian framework, researchers proposed several methods to address the issue, such as using power prior (Ibrahim and Chen, 2000; Ibrahim et al., 2015), meta-analytic predictive prior (Schmidli et al., 2014) and hierarchical models (Neuenschwander et al., 2010; Hobbs et al., 2011). However, almost all existing Bayesian methods depend on the choice of prior distributions and the parametric model assumptions, potentially introducing bias and subjectivity into the statistical inference. Furthermore, previous studies only used historical control data, ignoring the potential improvement from using historical treatment data.

Combining clinical trial and observation study. Another source of external data is real-world data, which primarily comes from observational studies, such as electronic health records, disease registry databases, wearable devices, etc. Observational studies typically have a much larger number of patients than clinical trials, which can empower the inference. However, they have disadvantages such as selection bias and unmeasured confounding, which may lead to biased estimates and identification problems for causal effects (Rubin, 1972; Colnet et al., 2024; Dahabreh and Bibbins-Domingo, 2024). Therefore, researchers have focused on effectively combining observational studies with clinical trials to leverage their respective strengths and compensate for weaknesses. For example, the integration can enhance the generalizability of clinical trials (Stuart et al., 2011; Dahabreh et al., 2020; Lee et al., 2023), help identify complex causal effects in observation studies (Athey et al., 2020; Jung et al., 2023; Jung and Bellot, 2024), improve estimation efficiency of treatment effect heterogeneity (Kallus et al., 2018; Yang et al., 2023, 2025; Wu and Yang, 2022), and identify

optimal individual regimes (Wu and Yang, 2023; Chu et al., 2023). For more discussion on combining randomized trials and observational studies, please refer to Colnet et al. (2024).

1.2 Paper outline

The remainder of this paper is organized as follows. Section 2 introduces the framework and notations for covariate-adaptive randomization. Before the main results, we generalize the lasso-adjusted treatment effect estimator for multiple treatments in Section 3. In Section 4, we develop our transfer learning algorithm and the asymptotic properties of the proposed transfer learning estimator. We propose nonparametric variance estimators for valid inference in Section 5. Simulation results are in Section 6. Section 7 provides a clinical trial example, and Section 8 summarizes our discussion. The Appendix includes a notation table, general asymptotic results for the regression-adjusted estimator, proofs of the main theorems, and additional simulations.

2. Framework and notation

We start by introducing a covariate-adaptive randomization procedure with n units. Suppose that the treatment a takes values in the set $\mathcal{A} = \{1, 2, \ldots, A\}$. Let $a = 0$ be the control group and $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$. Let A_i be the indicator variable, such that $A_i = a$ indicates that the ith unit is assigned to the ath treatment. Let $n_a = \sum_{i=1}^{n} I(A_i = a)$ be the number of units in each treatment group, where $I(\cdot)$ is the indicator function. Let B_i denote the stratum label, which takes values in $\mathcal{K} = \{1, 2, \ldots, K\}$. Here, K represents the total number of strata, which is fixed and finite. Define $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ as the p-dimensional additional baseline covariates not used in the randomization procedure. We consider a high-dimensional setting where p tends to infinity as n goes to infinity. Let $p_{[k]} = \mathrm{pr}(B_i = k) > 0$ be the expected proportion of stratum k and $\pi_{[k]a} = \mathrm{pr}(A_i = a \mid B_i = k)$ be the expected proportion of treatment a in stratum k. Let $n_{[k]} = \sum_{i=1}^{n} I(B_i = k)$ be the number of units in stratum k and $n_{[k]a} = \sum_{i=1}^{n} I(A_i = a, B_i = k)$ be the number of units in stratum k and treatment group a. Let $p_{n[k]} = n_{[k]}/n$ be the estimated proportion of stratum k.

We use the Neyman-Rubin model to define potential outcomes and treatment effects (Neyman, 1923; Rubin, 1974). Let $\{Y_i(a), a \in \mathcal{A}_0\}$ be the potential outcomes and $Y_i = \sum_{a \in \mathcal{A}_0} I(A_i = a) Y_i(a)$ be the observed outcomes. Let $\{W_i\}_{i=1}^{n} = \{Y_i(0), \ldots, Y_i(A), B_i, X_i\}_{i=1}^{n}$ be independent and identically distributed (i.i.d.) samples from the population distribution $W = \{Y(0), \ldots, Y(A), B, X\}$. Our goal is to estimate the treatment effect $\tau_a = E\{Y_i(a) - Y_i(0)\}$ for all $a \in \mathcal{A}$.

In the context of transfer learning, in addition to the observations $\{Y_i, X_i, A_i, B_i\}_{i=1}^{n}$ from the target population, we also have external trial data as source data. The treatments and the range of stratum labels in the source data are denoted as $\mathcal{A}_0^{(s)}$ and $\mathcal{K}^{(s)}$. Let $\{W_i^{(s)}\}_{i=1}^{n^{(s)}} = \{Y_i^{(s)}(0), \ldots, Y_i^{(s)}(A), B_i^{(s)}, X_i^{(s)}\}_{i=1}^{n^{(s)}}$ be $n^{(s)}$ i.i.d. samples from the source data population $W^{(s)} = \{Y^{(s)}(0), \ldots, Y^{(s)}(A), B^{(s)}, X^{(s)}\}$, which is independent of W and $(A_1, \ldots, A_n)$. Similarly, the notations for source data are superscripted by (s), such as $n_{[k]}^{(s)}$ indicates the number of units in stratum k in source data.

We denote the set of random variables with at least one positive stratum-specific variance as $\mathcal{R}_2 = \{V : \max_{k \in \mathcal{K}} \mathrm{var}(V | B_i = k) > 0\}$. Also, we assume that the stratum-specific

covariance matrix

$$\Sigma_{[k]XX} = E[\{X_i - E(X_i|B_i = k)\}\{X_i - E(X_i|B_i = k)\}^{\mathrm{T}}|B_i = k],\ k = 1, 2, \ldots, K,$$

is strictly positive definite. The following assumptions are made for the target data generation and covariate-adaptive randomization procedures.

Assumption 1 *$E\{Y_i^2(a)\} < \infty$ and $Y_i(a) \in \mathcal{R}_2$, for all $a \in \mathcal{A}_0$. $\{X_i\}_{i=1}^n$ are uniformly bounded by a constant, which means that there exists a constant M independent of n, such that $\max_{i=1,\ldots,n;j=1,\ldots,p} |X_{ij}| \leq M$.*

Assumption 2

1. *Conditional on $\{B_1, \ldots, B_n\}$, $\{A_1, \ldots, A_n\}$ is independent of $\{Y_i(0), \ldots, Y_i(A), X_i\}_{i=1}^n$.*
2. *$n_{[k]a}/n_{[k]} \xrightarrow{P} \pi_{[k]a}$ as $n \to \infty$, for all $a \in \mathcal{A}_0$ and $k \in \mathcal{K}$.*

The assumptions are similar to those proposed in Gu et al. (2023). The difference from Assumption 1 is that X_i is high-dimensional, so we assume that X_i is uniformly bounded, as motivated by Liu et al. (2023). This assumption may be stringent for X_i; however, it helps relax the requirements for the approximation error between the potential outcomes and the working model.

Assumption 2.1 requires that given the stratification B_i, treatment assignment A_i is conditionally independent of both the potential outcomes and covariates. Assumption 2.2 further requires that, in each stratum, the observed allocation proportion converges to the pre-specified target proportion. Assumption 2 is guaranteed by the mechanism of several well-known covariate-adaptive randomization methods, including stratified block randomization (Zelen, 1974) and stratified biased-coin randomization (Shao et al., 2010), and has been frequently used in recent studies on inference under covariate-adaptive randomization (Bugni et al., 2019; Ma et al., 2022a; Ye et al., 2022). Assumption 2.1 also guarantees the identification of the treatment effect, which is detailed in Remark 3. When $\pi_{[k]a}$ is identical across strata, Pocock and Simon's minimization (Pocock and Simon, 1975) also satisfies this assumption (Hu et al., 2023). Assumption 2 is also satisfied by simple randomization and permuted block randomization. Together with covariate-adaptive randomization, these designs account for approximately 90% of clinical trials (Ciolino et al., 2019).

These assumptions may fail in certain cases. Assumption 2.1 may be violated under treatment contamination or noncompliance (Sussman and Hayward, 2010; Huang, 2018), where patients do not receive the treatment to which they were randomized. Assumption 2.2 may not hold due to limited sample sizes in each stratum, for example, when a study site with many levels is used for stratification. Extending the results of this paper to these cases is beyond its scope and left for future research.

We impose similar assumptions on the source data; see Assumption 10 for details. Next, we introduce some notations. Let $r_i(a)$, for $i = 1, 2, \ldots, n$, $a \in \mathcal{A}_0$, be a transformed outcome, such as potential outcomes $Y_i(a)$, covariates X_i or their linear combinations. Also, we denote $\tilde{r}_i(a) = r_i(a) - E\{r_i(a)|B_i\}$. The sample means of the transformed outcomes are defined as $\bar{r}_a = (1/n_a) \sum_{i=1}^n I(A_i = a) r_i(a)$, $\bar{r}_k = (1/n_{[k]}) \sum_{i=1}^n I(B_i = k) r_i(a)$ and $\bar{r}_{[k]a} = (1/n_{[k]a}) \sum_{i=1}^n I(A_i = a, B_i = k) r_i(a)$. The population variance of a transformed

outcome $r_i(a)$ is denoted as $\sigma^2_{r(a)}$. We define $\sigma^2_{[k]r(a)} = \text{var}\{r_i(a) \mid B_i = k\}$ as the population variance of a transformed outcome in stratum k. Let $\|\cdot\|_1$ denote the l_1-norm of a vector. To enhance the clarity, we present a notation table (Table 3) in the Appendix A to summarize the frequently used notations in this paper.

Remark 3 *Based on Assumption 2.1, conditional on the stratum variable B_i, the treatment assignment A_i is independent of the potential outcomes. This condition implies that, within each stratum $B_i = k$, the conditional expectation of the observed outcome under treatment a equals the conditional expectation of the corresponding potential outcome:*

$$E[Y_i(a)|B_i = k] = E[Y_i(a)|A_i = a, B_i = k] = E[Y_i|A_i = a, B_i = k].$$

Consequently, the treatment effect for each treatment $a \in \mathcal{A}$ can be identified by taking the expectation over B_i as

$$\begin{aligned}\tau_a &= E[E\{Y_i(a)|B_i\}] - E[E\{Y_i(0)|B_i\}] \\ &= E[E\{Y_i|A_i = a, B_i\}] - E[E\{Y_i|A_i = 0, B_i\}] \\ &= \sum_{k=1}^{K} p_{[k]}[E\{Y_i|A_i = a, B_i = k\} - E\{Y_i|A_i = 0, B_i = k\}].\end{aligned}$$

3. Lasso-adjusted estimator under multiple treatments

The stratum-specific estimator under multiple treatments has been discussed in Gu et al. (2023). This estimator can guarantee an efficiency gain among the estimators they considered, regardless of whether the target allocation ratios across strata are the same or different. However, the OLS estimator does not work well in high-dimensional settings due to overfitting. Therefore, selecting covariates or some forms of regularization should be considered, such as lasso (Tibshirani, 1996). Liu et al. (2023) considered the stratum-specific lasso-adjusted treatment effect estimator and derived its asymptotic normality when there are two treatments. In this section, we generalize the stratum-specific lasso-adjusted treatment effect estimator to multiple treatments and derive its asymptotic normality.

We define the stratum-specific projection coefficient $\beta_{[k]\text{proj}}(a)$ as

$$\beta_{[k]\text{proj}}(a) = \arg\min_{\beta \in \mathbb{R}^p} E([\{Y_i(a) - E(Y_i(a)|B_i = k)\} - \{X_i - E(X_i|B_i = k)\}^{\mathrm{T}}\beta]^2|B_i = k).$$

The stratum-specific lasso-adjusted vectors can be calculated as

$$\hat{\beta}_{[k]\text{lasso}}(a) = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}\beta\}^2 + \lambda_{[k]a}\|\beta\|_1,$$

where $\lambda_{[k]a} > 0$ is a regularization parameter controlling the strength of the l_1 penalty. Therefore, the treatment effect estimator for each treatment a is

$$\hat{\tau}_{\text{lasso,a}} = \sum_{k \in \mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{lasso}}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{lasso}}(0)\}\big],$$

and the stratum-specific lasso-adjusted treatment effect estimator is

$$\hat{\tau}_{\text{lasso}} = (\hat{\tau}_{\text{lasso},1}, \hat{\tau}_{\text{lasso},2}, \ldots, \hat{\tau}_{\text{lasso,A}})^{\text{T}}.$$

To investigate the performance of $\hat{\tau}_{\text{lasso}}$, we define the transformed outcome $r_{i,\text{proj}}(a) = Y_i(a) - X_i^{\text{T}}\beta_{[k]\text{proj}}(a)$, for $i \in [k], A_i = a$. Here $i \in [k]$ means that unit i belongs to stratum k. Let $\beta_{j,[k]\text{proj}}(a)$ be the jth element of $\beta_{[k]\text{proj}}(a)$ and $S_{[k]} = \bigcup_{a\in\mathcal{A}_0}\{j \in \{1,\ldots,p\} : \beta_{j,[k]\text{proj}}(a) \neq 0\}$ be the union of the support of $\beta_{[k]\text{proj}}(a)$ for all $a \in \mathcal{A}_0$. Also, let $s_{[k]} = |S_{[k]}|$ be the number of relevant covariates in stratum k. We make the following assumptions within each stratum to obtain the l_1 convergence rate of vectors $\hat{\beta}_{[k]\text{lasso}}(a)$.

Assumption 4 *The minimum eigenvalue of the stratum-specific covariance matrix $\Sigma_{[k]XX}$ is $\Lambda_{\min,[k]}$ and $\Lambda_{\min,[k]} \geq \kappa_l > 0$, where κ_l is a constant independent of n.*

Assumption 5 *There exist constants $c_{\lambda[k]} > 0$, $c_M > 1$, and a sequence $M_{n[k]}$ in n for each stratum k such that $M_{n[k]} \to \infty$ with $M_{n[k]}s_{[k]}^2(\log p)^2/n \to 0, k = 1, 2, \ldots, K$. Then the tuning parameters $\lambda_{[k]a}$ belong to the interval*

$$\left[4\left(\frac{c_{\lambda[k]}p_{[k]}M_{n[k]}}{\pi_{[k]a}}\right)^{1/2}\left(\frac{\log p}{n}\right)^{1/2}, 4c_M\left(\frac{c_{\lambda[k]}p_{[k]}M_{n[k]}}{\pi_{[k]a}}\right)^{1/2}\left(\frac{\log p}{n}\right)^{1/2}\right].$$

Assumption 4 ensures that, within each stratum, the stratum-specific covariance matrix $\Sigma_{[k]XX}$ is positive definite with eigenvalues bounded away from zero. This prevents perfect collinearity among covariates and guarantees that stratum-specific estimators based on $\Sigma_{[k]XX}^{-1}$ are well-defined and stable. Meanwhile, it implies that the sample-version restricted eigenvalue condition holds with probability tending to one when $n \to \infty$, a typical assumption for obtaining l_1 convergence rate of the lasso under high-dimensional sparse linear regression models (e.g., Zhang and Huang, 2008; Huang et al., 2009; Meinshausen and Yu, 2009; Negahban et al., 2009). Assumption 4 closely resembles Assumption 4 in Liu et al. (2023), with minor modifications, as they require a restricted eigenvalue condition for $\Sigma_{[k]XX}$, whereas we assume full positive definiteness to ensure well-conditioning in all directions, including those induced by the source data. Assumption 4 is also commonly used in transfer learning settings (Li et al., 2022; Tian and Feng, 2023).

Assumption 5 specifies a range for the lasso tuning parameters $\lambda_{[k]a}$ in each stratum. Intuitively, it ensures that $\lambda_{[k]a}$ is large enough to control estimation error for sparse high-dimensional regression, but not so large as to overly shrink important coefficients, leading to the requirement that the tuning parameter be of the order of approximately $\{(\log p/n)\}^{1/2}$. It is a standard condition in high-dimensional analysis (Bickel et al., 2009; Bühlmann and van de Geer, 2011). It may only be restrictive in extremely dense or ultra-high-dimensional settings. An additional sequence $M_{n[k]}$, which diverges slowly in n for each stratum k (e.g., $M_{n[k]} = \log\log n$), is introduced to relax the condition on the transformed outcome $r_{i,\text{proj}}(a)$, thereby ensuring the robustness of our estimator. The scaling with $s_{[k]}$, p, and n guarantees consistent and reliable estimation as the sample size increases, which is essential for valid high-dimensional inference (van de Geer et al., 2014).

Let 1_A denote an A-dimensional vector with ones and $m_{[k]}(a) = E[Y_i(a)|B_i = k] - E[Y_i(a)]$. The asymptotic variance of $\hat{\tau}_{\text{lasso}}$ depends on the following quantities:

$$V_{H\tilde{Y}} = \sum_{k\in\mathcal{K}} p_{[k]} \left(m_{[k]}(a) - m_{[k]}(0) : a \in \mathcal{A}\right) \times \left(m_{[k]}(a') - m_{[k]}(0) : a' \in \mathcal{A}\right)^{\mathrm{T}},$$

$$V_{\tilde{r}_{\text{proj}}} = \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]0}} \sigma^2_{[k]\tilde{r}_{\text{proj}}(0)} 1_A 1_A^{\mathrm{T}} + \text{diag}\left\{\sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}} \sigma^2_{[k]\tilde{r}_{\text{proj}}(a)} : a \in \mathcal{A}\right\},$$

$$V_{\tilde{X}} = \Big\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta_{[k]\text{proj}}(a) - \beta_{[k]\text{proj}}(0)\}^{\mathrm{T}} \Sigma_{[k]XX} \{\beta_{[k]\text{proj}}(a') - \beta_{[k]\text{proj}}(0)\} : (a, a') \in \mathcal{A} \times \mathcal{A} \Big\}.$$

We assume that $V_{\tilde{r}_{\text{proj}}}$, $V_{\tilde{X}}$ have finite limits and the limit of $V_{\tilde{r}_{\text{proj}}} + V_{\tilde{X}} + V_{H\tilde{Y}}$, denoted by V_{proj}, is positive. The following theorem establishes the l_1 convergence rate of $\hat{\beta}_{[k]\text{lasso}}(a)$ and the asymptotic normality of $\hat{\tau}_{\text{lasso}}$.

Proposition 6 *Under Assumptions 1–5, if $r_{i,\text{proj}}(a) \in \mathcal{R}_2$, $E\{r^2_{i,\text{proj}}(a)\} < \infty$ for $a \in \mathcal{A}_0$, then $\|\hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)\|_1 = O_p[s_{[k]}(M_{n[k]}(\log p)/n)^{1/2}]$, for $k \in \mathcal{K}$ and $a \in \mathcal{A}_0$. Moreover,*

$$\sqrt{n}(\hat{\tau}_{\text{lasso}} - \tau) \xrightarrow{d} \text{N}(0, V_{\text{proj}}).$$

Remark 7 *A general theory in Liu et al. (2023) can guarantee the asymptotic normality for the regression-adjusted estimator as long as*

$$\sqrt{n}(\bar{X}_{[k]1} - \bar{X}_{[k]0})^{\mathrm{T}} \{\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)\} = o_p(1), \quad for \quad a = 0, 1,$$

is satisfied for some vector $\beta_{[k]}(a)$ and its estimator $\hat{\beta}_{[k]}(a)$. Under this condition, a regression-adjusted treatment effect estimator is asymptotically normal with mean $\tau = \tau_1 - \tau_0$. Therefore, the crucial step for proving the asymptotic normality is to derive the l_1 convergence rate of the lasso-adjusted vectors. We prove our result by generalizing the existing theory to settings with multiple treatments.

Remark 8 *The unadjusted estimator was studied in Bugni et al. (2019), which is*

$$\hat{\tau}_{\text{ben,a}} = \sum_{k\in\mathcal{K}} p_{n[k]} (\bar{Y}_{[k]a} - \bar{Y}_{[k]0}), \text{ for } a \in \mathcal{A},$$

and

$$\hat{\tau}_{\text{ben}} = (\hat{\tau}_{\text{ben},1}, \hat{\tau}_{\text{ben},2}, \ldots, \hat{\tau}_{\text{ben,A}})^{\mathrm{T}}.$$

To improve the efficiency, Gu et al. (2023) proposed a stratum-specific treatment effect estimator using linear regression in each stratum and treatment. Our proposed lasso-adjusted estimator shares the same asymptotic covariance matrix as the stratum-specific estimator, which had been certified to gain efficiency compared with the unadjusted estimator in Gu et al. (2023). This estimator will be used as the benchmark estimator in the numerical studies in Sections 6 and 7.

Remark 9 *Building upon the efficiency gain from the regression-adjusted estimator in Remark 8, the interpretability of the lasso adjustment in our high-dimensional framework is tied to the estimation of the sparse projection coefficient. Specifically, estimating this parsimonious linear representation enables the estimator to achieve the efficiency gains. In Appendix E.2, we conduct a corresponding interpretability study to demonstrate that the proposed method can select relevant baseline variables to improve efficiency.*

4. Transfer learning for treatment effect estimation

4.1 Context and algorithm

The validity of the stratum-specific estimator requires a sufficient number of units in each stratum, which may fail in high-dimensional settings, especially when there are many small strata. To obtain valid inference, we use transfer learning by leveraging source data in the estimation. In this paper, we consider the scenario where only one source dataset is used in the transfer learning algorithm. If multiple candidate source datasets exist, the one with the highest similarity to the target data may be pre-selected. In practice, this could be achieved through expert knowledge or data-driven methods. See Section 8 for a discussion on cases involving multiple source datasets. We make the following assumptions about the source data.

Assumption 10

1. *$X_i^{(s)}$ is a p-dimensional vector, $\mathcal{A}_0^{(s)} = \mathcal{A}_0$, and $\mathcal{K}^{(s)} = \mathcal{K}$.*

2. *Source data $\{W_i^{(s)}\}_{i=1}^{n^{(s)}}$ satisfy Assumption 1. In details, $E\big[\{Y_i^{(s)}(a)\}^2\big] < \infty$ and $Y_i^{(s)}(a) \in \mathcal{R}_2$, for all $a \in \mathcal{A}_0^{(s)}$. $\{X_i^{(s)}\}_{i=1}^{n^{(s)}}$ are uniformly bounded by a constant, which means that there exists a constant M independent of $n^{(s)}$ such that $\max_{i=1,\ldots,n^{(s)};j=1,\ldots,p} |X_{ij}^{(s)}| \leq M$.*

3. *The randomization procedure satisfies Assumption 2. In details, conditional on $\{B_1^{(s)}, \ldots, B_{n^{(s)}}^{(s)}\}$, $\{A_1^{(s)}, \ldots, A_{n^{(s)}}^{(s)}\}$ is independent of $\{Y_i^{(s)}(0), \ldots, Y_i^{(s)}(A), X_i^{(s)}\}_{i=1}^{n^{(s)}}$.*

4. *$n_{[k]a}^{(s)}/n_{[k]}^{(s)} \xrightarrow{P} \pi_{[k]a}^{(s)}$ as $n^{(s)} \to \infty$, for all $a \in \mathcal{A}_0^{(s)}$ and $k \in \mathcal{K}^{(s)}$.*

Assumption 10.1 requires that the source and target data share the same set of treatments and stratum labels. Specifically, having the same treatments may indicate that the two trials use the same drugs or closely related ones, such as drugs targeting the same biomarker. Likewise, having the same stratum labels means that both trials use the same covariates for stratification in the randomization process. We allow the distribution of $X_i^{(s)}$ to differ from that of X_i, which ensures that our theoretical results remain applicable under covariate shift. Assumption 10.2 is the same as Assumption 1. The upper bound M is uniform across both source and target data. The covariate-adaptive randomization procedure depends on Assumption 10.3 and 10.4. Here we allow the randomization procedures in the two clinical trials to be different, as $\pi_{[k]a}^{(s)}$ may not be equal to $\pi_{[k]a}$.

Before presenting the formal algorithm, we define the key ingredients in the transfer learning procedure. Let $\beta^{(s)}_{[k]\text{proj}}(a)$ denote the stratum-specific projection coefficient in the source data, defined by

$$\beta^{(s)}_{[k]\text{proj}}(a) = \arg\min_{\beta\in\mathbb{R}^p} E\bigg([\{Y_i^{(s)}(a) - E(Y_i^{(s)}(a)|B_i^{(s)}=k)\} - \{X_i^{(s)} - E(X_i^{(s)}|B_i^{(s)}=k)\}^{\mathrm{T}}\beta]^2 | B_i^{(s)}=k\bigg).$$

The bias induced by using the source data is defined as the difference between the target and source projection coefficients, which is given for each stratum k by

$$\delta_{[k]}(a) = \beta_{[k]\text{proj}}(a) - \beta^{(s)}_{[k]\text{proj}}(a).$$

Our transfer learning algorithm is implemented in a stratum-wise manner. Within each stratum, we adopt the two-stage idea of Li et al. (2022), first leveraging information from the source data and then using the target data to correct for bias. Following this overview, we formally present the algorithm in Algorithm 1. In Algorithm 1, we first estimate $\hat{\beta}^{(s)}_{[k]\text{proj}}(a)$ using the source data in each stratum k

$$\hat{\beta}^{(s)}_{[k]\text{proj}}(a) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n^{(s)}_{[k]a}} \sum_{i=1}^{n^{(s)}} I(A_i^{(s)}=a, B_i^{(s)}=k)\{Y_i^{(s)} - \bar{Y}^{(s)}_{[k]a} - (X_i^{(s)} - \bar{X}^{(s)}_{[k]a})^{\mathrm{T}}\beta\}^2 + \lambda^{(s)}_{[k]a}\|\beta\|_1,$$

and then obtain the bias estimator $\hat{\delta}_{[k]}(a)$ from the target data in each stratum k

$$\hat{\delta}_{[k]}(a) = \arg\min_{\delta\in\mathbb{R}^p} \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i=a, B_i=k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}(\hat{\beta}^{(s)}_{[k]\text{proj}}(a) + \delta)\}^2 + \lambda_{[k]a}\|\delta\|_1.$$

Combining these two estimators, we define

$$\hat{\beta}_{[k]\text{tl}}(a) = \hat{\delta}_{[k]}(a) + \hat{\beta}^{(s)}_{[k]\text{proj}}(a),$$

and the transfer learning estimator for each treatment $a \in \mathcal{A}$ is then given by

$$\hat{\tau}_{\text{tl,a}} = \sum_{k\in\mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(0)\}\big],$$

and the transfer learning estimator is

$$\hat{\tau}_{\text{tl}} = (\hat{\tau}_{\text{tl},1}, \hat{\tau}_{\text{tl},2}, \ldots, \hat{\tau}_{\text{tl,A}})^{\mathrm{T}}.$$

Remark 11 *Even when two trials evaluate the same drug or target the same biomarker, treatment effects may still differ, for example, because of differences in patient populations and in how the trials are conducted. We capture such differences using the projection coefficients estimated from each dataset, with $\delta_{[k]}(a)$ quantifying the discrepancy between the two domains. This approach, which has been used in recent transfer learning studies (e.g., Li et al., 2022; Tian and Feng, 2023), allows us to borrow information from the source data to improve the estimation of $\beta_{[k]\text{proj}}(a)$, particularly when the target sample size is small.*

Remark 12 *When using multiple datasets, the identification of treatment effects becomes more challenging, as it relies on causal transportability across different experimental domains (Jung et al., 2023; Jung and Bellot, 2024). This issue does not arise in our setting, as our transfer learning algorithm focuses on the treatment effects in the target trial, whose identification is directly ensured by the covariate-adaptive randomization procedure.*

Algorithm 1: Transfer learning algorithm for treatment effect estimation under covariate-adaptive randomization

1. Perform lasso regression on source data in each corresponding stratum k and treatment a. The coefficient estimators are derived by

$$\hat{\beta}^{(s)}_{[k]\text{proj}}(a) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n^{(s)}_{[k]a}} \sum_{i=1}^{n^{(s)}} I(A^{(s)}_i = a, B^{(s)}_i = k)\{Y^{(s)}_i - \bar{Y}^{(s)}_{[k]a} - (X^{(s)}_i - \bar{X}^{(s)}_{[k]a})^{\mathrm{T}}\beta\}^2 + \lambda^{(s)}_{[k]a}\|\beta\|_1.$$

2. Perform lasso regression on the bias δ using target data.

$$\hat{\delta}_{[k]}(a) = \arg\min_{\delta\in\mathbb{R}^p} \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}(\hat{\beta}^{(s)}_{[k]\text{proj}}(a) + \delta)\}^2 + \lambda_{[k]a}\|\delta\|_1.$$

3. Derive the coefficient estimator $\hat{\beta}_{[k]\text{tl}}(a) = \hat{\delta}_{[k]}(a) + \hat{\beta}^{(s)}_{[k]\text{proj}}(a)$.

4. Impute the coefficient estimators into the treatment effect estimator for each $a \in \mathcal{A}$:

$$\hat{\tau}_{\text{tl,a}} = \sum_{k\in\mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(0)\}\big],$$

and the transfer learning estimator is

$$\hat{\tau}_{\text{tl}} = (\hat{\tau}_{\text{tl},1}, \hat{\tau}_{\text{tl},2}, \ldots, \hat{\tau}_{\text{tl,A}})^{\mathrm{T}}.$$

4.2 Theoretical results

We consider $\delta_{[k]}(a)$ by its l_1 norm such that

$$\max_{k\in\mathcal{K},\ a\in\mathcal{A}_0} \|\delta_{[k]}(a)\|_1 \le h.$$

Here we use all components to evaluate the bias as the source data may have different supports from the target data. Recall that the transfer learning estimator for each treatment a is

$$\hat{\tau}_{\text{tl,a}} = \sum_{k\in\mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(0)\}\big],$$

and the transfer learning estimator is

$$\hat{\tau}_{\text{tl}} = (\hat{\tau}_{\text{tl},1}, \hat{\tau}_{\text{tl},2}, \ldots, \hat{\tau}_{\text{tl,A}})^{\mathrm{T}}.$$

To derive the asymptotic behavior of $\hat{\tau}_{\text{tl}}$, we first obtain the l_1 convergence rate of $\hat{\beta}_{[k]\text{tl}}(a)$. Even though a similar result has been derived in Li et al. (2022), it is not sufficient as

they assumed that the observations are independent, while our treatment assignments have dependency due to covariate-adaptive randomization. Consequently, their findings do not directly apply to our setting. To overcome this limitation, we incorporate the coupling framework proposed by Bugni et al. (2018) and derive the restricted strong convexity condition under covariate-adaptive randomization, which is detailed in Lemmas 30 and 31 in the Appendix C. The following assumptions are made for source data to reach our results.

Assumption 13 *The minimum eigenvalue of stratum-specific covariance matrix* $\Sigma^{(s)}_{[k]XX}$ *is* $\Lambda^{(s)}_{\min,[k]}$ *and* $\Lambda^{(s)}_{\min,[k]} \geq \kappa_l > 0$.

Assumption 14 *There exist constants* $c^{(s)}_{\lambda[k]} > 0$, $c^{(s)}_M > 1$ *and a sequence* $M^{(s)}_{n[k]} \to \infty$ *with* $M^{(s)}_{n[k]} s^2_{[k]} (\log p)^2 / n^{(s)} \to 0, k = 1, 2, \ldots, K$, *such that the tuning parameters* $\lambda^{(s)}_{[k]a}$ *belong to the interval*

$$\left[4 \left(\frac{c^{(s)}_{\lambda[k]} p^{(s)}_{[k]} M^{(s)}_{n[k]}}{\pi^{(s)}_{[k]a}} \right)^{1/2} \left(\frac{\log p}{n^{(s)}} \right)^{1/2}, 4c_M \left(\frac{c^{(s)}_{\lambda[k]} p^{(s)}_{[k]} M^{(s)}_{n[k]}}{\pi^{(s)}_{[k]a}} \right)^{1/2} \left(\frac{\log p}{n^{(s)}} \right)^{1/2} \right].$$

Assumptions 13 and 14 are the source data versions of Assumptions 4 and 5 respectively. $M^{(s)}_{n[k]}$ is the same sequence as $M_{n[k]}$ but with $n^{(s)}$ elements. The following theorem shows the l_1 convergence rate of $\hat{\beta}_{[k]\mathrm{tl}}(a)$.

Theorem 15 *Suppose that Assumptions 1–4 and 10–14 hold, and that* $h < s_{[k]} \sqrt{\log p / n}$, $n^{(s)} \gg n$, $h(\log p / n^{(s)})^{1/2} = o(1)$, *and the tuning parameters* $\lambda_{[k]a}$ *belong to the interval*

$$\left[4 \left(\frac{c_{\lambda[k]} p_{[k]} M_{n[k]}}{\pi_{[k]a}} \right)^{1/2} \left(\frac{\log p}{n} \right)^{1/2}, 4c_M \left(\frac{c_{\lambda[k]} p_{[k]} M_{n[k]}}{\pi_{[k]a}} \right)^{1/2} \left(\frac{\log p}{n} \right)^{1/2} \right], \quad k = 1, \ldots, K,$$

where $M_{n[k]} \to \infty$ *with* $M_{n[k]}/n \to 0$. *When* $r_{i,\mathrm{proj}}(a) \in \mathcal{R}_2$ *and* $E\{r_{i,\mathrm{proj}}(a)\}^2 < \infty$ *for* $a \in \mathcal{A}_0$, *we have*

$$\|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 = O_p \left(s_{[k]} \sqrt{\frac{M^{(s)}_{n[k]} \log p}{n + n^{(s)}}} + h \right).$$

Theorem 15 establishes the l_1 convergence rate of $\hat{\beta}_{[k]\mathrm{tl}}(a)$ under mild regularity conditions for $k \in \mathcal{K}, a \in \mathcal{A}_0$. We highlight the advantages of our transfer learning method when a large amount of source data is available and there is sufficient similarity between the two datasets. Specifically, when the source and target data are similar, in the sense that $h < s_{[k]} \sqrt{\log p / n}$, Theorem 15 shows that the convergence rate is nearly

$$O_p[s_{[k]} \{M^{(s)}_{n[k]} (\log p) / (n + n^{(s)})\}^{1/2}].$$

This result implies that our method allows for less sparsity in the target model compared with the lasso-adjusted estimator $\hat{\tau}_{\mathrm{lasso}}$ defined in Section 3, whose corresponding coefficient

estimator has l_1 convergence rate $O_p[s_{[k]}\{M_{n[k]}(\log p)/n\}^{1/2}]$ as shown in Proposition 6. The conditions for achieving this improvement are $n^{(s)} \gg n$ and $h < s_{[k]}\sqrt{\log p/n}$. These conditions can be satisfied when the source data has a large number of observations and is similar to the target data. The similarity requirement $h < s_{[k]}\sqrt{\log p/n}$ ensures that the heterogeneity between the source and target data is smaller than the l_1-convergence rate of applying lasso based on the target data alone. In this regime, transfer learning can improve estimation accuracy; otherwise, when the similarity requirement does not hold, even if $n^{(s)} \gg n$, the l_1 estimation error may increase. To make the results more informative, Appendix D.2 provides both the theoretical bounds and the corresponding finite-sample performances under varying h, demonstrating the convergence results from both theoretical and empirical perspectives. Meanwhile, the convergence rate and the sparsity condition are similar to the Oracle Trans-Lasso estimator in Li et al. (2022) except for a sequence $M^{(s)}_{n[k]}$. This sequence plays as the cost to relax the conditions on the transformed outcome $r_{i,\text{proj}}(a)$. $M^{(s)}_{n[k]}$ can grow very slowly such that $M^{(s)}_{n[k]}/n^{(s)}$ approaches zero when $n^{(s)}$ is large enough, for example, $M^{(s)}_{n[k]} = \log\log n^{(s)}$. Therefore, the main requirement on the tuning parameter $\lambda^{(s)}_{[k]a}$ is of the order $O\{(\log p/n^{(s)})^{1/2}\}$.

Now, we can obtain the asymptotic behavior of $\hat{\tau}_{\text{tl}}$.

Theorem 16 *Suppose that Assumptions 1–4, and 10–14 hold, and that $h < s_{[k]}\sqrt{\log p/n}$, $n^{(s)} \gg n$, $h(\log p)^{1/2} = o(1)$, and the tuning parameters $\lambda_{[k]a}$ satisfy the condition in Theorem 15. When $r_{i,\text{proj}}(a) \in \mathcal{R}_2$ and $E\{r_{i,\text{proj}}(a)\}^2 < \infty$ for $a \in \mathcal{A}_0$, we have*

$$\sqrt{n}(\hat{\tau}_{\text{tl}} - \tau) \xrightarrow{d} \text{N}(0, V_{\text{proj}}).$$

We compare the result in Theorem 16 with the result in Proposition 6 for lasso. The requirement on the sparsity $s_{[k]}$ is relaxed to $s_{[k]}\log p = o(\sqrt{n^{(s)}/M^{(s)}_{n[k]}})$, when $n^{(s)} \gg n$, $h < s_{[k]}\sqrt{\log p/n}$. This result allows fewer units in the target data, especially when many strata exist and a small number of units exist in each stratum. Also, in typical high-dimensional regression models, the valid inference requires $s_{[k]}\log p \ll \sqrt{n}$. In Theorem 16, $s_{[k]}$ can be greater than $\sqrt{n}$ and of order $o(\sqrt{n^{(s)}})$.

Remark 17 *It is worth noting that, although the proposed framework naturally bridges covariate-adaptive randomization and transfer learning, several additional challenges arise when establishing its theoretical properties. In particular, the adaptive procedure induces dependence among units, complicating the derivation of asymptotic results in the transfer learning setting. To address this issue, we generalize the coupling method proposed in Bugni et al. (2018, 2019) to include high-dimensional covariates X_i alongside outcomes Y_i, thereby accounting for the joint dependence structure inherent in covariate-adaptive randomization. Moreover, robustness analysis becomes substantially more difficult than in standard single-domain problems, since our estimator must remain valid even when the working models in either the source or the target domain, or both, are misspecified. These challenges reflect the additional technical considerations required in our framework, as existing approaches often assume correct model specification in at least one domain (Li et al., 2022; Tian and*

Feng, 2023). To this end, we establish a generalized restricted strong convexity condition that accommodates model misspecification in either the source or target domain, or both. This condition provides the theoretical foundation for establishing the l_1-convergence rate and asymptotic normality results in Theorems 15 and 16. To validate these theoretical derivations, we provide an empirical verification of the restricted strong convexity condition in Appendix D.4 under our assumptions.

Remark 18 *Theorems 15 and 16 hold under Assumptions 1 and 10 regarding the uniform boundedness of X_i and Assumptions 4 and 13 concerning the strictly positive definite population covariance matrix $\Sigma_{[k]XX}$. The boundedness assumption is naturally justified in clinical settings because biological indicators such as age and blood pressure reside within finite ranges. Regarding the positive definiteness, this condition is guaranteed by the selection of non-redundant covariates by clinical experts and the subsequent removal of highly correlated features during data pre-processing. These procedures ensure that the chosen covariates are informative and avoid perfect multicollinearity.*

Remark 19 *We discuss the conditions on h in Theorem 16. The typical condition for the inference using the debiased lasso (Zhang and Zhang, 2014) is $s_{[k]} \log p = o(\sqrt{n})$, which leads to h being of order $o(1/\sqrt{\log p})$. Meanwhile, the larger $s_{[k]}$ is, the weaker the conditions on h. The condition $h(\log p)^{1/2} = o(1)$ coincides with conditions of Theorem 4.1 in Li et al. (2024), which was required for inference on each component of the debiased transfer learning coefficient estimator in the generalized linear model.*

Remark 20 *Our transfer learning approach requires the source and target data to be similar with respect to the l_1 norm of the difference in regression coefficients. However, it is possible that the external trial data may differ substantially from the current trial data, even if the two trials have similar characteristics. In this case, it is still feasible to use our transfer learning estimator; however, the asymptotic variance may differ from that in the case with sufficient similarity. Proposition 26 in Appendix B provides a way to obtain valid inference for the transfer learning estimator. Nevertheless, this change in variance could reduce efficiency relative to the ideal setting with high similarity between the two trials. The simulation results in Appendix D.3 demonstrate that even when large l_1 differences indicate a risk of negative transfer, our transfer learning estimator can still provide valid inference.*

5. Consistent variance estimator

A crucial step in drawing valid inference is constructing a consistent asymptotic variance estimator. Specifically, we consider the difference in the potential outcomes for any two treatments b and c. Let e_{bc} be an A-dimensional vector, such that all elements are zero, except for the bth element, which is 1, and the cth element, which is -1. Let $\tau_{bc} = E\{Y(b) - Y(c)\} = \tau_b - \tau_c,\ b, c \in \mathcal{A}$, be the real difference in potential outcomes, and $\hat{\tau}_{\dagger,bc}$ be the estimator for τ_{bc} generated from the estimators above, for $\dagger = \text{lasso}, \text{tl}$. Based on simple calculations,

$$\sqrt{n}(\hat{\tau}_{\dagger,bc} - \tau_{bc}) \xrightarrow{d} \mathrm{N}(0, \sigma^2_{\text{proj,bc}}), \quad \dagger = \text{lasso}, \text{tl},$$

where

$$\sigma^2_{\text{proj,bc}} = e_{bc}^{\mathrm{T}} V_{\text{proj}} e_{bc} = V_{\text{proj}}(b,b) - 2V_{\text{proj}}(b,c) + V_{\text{proj}}(c,c).$$

Here $V_{\text{proj}}(i,j)$ denotes the element in the ith row and jth column of V_{proj}. By using $r_i(a)$ as the transformed outcomes again and $\hat{r}_i(a)$ as an estimator of $r_i(a)$, $\hat{r}_i = \sum_{a\in\mathcal{A}_0} I(A_i = a)\hat{r}_i(a)$. Let $\bar{r}_{[k]a} = 1/n_{[k]a}\sum_{i=1}^n I(A_i = a, B_i = k)\hat{r}_i$. For the estimators above, we use $\hat{r}_{i,\dagger} = Y_i - X_i^{\mathrm{T}}\hat{\beta}_{[k],\dagger}(a)$, where $\dagger = \text{lasso}, \text{tl}$, respectively, for $i \in [k],\ A_i = a$. We define

$$\hat{\sigma}^2_{r_\dagger(a)} = \sum_{k\in\mathcal{K}} \frac{p_{n[k]} n_{[k]}}{n_{[k]a}} \left\{ \frac{1}{n_{[k]a}} \sum_{i=1}^n I(A_i = a, B_i = k)(\hat{r}_{i,\dagger} - \bar{r}_{[k]a,\dagger})^2 \right\}, \quad \dagger = \text{lasso}, \text{tl},$$

and

$$\hat{\sigma}^2_{bc,HY} = \sum_{k\in\mathcal{K}} p_{n[k]} \left\{ \left(\bar{Y}_{[k]b} - \sum_{k'\in\mathcal{K}} p_{n[k']} \bar{Y}_{[k']b} \right) - \left(\bar{Y}_{[k]c} - \sum_{k'\in\mathcal{K}} p_{n[k']} \bar{Y}_{[k']c} \right) \right\}^2.$$

Let $S_{[k]XX} = 1/n_{[k]}\ \sum_{i=1}^n I(B_i = k)(X_i - \bar{X}_{[k]})(X_i - \bar{X}_{[k]})^{\mathrm{T}}$ denote the stratum-specific sample covariance matrix. Let $\hat{\sigma}^2_{\dagger,bc}$ be the estimator of $\sigma^2_{\text{proj,bc}}$ for $\dagger = \text{lasso}, \text{tl}$, respectively.

Theorem 21

1. *Under Assumptions 1–5, the variance estimator*

$$\begin{aligned}\hat{\sigma}^2_{\text{lasso,bc}} = {} & \hat{\sigma}^2_{r_{\text{lasso}}(b)} + \hat{\sigma}^2_{r_{\text{lasso}}(c)} + \hat{\sigma}^2_{bc,HY} \\ & + \sum_{k\in\mathcal{K}} p_{n[k]} \left\{ \hat{\beta}_{[k]\text{lasso}}(b) - \hat{\beta}_{[k]\text{lasso}}(c) \right\}^{\mathrm{T}} S_{[k]XX} \left\{ \hat{\beta}_{[k]\text{lasso}}(b) - \hat{\beta}_{[k]\text{lasso}}(c) \right\}\end{aligned}$$

   is a consistent estimator for $\sigma^2_{\text{proj,bc}}$.

2. *Under conditions in Theorem 16, the variance estimator*

$$\begin{aligned}\hat{\sigma}^2_{\text{tl,bc}} = {} & \hat{\sigma}^2_{r_{\text{tl}}(b)} + \hat{\sigma}^2_{r_{\text{tl}}(c)} + \hat{\sigma}^2_{bc,HY} \\ & + \sum_{k\in\mathcal{K}} p_{n[k]} \left\{ \hat{\beta}_{[k]\text{tl}}(b) - \hat{\beta}_{[k]\text{tl}}(c) \right\}^{\mathrm{T}} S_{[k]XX} \left\{ \hat{\beta}_{[k]\text{tl}}(b) - \hat{\beta}_{[k]\text{tl}}(c) \right\}\end{aligned}$$

   is a consistent estimator for $\sigma^2_{\text{proj,bc}}$.

Remark 22 *A concern regarding Theorem 16 is that the condition* $h < s_{[k]}\sqrt{\log p/n}$ *may not hold, as discussed in Remark 20. However, in this case, we can still obtain valid inference for* $\hat{\tau}_{\text{tl}}$ *using a debiased variance estimator. Assume there exists some coefficient vector* $\beta_{[k]}(a)$ *such that*

$$\sqrt{n}(\bar{X}_{[k]1} - \bar{X}_{[k]0})^{\mathrm{T}}\{\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]}(a)\} = o_p(1).$$

Then under Assumptions 1 and 2, we can show that $\sqrt{n}(\hat{\tau}_{\text{tl}} - \tau)$ *is asymptotically normal with a covariance matrix depending on* $\beta_{[k]}(a)$. *At the same time, we can provide a consistent variance estimator for this new asymptotic variance as we did in this section. The related theorems and proofs are provided in the Appendix B. A direct result is that when Assumptions 10-14 hold,* $\hat{\beta}_{[k]\text{tl}}(a)$ *can be replaced by* $\hat{\beta}^{(s)}_{[k]\text{proj}}(a)$ *to obtain consistent treatment effect estimator and variance estimator. However, the efficiency of this source-only estimator could be lower. In the next section, we present simulations to compare the performance of these estimators.*

6. Simulation

In this section, we evaluate the performance of our proposed estimators compared to a benchmark estimator across various scenarios. The benchmark estimator is the unadjusted estimator in Bugni et al. (2019), as discussed in Remark 8. We compare the lasso-adjusted treatment effect estimator $\hat{\tau}_{\text{lasso}}$, the transfer learning treatment effect estimator $\hat{\tau}_{\text{tl}}$, and the source-only treatment effect estimator $\hat{\tau}_{\text{so}}$ as we mentioned in Remark 22. The term "source-only estimator" refers to an estimator whose projection coefficients are estimated solely from the source data. The source-only treatment effect estimator for each treatment a is defined as

$$\hat{\tau}_{\text{so,a}} = \sum_{k\in\mathcal{K}} p_{n[k]} \big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}} \hat{\beta}^{(s)}_{[k]\text{proj}}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}} \hat{\beta}^{(s)}_{[k]\text{proj}}(0)\}\big]$$

and $\hat{\tau}_{\text{so}}$ is

$$\hat{\tau}_{\text{so}} = (\hat{\tau}_{\text{so},1}, \hat{\tau}_{\text{so},2}, \ldots, \hat{\tau}_{\text{so,A}})^{\mathrm{T}}.$$

Note that, $\hat{\tau}_{\text{so}}$ can be seen as we transfer the coefficient estimator in source data without any bias correction.

The following model is used to generate the potential outcomes for both source and target data:

$$Y_i(a) = \mu_a + g_a(X_i) + \epsilon_{a,i}, \text{ for } a \in \mathcal{A}_0,$$

where X_i, $g_a(X_i)$ are specified below. We consider $|\mathcal{A}_0| = 3$, that is, 2 treatment groups and a control group. In each model, $(X_i, \epsilon_{0,i}, \epsilon_{1,i}, \epsilon_{2,i})$ are i.i.d., and $\epsilon_{a,i}$ follows standard normal distribution, for $a = 0, 1, 2$. Covariates other than X_i are generated such that the total dimension is $p = 100$ to consider the high-dimensional setting. The details of the models are as follows:

Model 1: We consider s-dimensional X_i such that $g_a(X_i) = \beta_1 X_{i1} + \sum_{j=2}^{s} \beta_j X_{i1} X_{ij}$, for $a = 0, 1, 2$, where X_{i1} takes value in $\{1, 2\}$ with probabilities 0.4 and 0.6, respectively. $X_{ij} \sim \text{Unif}[-2, 2]$, for $j = 2, 3, \ldots, s$, and they are independent of each other. The coefficient β will be specified later for source data and target data separately. Here X_{i1} is used for randomization, resulting in two strata. The additional covariates are independent of X_i and follow a multivariate normal distribution with mean zero and a diagonal covariance matrix where all diagonal elements are 2.

Model 2: Similar to Model 1, we consider s-dimensional X_i, where X_{i1} takes value in $\{1, 2\}$ with probabilities 0.4 and 0.6, respectively. $X_{ij} \sim \text{Beta}(2, 2)$ for $j = 2, 3, \ldots, s$, and they are independent of each other. For different treatments, we set $g_0(X_i) = \beta_1 X_{i1} + \sum_{j=2}^{s} \beta_j X_{i1}(X_{ij} - 0.5)$, $g_1(X_i) = \beta_1 X_{i1} + \sum_{j=2}^{s} \beta_j X_{i1}(X_{ij}^2 - 0.3)$, and $g_2(X_i) = g_0(X_i)$. The coefficients β and the additional covariates are set the same as in Model 1.

The source and target data are generated from either Model 1 (a linear model) or Model 2 (a nonlinear model) for the potential outcomes. Since our transfer learning estimator is derived under a high-dimensional linear working model, we evaluate its performance under both correctly specified and misspecified scenarios, where "misspecified" means that the true data-generating model differs from the linear model. To this end, we consider three cases. Case 1 considers a fully linear setting for both the source and target data, while Case 2 and Case 3 introduce model misspecification in the source and target data, respectively. The details for each case are as follows:

Case 1: Both the source and target data use Model 1. We examine the results as the bias increases in a linear model setting. Let $s = \lceil n^r \rceil$ where $r = 0, 0.05, \ldots, 0.7$. In target data, $\beta_j = 2$ for $j = 1, 2, \ldots, s$, and in source data, $\beta_j^{(s)} = 2 + \frac{hj}{s}$ for $j = 1, 2, \ldots, s$, $h = 0.2, 0.5, 1$. This setup leads to $\|\beta - \beta^{(s)}\|_1 = h(s+1)/2$.

Case 2: We use Model 2 as the source data and Model 1 as the target data. This case demonstrates the compatibility of our transfer learning estimator when the source data is misspecified. The coefficients β, $\beta^{(s)}$, and dimension s are the same as in Case 1.

Case 3: We use Model 1 as the source data and Model 2 as the target data. This case illustrates the robustness of our proposed estimators when the target data is misspecified. The coefficients β, $\beta^{(s)}$, and dimension s are the same as in Case 1.

We represent the simulation results for four treatment effect estimators under stratified block randomization. The target data size is $n = 300$ and the source data size is $n^{(s)} = 1200$. The allocation ratio is 1:1:1 among treatments, with a block size of 6. Our estimators are calculated in $R = 2000$ replicates. The performances are measured by relative bias, standard deviation, and coverage probability of 95% confidence interval. The relative bias is defined as $\sqrt{n}R^{-1}\sum_{r=1}^{R}(\hat{\tau}_{\dagger,r} - \tau)/\hat{\sigma}$, where $\hat{\tau}_{\dagger,r}$ denotes any of our treatment effect estimators in the rth replicate, τ is the true value of treatment effect, and $\hat{\sigma}$ is the sampling standard deviation. An additional result for the unequal allocation ratio, which is 1:1:1 in the first stratum and 2:2:1 in another stratum, is in Appendix D.1. The results are summarized as follows:

1. For all cases, the relative bias for all estimators is close to 0.
2. In Figure 1(B), the sampling standard deviation (SD) of all estimators increases as s increases. When the difference between the true projection coefficients in the source and target data is small, $\hat{\tau}_{\text{so}}$ behaves similarly to $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{tl}}$, as expected. As h increases, the gap in SD between $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{lasso}}$ widens, reflecting that the source data becomes less informative for the target domain. This gap can be reduced by our transfer learning approach. As we expected, when s is small, the sampling standard deviation for $\hat{\tau}_{\text{tl}}$ is similar to $\hat{\tau}_{\text{lasso}}$, which verifies Theorem 16; when s is sufficiently large, $\hat{\tau}_{\text{tl}}$ shows an efficiency gain compared to $\hat{\tau}_{\text{lasso}}$.
3. In Figure 1(C), $\hat{\tau}_{\text{lasso}}$ shows a substantial decrease in coverage probability when s increases, which may be caused by the overfitting when the condition $(s \log p < \sqrt{n})$ is violated, leading to an inconsistent variance estimator. As we have mentioned in Remark 22, the coverage probability is approximately 95% for $\hat{\tau}_{\text{so}}$. For $\hat{\tau}_{\text{tl}}$, the estimation is conservative when $h = 0.2$. When $h = 0.5$ and s increases, $\hat{\tau}_{\text{tl}}$ maintains a 95% coverage probability as the conditions in Theorem 16 hold. Moreover, even when the condition is violated at $h = 1$, $\hat{\tau}_{\text{tl}}$ shows a slower decrease in coverage probability compared to $\hat{\tau}_{\text{lasso}}$.
4. In Figures 2 and 3, the underlying model is misspecified from the linear model for source data or target data, respectively. The results for all measures are consistent with those presented in Figure 1, which shows that our transfer learning algorithm is robust regardless of whether the outcome model for the target data or the source data is misspecified. Especially, our transfer learning approach performs better in Cases 2 and 3, with the coverage probabilities around 95%. This confirms that our proposed approach can provide valid inference when conditions are satisfied.

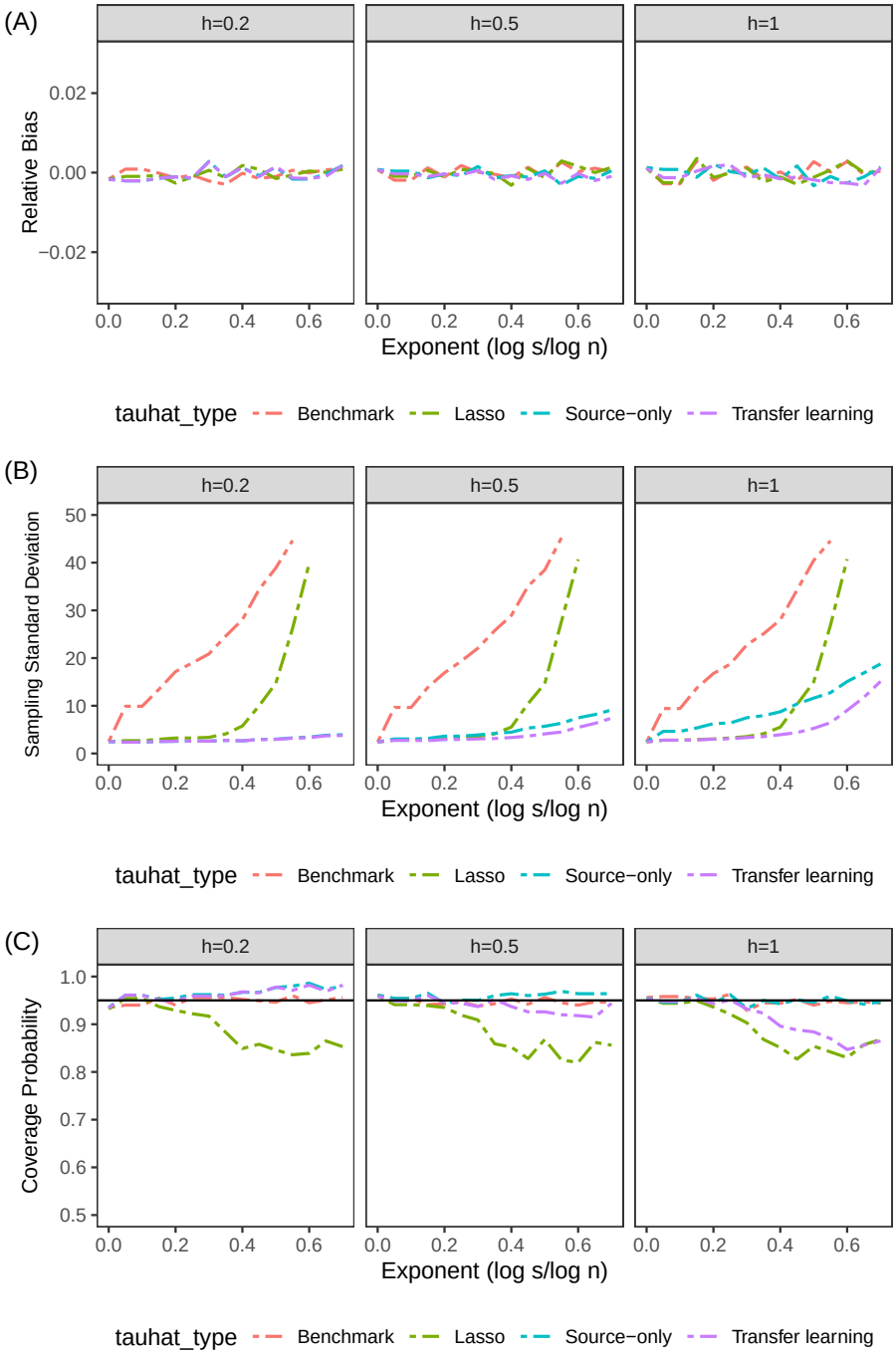

Figure 1: Simulations in Case 1 under stratified block randomization. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

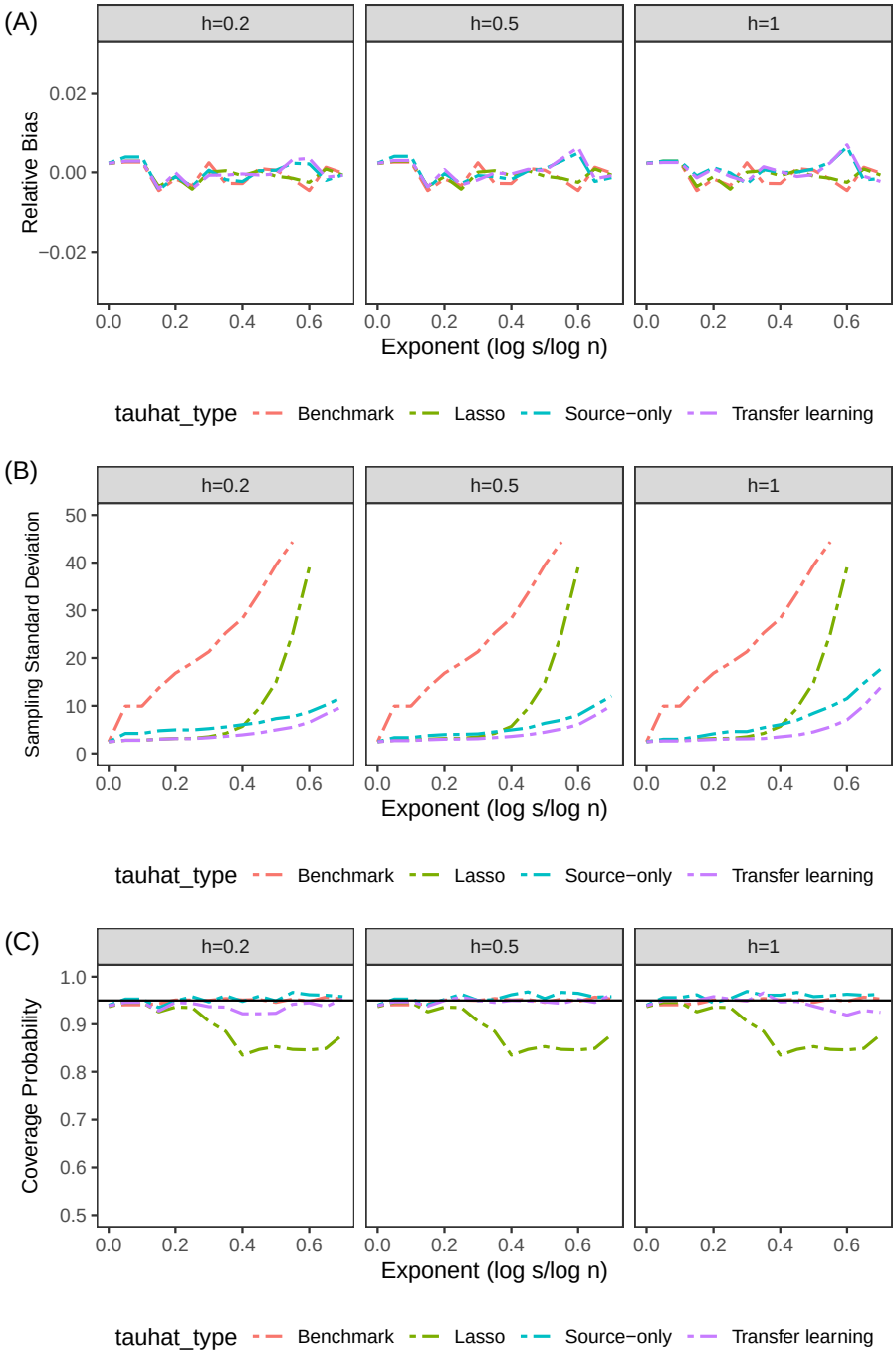

Figure 2: Simulations in Case 2 under stratified block randomization. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

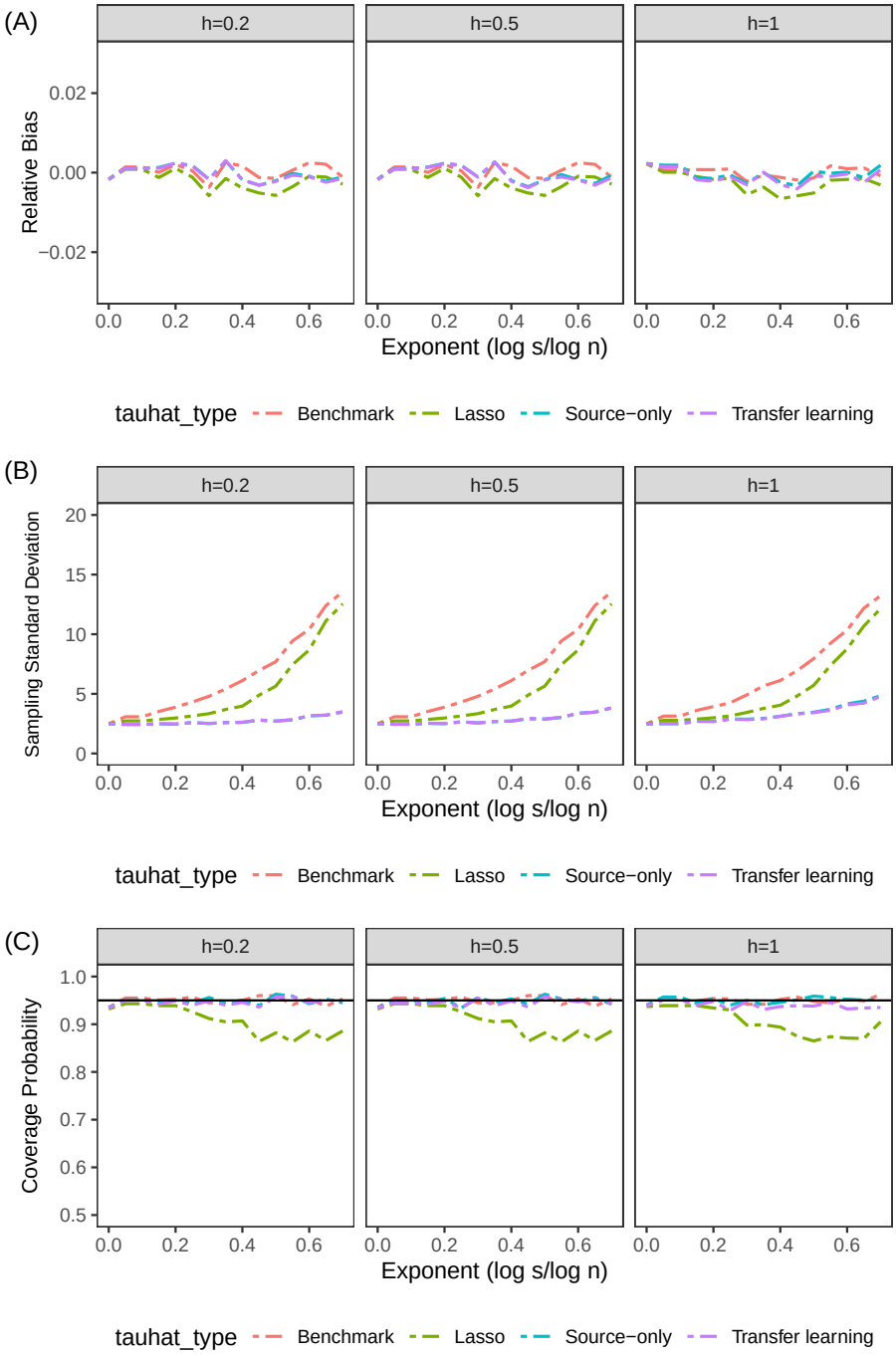

Figure 3: Simulations in Case 3 under stratified block randomization. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

7. Clinical trial example

We consider an example of a clinical trial solely to generate synthetic data to illustrate the capability of our proposed estimators to improve efficiency compared with the benchmark and provide valid inference. Additionally, the application of our proposed method to the original dataset is detailed in Appendix E. We use synthetic data from a trial using nefazodone and the cognitive behavioral-analysis system of psychotherapy (CBASP) (Keller et al., 2000). The trial aimed to compare the efficacy of nefazodone, CBASP, and their combination in the treatment of chronic depression. We consider the combination as the control group indexed as 0, nefazodone as the treatment indexed as 1, and the CBASP as the treatment indexed as 2. The outcome of interest was FinalHAMD, the final score of the 24-item Hamilton rating scale for depression. We evaluate two treatment effects: the effect of nefazodone versus the combination, denoted as τ_1, and the effect of CBASP versus the combination, denoted as τ_2. To generate the synthetic data, we first fit the data with a nonparametric spline model using the function `bigssa` from the `R` package `bigspline`. Based on this model, we generate multiple synthetic datasets to evaluate bias, variance, and coverage probabilities. We stratify the data using GENDER and select eight covariates to fit the model: AGE, HAMA SOMATI, HAMD17, HAMD24, HAMD COGIN, Mstatus2, NDE and TreatPD, which are detailed in Table 1. We take linear, quadratic, cubic, and interaction terms for continuous covariates, linear and interaction terms for binary covariates, and interaction terms for the above two sets of coordinates as the covariates, resulting in $p = 135$ for treatment effect estimators other than the benchmark estimator.

Table 1: Description of baseline covariates

Variable	Description
AGE	Age of patients in years
GENDER	1 female and 0 male
HAMA_SOMATI	HAMA somatic anxiety score
HAMD17	Total HAMD-17 score
HAMD24	Total HAMD-24 score
HAMD_COGIND	HAMD cognitive disturbance score
Mstatus2	Marriage status: 1 if married or living with someone and 0 otherwise
NDE	Number of depressive episodes
TreatPD	Treated past depression: 1 yes and 0 no

Note: HAMD, Hamilton Depression Scale; HAMA, Hamilton Anxiety Scale.

To generate the source and target data, we split the data based on the HAMD17 score such that the target population has HAMD17 < 18 and the source population has HAMD17 ≥ 18. The purpose is to create a difference in outcomes between the source and target data, as shown by a boxplot in Figure 4. Then, we sample 300 units among the target population with replacement as the target dataset. Similarly, we sample 1200 units from the source population with replacement as the source data. Next, we implement stratified block randomization on both the source and target data, using GENDER as the stratification variable with an allocation ratio of 1:1:1 for treatments and a block size of 6. In this

section, we also compare the performance of the same four estimators as in Section 6. The treatment effects and variances are evaluated based on 1000 replicates. The coverage probabilities of the 95% confidence interval are assessed based on the estimates and variance estimators. The results are shown in Table 2.

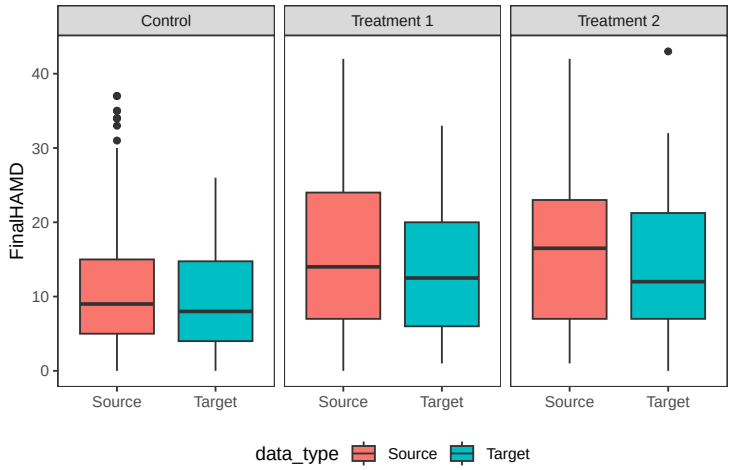

Figure 4: Difference in potential outcomes (FinalHAMD) between source data and target data.

Table 2: Treatment effect estimates, sampling variances, variance reductions and coverage probabilities under stratified block randomization for synthetic nefazodone CBASP trial data.

	τ_1				τ_2			
Estimator	Estimate	Variance ($\times 10^{-2}$)	VR (%)	CP	Estimate	Variance ($\times 10^{-2}$)	VR (%)	CP
$\hat{\tau}_{\text{ben}}$	4.46	15.97	—	0.956	4.29	22.70	—	0.954
$\hat{\tau}_{\text{so}}$	4.44	21.57	-35.08	0.950	4.30	19.13	15.75	0.954
$\hat{\tau}_{\text{lasso}}$	4.46	11.89	25.53	0.933	4.30	15.38	32.25	0.913
$\hat{\tau}_{\text{tl}}$	4.46	11.28	29.36	0.944	4.29	13.58	40.19	0.950

Note: τ_1, the treatment effect of nefazodone versus combination therapy; τ_2, the treatment effect of CBASP versus combination therapy; Estimates: treatment effect estimates; CI, confidence interval; VR, variance reduction; CP, coverage probability.

As shown in Table 2, the treatment effect estimation for all four estimators is similar, with positive results. The source-only estimator $\hat{\tau}_{\text{so}}$ has the worst performance in terms of efficiency, which may decrease the efficiency compared with the benchmark estimator $\hat{\tau}_{\text{ben}}$ (variance reduction: -35.08%). Both $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{tl}}$ show efficiency gains compared with $\hat{\tau}_{\text{ben}}$ as the variance reductions range from 25.53% to 40.19%, and $\hat{\tau}_{\text{tl}}$ also shows an efficiency gain compared with $\hat{\tau}_{\text{lasso}}$. Moreover, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{so}}$ have better performance on coverage probabilities than $\hat{\tau}_{\text{lasso}}$, which are closer to 95%.

8. Discussion

In this paper, we incorporate external trial data into the inference under covariate-adaptive randomization by transfer learning. We first generalize the lasso-adjusted stratum-specific treatment effect estimator under multiple treatments. Then, we apply a transfer learning approach to develop the transfer learning estimator. The theoretical results show that our proposed estimator can enhance the lasso-adjusted stratum-specific estimator to have a faster l_1 convergence rate of regression coefficient estimators and weaken the requirements on the sample size for asymptotic normality when the source data and target data are similar under mild conditions. We also propose a consistent variance estimator to provide valid inference. The simulation results and a clinical trial example confirm our findings. By proposing an effective transfer algorithm, offering theoretical guarantees, and developing a robust inference method, our study sheds light on the potential of transfer learning in analyzing randomized clinical trials.

Our method relies on the assumption that the l_1-norm of the coefficient difference is small. When there are multiple heterogeneous source datasets, some of them may not satisfy our conditions on the similarity. In this case, one possible solution is to use a data-driven source detection mechanism to exclude datasets with poor similarity. However, this could be complex, as it may introduce correlated selection procedures across datasets. Therefore, identifying similar parts is a potential direction for future work. Meanwhile, the efficiency gain of the treatment effect estimation can be improved by estimating the underlying model between the response and the baseline covariates (Tu et al., 2024). A potential improvement for our work is to replace the linear model with a general function for covariates. In this setting, future work may focus on characterizing the difference between the source data and the target data and proposing a proper transfer learning approach to develop a robust treatment effect estimator and ensure valid inference. Additionally, a promising direction for future research is extending our transfer learning framework to accommodate more complex adaptive randomization procedures. For example, response-adaptive randomization (RAR) in clinical trials introduces outcome-dependent allocations that violate Assumption 2. Despite this violation, as shown in Appendix D.5, our framework can be empirically applied to RAR and continues to demonstrate noticeable advantages. Establishing rigorous theoretical guarantees under RAR remains an important avenue for our future work. Beyond clinical trials, generalizing our methodology to randomization schemes frequently used in other intervention studies, such as rerandomization (Zhu et al., 2025) and matched-pair designs (Bai et al., 2024), represents another valuable opportunity for future investigation. Overall, these potential works will help broaden the scope and enhance the utility of transfer learning in the context of randomized trials.

Acknowledgments

We thank the Action Editor and the referees for their helpful comments, which have significantly improved our manuscript. Wei Ma was supported by the National Natural Science Foundation of China (Grant No. 12571277). Hanzhong Liu was supported by the National Natural Science Foundation of China (Grant No. 12531012) and the Beijing Natural Science Foundation (Grant No. F251001). Wei Ma is the corresponding author for this paper.

Appendix A. A notation table

We summarize the frequently used notations in Table 3.

Table 3: Notations used in this paper

Notations	Definitions
n	Number of units in the trial
$n_{[k]}$	Number of units in stratum k
n_a	Number of units in the treatment group a
$n_{[k]a}$	Number of units in the stratum k and treatment group a
$p_{n[k]} = n_{[k]}/n$	Estimated proportion in each stratum k
$p_{[k]} = \Pr(B_i = k)$	Expected proportion in each stratum k
$\pi_{[k]a} = \Pr(A_i = a \mid B_i = k)$	Expected proportion in stratum k for treatment group a
$\mathcal{A}$	A set of treatment groups, $\{1, 2, \ldots, A\}$
$\mathcal{A}_0$	A set of treatment groups including the control group, $\{0, 1, 2, \ldots, A\}$
$r_i(a)$	A transformed outcome
$\tilde{r}_i(a) = r_i(a) - E\{r_i(a)\|B_i\}$	A stratum-centered transformed outcome
$r_i = \sum_{a=0}^{A} I(A_i = a) r_i(a)$	Observed transformed outcome
$\bar{r}_a$	The sample mean of r_i in treatment group a
$\bar{r}_{[k]}$	The sample mean of r_i in stratum k
$\bar{r}_{[k]a}$	The sample mean of r_i in stratum k and treatment group a
τ_a	The treatment effect for treatment a
τ	The treatment effect for all treatments, $(\tau_1, \ldots, \tau_A)^{\mathrm{T}}$
$\hat{\tau}_{\mathrm{ben}}$	The benchmark estimator for τ, defined in Remark 8
$\hat{\tau}_{\mathrm{lasso}}$	The lasso-adjusted estimator for τ, defined in Section 3
$\hat{\tau}_{\mathrm{tl}}$	The transfer learning estimator for τ, defined in Algorithm 1
$\beta_{[k]\mathrm{proj}}(a)$	The population-level coefficient in stratum k and treatment group a
$\hat{\beta}_{[k]\mathrm{lasso}}(a)$	The lasso vector to estimate $\beta_{[k]\mathrm{proj}}(a)$ in stratum k and treatment group a
$\hat{\beta}_{[k]\mathrm{tl}}(a)$	The transfer learning vector to estimate $\beta_{[k]\mathrm{proj}}(a)$ in stratum k and treatment group a
M	The uniform bound for each element of X_i
$M_{n[k]}$	A sequence with n in stratum k, satisfying $M_{n[k]} \to \infty$ while $M_{n[k]}/n \to 0$ when $n \to \infty$.
$S_{[k]}$	The union of the support of $\beta_{[k]\mathrm{proj}}(a)$ for all $a \in \mathcal{A}_0$
$s_{[k]} = \|S_{[k]}\|$	The number of relevant covariates in stratum k
$(\cdot)^{(s)}$	Expression for notations in source data. For example, $Y_i^{(s)}(a)$, $X_i^{(s)}$ and $\beta^{(s)}_{[k]\mathrm{proj}}(a)$.

Appendix B. General theory for regression adjustment under covariate-adaptive randomization with multiple treatments

We develop a general theory for regression-adjusted treatment effect estimators, which we can apply to the lasso-adjusted and transfer learning estimators.

Assumption 23 *For $k = 1, 2, \ldots, K$ and $a \in \mathcal{A}_0$, there exist coefficient vectors $\beta_{[k]}(a)$ such that*

$$\sqrt{n}(\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\{\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)\} = o_p(1).$$

Remark 24 *We develop the asymptotic normality by first deriving the asymptotic normality for the mean potential outcomes $\tau^* = (\tau_0^*, \ldots, \tau_A^*)^{\mathrm{T}}$ and then using linear transformation to obtain the result, where $\tau_a^* = E\{Y_i(a)\}$ for $a \in \mathcal{A}_0$. Therefore, Assumption 23 is the remaining term for the first step and needs to be $o_p(1)$.*

Remark 25 *When the dimension p is fixed, $\sqrt{n}(\bar{X}_{[k]a} - \bar{X}_{[k]}) = O_p(1)$ by Bugni et al. (2018). Therefore, when $\hat{\beta}_{[k]}(a)$ element-wisely converges to $\beta_{[k]}(a)$, Assumption 23 holds. Remark 2 in Gu et al. (2023) also explained this result when using the stratum-common estimator. However, the OLS estimator may not satisfy this condition when p is diverging or even larger than the sample size.*

Let $\hat{\tau}_{\mathrm{gen}}$ be a general estimator for treatment effect τ. That is,

$$\hat{\tau}_{\mathrm{gen,a}} = \sum_{k\in\mathcal{K}} p_{n[k]}\left[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]}(0)\}\right]$$

and

$$\hat{\tau}_{\mathrm{gen}} = (\hat{\tau}_{\mathrm{gen},1}, \ldots, \hat{\tau}_{\mathrm{gen,A}}).$$

Also, we define the transformed outcome $r_{i,\mathrm{gen}}(a) = Y_i(a) - X_i^{\mathrm{T}}\beta_{[k]}(a)$, for $i \in [k], A_i = a$. Recall that 1_A is an A-dimensional vector with ones and $m_{[k]}(a) = E[Y_i(a)|B_i = k] - E[Y_i(a)]$. Define

$$\begin{aligned}
V_{H\tilde{Y}} &= \sum_{k\in\mathcal{K}} p_{[k]}\left(m_{[k]}(a) - m_{[k]}(0) : a \in \mathcal{A}\right) \times \left(m_{[k]}(a') - m_{[k]}(0) : a' \in \mathcal{A}\right)^{\mathrm{T}}, \\
V_{\tilde{r}_{\mathrm{gen}}} &= \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]0}}\sigma^2_{[k]\tilde{r}_{\mathrm{gen}}(0)} 1_A 1_A^{\mathrm{T}} + \mathrm{diag}\left\{\sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}}\sigma^2_{[k]\tilde{r}_{\mathrm{gen}}(a)} : a \in \mathcal{A}\right\}, \\
V_{\tilde{X}} &= \Big\{\sum_{k\in\mathcal{K}} p_{[k]}\{\beta_{[k]}(a) - \beta_{[k]}(0)\}^{\mathrm{T}}\Sigma_{[k]XX}\{\beta_{[k]}(a') - \beta_{[k]}(0)\} : (a, a') \in \mathcal{A}\times\mathcal{A}\Big\}, \\
V_{\tilde{r}_{\mathrm{gen}}\tilde{X}} &= \Big\{\sum_{k\in\mathcal{K}} p_{[k]}\{\beta_{[k]}(a') - \beta_{[k]}(0)\}^{\mathrm{T}}\left[(\Sigma_{[k]XY(a)} - \Sigma_{[k]XY(0)}) - \Sigma_{[k]XX}\{\beta_{[k]}(a) - \beta_{[k]}(0)\}\right] : \\
&\qquad (a, a') \in \mathcal{A}\times\mathcal{A}\Big\}.
\end{aligned}$$

Proposition 26 *Under Assumptions 1, 2 and 23, if $V_{\tilde{r}_{\mathrm{gen}}}$, $V_{\tilde{X}}$, and $V^{\mathrm{T}}_{\tilde{r}_{\mathrm{gen}}\tilde{X}}$ have finite limits, and the limit of $V_{H\tilde{Y}} + V_{\tilde{r}_{\mathrm{gen}}} + V_{\tilde{X}} + V^{\mathrm{T}}_{\tilde{r}_{\mathrm{gen}}\tilde{X}}$, denoted by V_{gen}, is positive, $r_{i,\mathrm{gen}}(a) \in \mathcal{R}_2$, and $E\{r_{i,\mathrm{gen}}(a)\}^2 < \infty$ for $a \in \mathcal{A}_0$, we have*

$$\sqrt{n}(\hat{\tau}_{\mathrm{gen}} - \tau) \xrightarrow{d} \mathrm{N}(0, V_{\mathrm{gen}}).$$

Proof The asymptotic normality can be derived from Theorem 2 in Gu et al. (2023) under Assumption 23 with minor changes. Recall that

$$\hat{\tau}_{\text{gen,a}} = \sum_{k\in\mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]}(0)\}\big].$$

Then we have

$$\begin{aligned}
\hat{\tau}_{\text{gen}} &= \Big(\sum_{k\in\mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]}(a)\} - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\hat{\beta}_{[k]}(0)\}\big] : a\in\mathcal{A}\Big) \\
&= \Big(\sum_{k\in\mathcal{K}} p_{n[k]}\big[\{\bar{Y}_{[k]a} - (\bar{X}_{[k]a} - E(X_i|B_i=k))^{\mathrm{T}}\beta_{[k]}(a)\} \\
&\quad - \{\bar{Y}_{[k]0} - (\bar{X}_{[k]0} - E(X_i|B_i=k))^{\mathrm{T}}\beta_{[k]}(0)\}\big] : a\in\mathcal{A}\Big) \\
&\quad - \Big(\sum_{k\in\mathcal{K}} p_{n[k]}\{E(X_i|B_i=k) - \bar{X}_{[k]}\}^{\mathrm{T}}\{\beta_{[k]}(a) - \beta_{[k]}(0)\} : a\in\mathcal{A}\Big) \\
&\quad + \Big(\sum_{k\in\mathcal{K}} p_{n[k]}[(\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\{\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)\} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\{\hat{\beta}_{[k]}(0) - \beta_{[k]}(0)\}] : a\in\mathcal{A}\Big).
\end{aligned}$$

The proof reduces to showing that the first two terms are asymptotically normal and the last term is $o_p(n^{-1/2})$. By Assumption 23, we have

$$(\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\{\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)\} = o_p(n^{-1/2})$$

for all $a\in\mathcal{A}_0$. As the stratum size K is finite, we have

$$\sum_{k\in\mathcal{K}} p_{n[k]}[(\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}\{\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)\} - (\bar{X}_{[k]0} - \bar{X}_{[k]})^{\mathrm{T}}\{\hat{\beta}_{[k]}(0) - \beta_{[k]}(0)\}] = o_p(n^{-1/2})$$

for all $a\in\mathcal{A}$.

Then we show that the first two terms are asymptotically normal. Simple calculation shows that

$$E(X_i|B_i=k) = \frac{1}{n_{[k]}}\sum_{i=1}^{n} I(B_i=k)E(X_i|B_i=k) = \frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i=a,B_i=k)E(X_i|B_i=k),$$

$$\bar{X}_{[k]} = \frac{1}{n_{[k]}}\sum_{i=1}^{n} I(B_i=k)X_i.$$

Define

$$\begin{aligned}
\mathbb{L}^{(1)}_{n[k],r_i(a)} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(A_i=a,B_i=k)\tilde{r}_i(a), \\
\mathbb{L}^{(1)}_{n[k],X_i} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(B_i=k)\tilde{X}_i, \\
\mathbb{L}^{(2)}_{n[k]} &= \sqrt{n}(p_{[k]} - p_{n[k]}).
\end{aligned}$$

Here $r_i(a)$ is a transformed outcome, and

$$\tilde{r}_i(a) = r_i(a) - E[r_i(a)|B_i = k], \quad i \in [k].$$

In Bugni et al. (2019), it takes $r_i(a) = Y_i(a)$. Some calculation for the first two parts of $\hat{\tau}_{\text{gen}}$ leads to

$$\begin{aligned}\sqrt{n}(\hat{\tau}_{\text{gen}} - \tau) =& \Big\{ \sum_{k\in\mathcal{K}} \big(\frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r_{i,\text{gen}}(a)} - \frac{1}{\pi_{[k]0}} \mathbb{L}^{(1)}_{n[k],r_{i,\text{gen}}(0)}\big), a \in \mathcal{A} \Big\} \\ &+ \Big\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(1)\text{T}}_{n[k],X_i} \{\beta_{[k]}(a) - \beta_{[k]}(0)\}, a \in \mathcal{A} \Big\} \\ &+ \Big\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(2)}_{n[k]} \{m_{[k]}(a) - m_{[k]}(0)\}, a \in \mathcal{A} \Big\} + o_p(1),\end{aligned}$$

where $r_{i,\text{gen}}(a) = Y_i(a) - X_i^{\text{T}}\beta_{[k]}(a)$ and $m_{[k]}(a) = E[Y_i(a) \mid B_i = k] - E[Y_i(a)]$. By Lemma 27 and some additional calculations, we have

$$\begin{pmatrix} \Big\{ \sum_{k\in\mathcal{K}} \big(\frac{1}{\pi_{[k]a}} \mathbb{L}^{(1)}_{n[k],r_{i,\text{gen}}(a)} - \frac{1}{\pi_{[k]0}} \mathbb{L}^{(1)}_{n[k],r_{i,\text{gen}}(0)}\big), a \in \mathcal{A} \Big\} \\ \Big\{ \sum_{k\in\mathcal{K}} (\mathbb{L}^{(1)\text{T}}_{n[k],X_i} \{\beta_{[k]}(a) - \beta_{[k]}(0)\}, a \in \mathcal{A} \Big\} \\ \Big\{ \sum_{k\in\mathcal{K}} \mathbb{L}^{(2)}_{n[k]} \{m_{[k]}(a) - m_{[k]}(0)\}, a \in \mathcal{A} \Big\} \end{pmatrix} \xrightarrow{d} \text{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{\tilde{r}_{\text{gen}}} & V_{\tilde{r}_{\text{gen}}\tilde{X}} & 0 \\ V^{\text{T}}_{\tilde{r}_{\text{gen}}\tilde{X}} & V_{\tilde{X}} & 0 \\ 0 & 0 & V_{H\tilde{Y}} \end{pmatrix} \right),$$

where

$$\begin{aligned} V_{H\tilde{Y}} &= \sum_{k\in\mathcal{K}} p_{[k]} \left(m_{[k]}(a) - m_{[k]}(0) : a \in \mathcal{A}\right) \times \left(m_{[k]}(a') - m_{[k]}(0) : a' \in \mathcal{A}\right)^{\text{T}}, \\ V_{\tilde{r}_{\text{gen}}} &= \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]0}} \sigma^2_{[k]\tilde{r}_{\text{gen}}(0)} 1_A 1_A^{\text{T}} + \text{diag}\left\{ \sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}} \sigma^2_{[k]\tilde{r}_{\text{gen}}(a)} : a \in \mathcal{A} \right\}, \\ V_{\tilde{X}} &= \Big\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta_{[k]}(a) - \beta_{[k]}(0)\}^{\text{T}} \Sigma_{[k]XX} \{\beta_{[k]}(a') - \beta_{[k]}(0)\} : (a, a') \in \mathcal{A} \times \mathcal{A} \Big\}, \\ V_{\tilde{r}_{\text{gen}}\tilde{X}} &= \Big\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta_{[k]}(a') - \beta_{[k]}(0)\}^{\text{T}} \big[(\Sigma_{[k]XY(a)} - \Sigma_{[k]XY(0)}) - \Sigma_{[k]XX} \{\beta_{[k]}(a) - \beta_{[k]}(0)\} \big] : \\ &\qquad (a, a') \in \mathcal{A} \times \mathcal{A} \Big\}. \end{aligned}$$

Let $V_{\text{gen}} = V_{H\tilde{Y}} + V_{\tilde{r}_{\text{gen}}} + V_{\tilde{X}} + V_{\tilde{r}_{\text{gen}}\tilde{X}} + V^{\text{T}}_{\tilde{r}_{\text{gen}}\tilde{X}}$, we have

$$\sqrt{n}(\hat{\tau}_{\text{gen}} - \tau) \xrightarrow{d} \text{N}(0, V_{\text{gen}}).$$

■

Appendix C. Proof of main theorems

As we assume that $V_{\tilde{r}_{\text{proj}}}$, $V_{\tilde{X}}$ have finite limits, when no confusion arises, we use the same notation to denote their respective limits.

C.1 Proof of Proposition 1

Proof By the definition of lasso estimator, we have

$$\hat{\beta}_{[k]\text{lasso}}(a) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}\beta\}^2 + \lambda_{[k]a}\|\beta\|_1.$$

Therefore, we have

$$\begin{aligned}
&\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}\hat{\beta}_{[k]\text{lasso}}(a)\}^2 + \lambda_{[k]a}\|\hat{\beta}_{[k]\text{lasso}}(a)\|_1 \\
\leq &\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}\beta_{[k]\text{proj}}(a)\}^2 + \lambda_{[k]a}\|\beta_{[k]\text{proj}}(a)\|_1.
\end{aligned}$$

Let $u = \hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)$. Let $\epsilon_i(a) = Y_i(a) - X_i^{\mathrm{T}}\beta_{[k]\text{proj}}(a)$, for $i \in (a, k)$, be a transformed outcome. Here $i \in (a, k)$ means the units such that $A_i = a, B_i = k$. Also, we denote

$$\begin{aligned}
S_{[k]XX(a)} &= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(X_i - \bar{X}_{[k]a})(X_i - \bar{X}_{[k]a})^{\mathrm{T}} \\
S_{[k]X\epsilon(a)} &= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(X_i - \bar{X}_{[k]a})(\epsilon_i(a) - \bar{\epsilon}_{[k]a}).
\end{aligned}$$

Then by some calculations,

$$u^{\mathrm{T}} S_{[k]XX(a)} u \leq \lambda_{[k]a}\|\beta_{[k]\text{proj}}(a)\|_1 - \lambda_{[k]a}\|\hat{\beta}_{[k]\text{lasso}}(a)\|_1 + 2u^{\mathrm{T}} S_{[k]X\epsilon(a)}.$$

By Lemma 29, with high probability that

$$2\|S_{[k]X\epsilon(a)}\|_\infty \leq \frac{\lambda_{[k]a}}{2}.$$

Therefore, we have

$$u^{\mathrm{T}} S_{[k]XX(a)} u \leq \lambda_{[k]a}\|\beta_{[k]\text{proj}}(a)\|_1 - \lambda_{[k]a}\|\hat{\beta}_{[k]\text{lasso}}(a)\|_1 + \frac{\lambda_{[k]a}}{2}\|u\|_1.$$

As $S_{[k]}$ is the union of the support of $\beta_{[k]\text{proj}}(a)$ for $a \in \mathcal{A}_0$, we have

$$\|[\beta_{[k]\text{proj}}(a)]_{S_{[k]}^C}\|_1 = 0.$$

Therefore, we split the l_1 norm into the parts in $S_{[k]}$ and $S_{[k]}^C$, and by triangle inequality:

$$\begin{aligned}
\|[[\beta_{[k]\text{proj}}(a)]_{S_{[k]}}\|_1 - \|[\hat{\beta}_{[k]\text{lasso}}(a)]_{S_{[k]}}\|_1 \leq \|[u]_{S_{[k]}}\|_1, \\
\|[\hat{\beta}_{[k]\text{lasso}}(a)]_{S_{[k]}^C}\|_1 \geq \|[u]_{S_{[k]}^C}\|_1 - \|[\beta_{[k]\text{proj}}(a)]_{S_{[k]}^C}\|_1,
\end{aligned}$$

we have

$$u^{\mathrm{T}} S_{[k]XX(a)} u \leq \frac{3}{2}\lambda_{[k]a}\|u_{S_{[k]}}\|_1 - \frac{1}{2}\lambda_{[k]a}\|u_{S_{[k]}^C}\|_1.$$

As all three terms are greater than or equal to zero, we have $\|u_{S_{[k]}^C}\|_1 \leq 3\|u_{S_{[k]}}\|_1$. Let

$$E_3 = \{u \in V :\ u^{\mathrm{T}} S_{[k]XX(a)} u \geq \kappa \|u\|_2^2\},$$

where $V = \{h \in \mathbb{R}^p :\ \|h_{S_{[k]}^C}\|_1 \leq 3\|h_{S_{[k]}}\|_1\}$. By Lemma 30,

$$\Pr(E_3) \to 1.$$

Under E_3, we have

$$\kappa\|u\|_2^2 \leq \frac{3}{2}\lambda_{[k]a}\|u_{S_{[k]}}\|_1 \leq \frac{3}{2}\sqrt{s_{[k]}}\lambda_{[k]a}\|u_{S_{[k]}}\|_2 \leq \frac{3}{2}\sqrt{s_{[k]}}\lambda_{[k]a}\|u\|_2.$$

That is

$$\|u\|_2 \leq \frac{3}{2\kappa}\sqrt{s_{[k]}}\lambda_{[k]a}.$$

Then

$$\|u\|_1 = \|u_{S_{[k]}}\|_1 + \|u_{S_{[k]}^C}\|_1 \leq 4\|u_{S_{[k]}}\|_1 \leq 4\sqrt{s_{[k]}}\|u\|_2 \leq \frac{6}{\kappa} s_{[k]}\lambda_{[k]a}.$$

Therefore, as $\lambda_{[k]a} = O_p(\sqrt{M_{n[k]}\log p/n})$, we have

$$\|\hat{\beta}_{[k]\mathrm{lasso}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 = O_p\left(s_{[k]}\sqrt{\frac{M_{n[k]}\log p}{n}}\right).$$

Then by Lemma 28, $\|\bar{X}_{[k]a} - \bar{X}_{[k]}\|_\infty = O_p(\sqrt{\log p/n})$, we have

$$\begin{aligned}
|\sqrt{n}(\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}}(\hat{\beta}_{[k]\mathrm{lasso}}(a) - \beta_{[k]\mathrm{proj}}(a))| &= \sqrt{n}\cdot O_p\left(s_{[k]}\sqrt{\frac{M_{n[k]}\log p}{n}}\right)\cdot O_p\left(\sqrt{\frac{\log p}{n}}\right) \\
&= O_p\left(s_{[k]}\log p\sqrt{\frac{M_{n[k]}}{n}}\right) \\
&= o_p(1).
\end{aligned}$$

The last line is due to Assumption 5. The assumptions for Proposition 26 hold, so we have

$$\sqrt{n}(\hat{\tau}_{\mathrm{lasso}} - \tau) \xrightarrow{d} \mathrm{N}(0, V_{\mathrm{gen}}).$$

In Proposition 26, when $\beta_{[k]}(a) = \beta_{[k]\mathrm{proj}}(a)$ for $k \in \mathcal{K}, a \in \mathcal{A}_0$, we have

$$(\Sigma_{[k]XY(a)} - \Sigma_{[k]XY(0)}) - \Sigma_{[k]XX}\{\beta_{[k]}(a) - \beta_{[k]}(0)\} = 0.$$

Therefore, the asymptotic covariance matrix is $V_{\mathrm{proj}} = V_{\tilde{r}_{\mathrm{proj}}} + V_{\tilde{X}} + V_{H\tilde{Y}}$. ■

C.2 Proof of Theorem 1

Proof ***Transfer Step:***

By the definition of $\hat{\beta}^{(s)}_{[k]\text{proj}}(a)$, we have

$$\frac{1}{n^{(s)}_{[k]a}}\sum_{i\in(a,k)}\{Y^{(s)}_i-\bar{Y}^{(s)}_{[k]a}-(X^{(s)}_i-\bar{X}^{(s)}_{[k]a})^{\mathrm{T}}\hat{\beta}^{(s)}_{[k]\text{proj}}(a)\}^2+\lambda^{(s)}_{[k]a}\|\hat{\beta}^{(s)}_{[k]\text{proj}}(a)\|_1$$
$$\leq\frac{1}{n^{(s)}_{[k]a}}\sum_{i\in(a,k)}\{Y^{(s)}_i-\bar{Y}^{(s)}_{[k]a}-(X^{(s)}_i-\bar{X}^{(s)}_{[k]a})^{\mathrm{T}}\beta^{(s)}_{[k]\text{proj}}(a)\}^2+\lambda^{(s)}_{[k]a}\|\beta^{(s)}_{[k]\text{proj}}(a)\|_1.$$

Let $u^{(s)}=\hat{\beta}^{(s)}_{[k]\text{proj}}(a)-\beta^{(s)}_{[k]\text{proj}}(a)$. By some calculations, we have

$$(u^{(s)})^{\mathrm{T}}S^{(s)}_{[k]XX(a)}u^{(s)}\leq\lambda^{(s)}_{[k]a}\|\beta^{(s)}_{[k]\text{proj}}(a)\|_1-\lambda^{(s)}_{[k]a}\|\hat{\beta}^{(s)}_{[k]\text{proj}}(a)\|_1+2(u^{(s)})^{\mathrm{T}}S^{(s)}_{[k]X\epsilon(a)},$$

where

$$S^{(s)}_{[k]XX(a)}=\frac{1}{n^{(s)}_{[k]a}}\sum_{i=1}^{n^{(s)}}I(A^{(s)}_i=a,B^{(s)}_i=k)(X^{(s)}_i-\bar{X}^{(s)}_{[k]a})(X^{(s)}_i-\bar{X}^{(s)}_{[k]a})^{\mathrm{T}},$$

$$S^{(s)}_{[k]X\epsilon(a)}=\frac{1}{n^{(s)}_{[k]a}}\sum_{i=1}^{n^{(s)}}I(A^{(s)}_i=a,B^{(s)}_i=k)(X^{(s)}_i-\bar{X}^{(s)}_{[k]a})(\epsilon^{(s)}_i(a)-\bar{\epsilon}^{(s)}_{[k]a}).$$

Here $\epsilon^{(s)}_i(a)=Y^{(s)}_i(a)-X^{(s)}_i\beta_{[k]\text{proj}}(a)^{(s)}$, for $i\in(a,k)$.

Let $E_1=\{\|2S^{(s)}_{[k]X\epsilon(a)}\|_\infty\leq\lambda^{(s)}_{[k]a}/2\}$. By Lemma 29, we have

$$\Pr(E_1)\geq1-\frac{1}{M^{(s)}_{n[k]}}\to1.$$

Then under E_1, we have

$$(u^{(s)})^{\mathrm{T}}S^{(s)}_{[k]XX(a)}u^{(s)}\leq\lambda^{(s)}_{[k]a}\|\beta^{(s)}_{[k]\text{proj}}(a)\|_1-\lambda^{(s)}_{[k]a}\|\hat{\beta}^{(s)}_{[k]\text{proj}}(a)\|_1+\frac{\lambda^{(s)}_{[k]a}}{2}\|u^{(s)}\|_1.$$

Recall that $S_{[k]}=\bigcup_{a\in\mathcal{A}}\{j\in\{1,\ldots,p\}:\ \beta_{j,[k]\text{proj}}(a)\neq0\}$ and $S^C_{[k]}$ is the complementary set of $S_{[k]}$. Let $[u]_S=(u_j,j\in S_{[k]})$ and $[u]_{S^C}=(u_j,j\in S^C_{[k]})$ for any vector u. We split $S_{[k]}$ and $S^C_{[k]}$ apart and take the following inequalities: for $S=S_{[k]}$,

$$\|[\beta^{(s)}_{[k]\text{proj}}(a)]_S\|_1-\|[\hat{\beta}^{(s)}_{[k]\text{proj}}(a)]_S\|_1\leq\|[u^{(s)}]_S\|_1,$$

$$\|[u^{(s)}]_{S^C}\|_1\leq\|[\hat{\beta}^{(s)}_{[k]\text{proj}}(a)]_{S^C}\|_1+\|[\beta^{(s)}_{[k]\text{proj}}(a)]_{S^C}\|_1.$$

Then we have

$$(u^{(s)})^{\mathrm{T}}S^{(s)}_{[k]XX(a)}u^{(s)}\leq\frac{3}{2}\lambda^{(s)}_{[k]a}\|[u^{(s)}]_S\|_1+\frac{3}{2}\lambda^{(s)}_{[k]a}\|[\beta^{(s)}_{[k]\text{proj}}(a)]_{S^C}\|_1-\frac{1}{2}\lambda^{(s)}_{[k]a}\|[u^{(s)}]_{S^C}\|_1.$$

Then we compare the first term and the second term on the RHS.

Case 1 :

$$\frac{3}{2}\lambda_{[k]a}^{(s)}\|[u^{(s)}]_S\|_1 \geq \frac{3}{2}\lambda_{[k]a}^{(s)}\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1.$$

Then we have $\|[u^{(s)}]_{S^C}\|_1 \leq 6\|[u^{(s)}]_S\|_1$. By Lemma 30, we have

$$c_0\|u^{(s)}\|_2^2 \leq 3\lambda_{[k]a}^{(s)}\|[u^{(s)}]_S\|_1 \leq 3\sqrt{s_{[k]}}\lambda_{[k]a}^{(s)}\|[u^{(s)}]_S\|_2,$$

which leads to

$$\|u^{(s)}\|_2 \leq C_1\sqrt{s_{[k]}}\lambda_{[k]a}^{(s)}$$

and

$$\|u^{(s)}\|_1 \leq 7\|[u^{(s)}]_S\|_1 \leq 7\sqrt{s_{[k]}}\|[u^{(s)}]_S\|_2 \leq C_2 s_{[k]}\lambda_{[k]a}^{(s)}.$$

Case 2 :

$$\frac{3}{2}\lambda_{[k]a}^{(s)}\|[u^{(s)}]_S\|_1 < \frac{3}{2}\lambda_{[k]a}^{(s)}\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1.$$

Then

$$(u^{(s)})^{\mathrm{T}}S_{[k]XX(a)}^{(s)}u^{(s)} \leq 3\lambda_{[k]a}^{(s)}\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1 - \frac{1}{2}\lambda_{[k]a}^{(s)}\|[u^{(s)}]_{S^C}\|_1.$$

A direct bound for $\|u^{(s)}\|_1$ is

$$\|u^{(s)}\|_1 \leq \|[u^{(s)}]_S\|_1 + \|[u^{(s)}]_{S^C}\|_1 \leq 7\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1 \leq 7\|[\delta_{[k]}(a)]_{S^C}\|_1 \leq 7h.$$

The third inequality holds as

$$\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1 \leq \|[\beta_{[k]\text{proj}}(a)]_{S^C}\|_1 + \|[\delta_{[k]}(a)]_{S^C}\|_1 = \|[\delta_{[k]}(a)]_{S^C}\|_1.$$

Also, by Lemma 31, we have

$$\kappa_1\|u^{(s)}\|_2^2 - \kappa_3\frac{\log p}{n_{[k]a}^{(s)}}\|u^{(s)}\|_1^2 \leq 3\lambda_{[k]a}^{(s)}\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1 - \frac{1}{2}\lambda_{[k]a}^{(s)}\|[u^{(s)}]_{S^C}\|_1.$$

Under the condition that $h(\log p/n^{(s)})^{1/2} = o(1)$, we have

$$\kappa_1\|u^{(s)}\|_2^2 \leq 3\lambda_{[k]a}^{(s)}\|[\beta_{[k]\text{proj}}^{(s)}(a)]_{S^C}\|_1 \leq C_3\lambda_{[k]a}^{(s)}h.$$

This is used for the proof in the debiasing step. Therefore, we have

$$\|u^{(s)}\|_1 \leq C_2 s_{[k]}\lambda_{[k]a}^{(s)} + C_4 h.$$

Debiasing Step:

Let $u = \hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)$. By the definition of $\hat{\beta}_{[k]\text{tl}}(a) = \hat{\delta}_{[k]}(a) + \hat{\beta}_{[k]\text{proj}}^{(s)}(a)$, we have

$$\frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}\hat{\beta}_{[k]\text{tl}}(a)\}^2 + \lambda_{[k]a}\|\hat{\delta}_{[k]}(a)\|_1$$

$$\leq \frac{1}{n_{[k]a}}\sum_{i=1}^{n} I(A_i = a, B_i = k)\{Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\mathrm{T}}\beta_{[k]\text{proj}}(a)\}^2 + \lambda_{[k]a}\|\beta_{[k]\text{proj}}(a) - \hat{\beta}_{[k]\text{proj}}^{(s)}(a)\|_1.$$

Then, by similar calculations in the Transfer step, we have

$$u^{\mathrm{T}} S_{[k]XX(a)} u \leq \lambda_{[k]a} \|\beta_{[k]\mathrm{proj}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a)\|_1 - \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + 2u^{\mathrm{T}} \left[\frac{1}{n_{[k]a}} \sum_{i \in (a,k)} (X_i - \bar{X}_{[k]a}) \{ Y_i - \bar{Y}_{[k]a} - (X_i - \bar{X}_{[k]a})^{\top} \beta_{[k]\mathrm{proj}}(a) \} \right].$$

Recall that $\epsilon_i(a) = Y_i(a) - X_i^{\mathrm{T}} \beta_{[k]\mathrm{proj}}(a)$, then we have

$$u^{\mathrm{T}} S_{[k]XX(a)} u \leq \lambda_{[k]a} \|\beta_{[k]\mathrm{proj}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a)\|_1 - \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + 2u^{\mathrm{T}} S_{[k]X\epsilon(a)}.$$

Here $S_{[k]XX(a)}$ and $S_{[k]X\epsilon(a)}$ are the same as the ones defined in the Proof of Proposition 1. Let

$$E_2 = \left\{ \|2S_{[k]X\epsilon(a)}\|_\infty \leq \frac{\lambda_{[k]a}}{2} \right\} \cap E_1.$$

By Lemma 29 we have

$$\Pr(E_2) \geq 1 - \frac{1}{M_{n[k]}} \to 1.$$

Therefore, under E_2, we have

$$\begin{aligned} u^{\mathrm{T}} S_{[k]XX(a)} u &\leq \lambda_{[k]a} \|\beta_{[k]\mathrm{proj}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a)\|_1 - \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + 2u^{\mathrm{T}} S_{[k]X\epsilon(a)} \\ &\leq \lambda_{[k]a} \|\beta_{[k]\mathrm{proj}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a)\|_1 - \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + \|u\|_1 \|2S_{[k]X\epsilon(a)}\|_\infty \\ &\leq \lambda_{[k]a} \|\beta_{[k]\mathrm{proj}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a)\|_1 - \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + \frac{\lambda_{[k]a}}{2} \|u\|_1. \end{aligned}$$

We have the following inequalities

$$\begin{aligned} \|u\|_1 &= \|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 \\ &= \|\hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a) + \hat{\delta}_{[k]}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 \\ &\leq \|\hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 + \|\hat{\delta}_{[k]}(a)\|_1, \end{aligned}$$

and

$$\|\hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 = \|u^{(s)} - \delta_{[k]}(a)\|_1 \leq \|\delta_{[k]}(a)\|_1 + \|u^{(s)}\|_1.$$

Then we have

$$\begin{aligned} u^{\mathrm{T}} S_{[k]XX(a)} u &\leq \lambda_{[k]a} \|\beta_{[k]\mathrm{proj}}(a) - \hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a)\|_1 - \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + \frac{\lambda_{[k]a}}{2} \|u\|_1 \\ &\leq \frac{3}{2} \lambda_{[k]a} \|\delta_{[k]}(a)\|_1 - \frac{1}{2} \lambda_{[k]a} \|\hat{\delta}_{[k]}(a)\|_1 + \frac{3}{2} \lambda_{[k]a} \|u^{(s)}\|_1 \end{aligned}$$

As the LHS is greater than or equal to zero, we have

$$\|\hat{\delta}_{[k]}(a)\|_1 \leq 3\|\delta_{[k]}(a)\|_1 + 3\|u^{(s)}\|_1 \leq (3 + C_4)h + 3C_2 s_{[k]} \lambda^{(s)}_{[k]a}.$$

Also, we can derive that

$$\|\hat{\delta}_{[k]}(a) - \delta_{[k]}(a)\|_1 \leq \|\hat{\delta}_{[k]}(a)\|_1 + \|\delta_{[k]}(a)\|_1 \leq (4 + C_4)h + 3C_2 s_{[k]} \lambda^{(s)}_{[k]a}.$$

Therefore, according to the inequalities for $\|u^{(s)}\|_1$ and $\|\hat{\delta}_{[k]}(a) - \delta_{[k]}(a)\|_1$, with high probability, we have

$$\begin{aligned}\|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 &= \|\{\hat{\beta}^{(s)}_{[k]\mathrm{proj}}(a) - \beta^{(s)}_{[k]\mathrm{proj}}(a)\} + \{\hat{\delta}_{[k]}(a) - \delta_{[k]}(a)\}\|_1 \\ &\leq K_1 s_{[k]} \sqrt{\frac{M^{(s)}_{n[k]} \log p}{n^{(s)}}} + K_2 h\end{aligned}$$

for some constant K_1 and K_2 independent of $n, n^{(s)}$, and p. When $h < s_{[k]}\sqrt{\log p/n}$, we can write the above term as

$$\|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 \leq K_1 s_{[k]} \sqrt{\frac{M^{(s)}_{n[k]} \log p}{n^{(s)} + n}} + K_2 \left(h \wedge s_{[k]} \sqrt{\frac{M_{n[k]} \log p}{n}} \right).$$

■

C.3 Proof of Theorem 2

Proof According to Theorem 15 and Lemma 28, we have

$$\|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 = O_p\left(s_{[k]} \sqrt{\frac{M^{(s)}_{n[k]} \log p}{n + n^{(s)}}} + h \right),$$

and

$$\|\bar{X}_{[k]a} - \bar{X}_{[k]}\|_\infty = O_p\left\{ \left(\frac{\log p}{n} \right)^{1/2} \right\}$$

for all $k \in \mathcal{K}$ and $a \in \mathcal{A}_0$. By Hölder's inequality, we have

$$\begin{aligned}&|\sqrt{n}(\bar{X}_{[k]a} - \bar{X}_{[k]})^{\mathrm{T}} \{\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\}| \\ &\leq \sqrt{n} \|\bar{X}_{[k]a} - \bar{X}_{[k]}\|_\infty \|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{proj}}(a)\|_1 \\ &= \sqrt{n} \cdot O_p\left(s_{[k]} \sqrt{\frac{M^{(s)}_{n[k]} \log p}{n + n^{(s)}}} + h \right) \cdot O_p\left\{ \left(\frac{\log p}{n} \right)^{1/2} \right\} \\ &= O_p\left(s_{[k]} \log p \sqrt{\frac{M_n}{n + n^{(s)}}} + h\sqrt{\log p} \right) \\ &= o_p(1)\end{aligned}$$

The last line is due to Assumption 14 and the condition that $h(\log p)^{1/2} = o(1)$. Therefore, by Proposition 26 and similar arguments for the asymptotic covariance matrix in the proof of Proposition 6, we have

$$\sqrt{n}(\hat{\tau}_{\text{tl}} - \tau) \xrightarrow{d} \text{N}(0, V_{\text{proj}}).$$

■

C.4 Proof of Theorem 3

Proof By Lemma 32, we only need to show that

$$\frac{1}{n_{[k]}} \sum_{i \in [k]} [X_i^{\text{T}} \{\hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)\}]^2 = o_p(1)$$

and

$$\frac{1}{n_{[k]}} \sum_{i \in [k]} [X_i^{\text{T}} \{\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)\}]^2 = o_p(1),$$

respectively. By Proposition 6 and Theorem 16, we have

$$\|\hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)\|_1 = O_p\left(s_{[k]}\sqrt{\frac{M_n \log p}{n}}\right) = o_p(1)$$

and

$$\|\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)\|_1 = O_p\left(s_{[k]}\sqrt{\frac{M_n^{(s)} \log p}{n + n^{(s)}}} + h\right) = o_p(1).$$

By Hölder inequality and X_i is uniformly bounded, we have

$$\frac{1}{n_{[k]}} \sum_{i \in [k]} [X_i^{\text{T}} \{\hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)\}]^2 \leq M^2 \|\hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)\|_1^2 = o_p(1)$$

and

$$\frac{1}{n_{[k]}} \sum_{i \in [k]} [X_i^{\text{T}} \{\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)\}]^2 \leq M^2 \|\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)\|_1^2 = o_p(1).$$

Therefore our variance estimators are consistent. ■

C.5 Lemmas with proof

Lemma 27 *Let* $\mathbb{L}_n^{(2)} = \{\mathbb{L}_{n[k]}^{(2)}, k \in \mathcal{K}\}$ *and*

$$\mathbb{L}_n = \left(\left\{ \sum_{k \in \mathcal{K}} \frac{1}{\pi_{[k]a}} \mathbb{L}_{n[k], r_{i,\text{gen}}(a)}^{(1)}, a \in \mathcal{A}_0 \right\}, \left\{ \sum_{k \in \mathcal{K}} \mathbb{L}_{n[k], X_i}^{(1)\text{T}} \beta_{[k]}(a), a \in \mathcal{A}_0 \right\}, \mathbb{L}_n^{(2)} \right).$$

Under Assumptions 1–2, we have

$$\mathbb{L}_n \xrightarrow{d} \mathrm{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}\Sigma_1 & \Sigma_{12} & 0\\ \Sigma_{12}^{\mathrm{T}} & \Sigma_2 & 0\\ 0 & 0 & \Sigma_3\end{pmatrix}\right),$$

where

$$\begin{aligned}
\Sigma_1 &= \mathrm{diag}\left\{\sum_{k\in\mathcal{K}} \frac{p_{[k]}}{\pi_{[k]a}}\sigma^2_{[k]\tilde{r}_{\mathrm{gen}}(a)}, a\in\mathcal{A}_0\right\},\\
\Sigma_{12} &= \left\{\sum_{k\in\mathcal{K}} p_{[k]}\beta^{\mathrm{T}}_{[k]}(a')\left(\Sigma_{[k]XY(a)} - \Sigma_{[k]XX}\beta_{[k]}(a)\right), (a,a')\in\mathcal{A}_0\times\mathcal{A}_0\right\},\\
\Sigma_2 &= \left\{\sum_{k\in\mathcal{K}} \beta^{\mathrm{T}}_{[k]}(a)\Sigma_{[k]XX}\beta_{[k]}(a'), (a,a')\in\mathcal{A}_0\times\mathcal{A}_0\right\},\\
\Sigma_3 &= \mathrm{diag}\{p_{[k]} : k\in\mathcal{K}\} - \{p_{[k]}p_{[k']}, (k,k')\in\mathcal{K}\times\mathcal{K}\}.
\end{aligned}$$

Proof Most of the proof is similar to the proof for Lemma 5 in Gu et al. (2023). The difference is that in this Lemma, $\beta_{[k]}(a)$ may not be $\beta_{[k]\mathrm{proj}}(a)$ as we define in the context. The consequence is that $\Sigma_{12}\neq 0$ for most cases. The remaining parts are the same and we omit the proof. ■

Lemma 28 *Under Assumptions 1 and 2, we have*

$$\|\bar{X}_{[k]a} - \bar{X}_{[k]}\|_\infty = O_p\left\{\left(\frac{\log p}{n}\right)^{1/2}\right\}.$$

for all $a\in\mathcal{A}_0,\ k\in\mathcal{K}$.

Proof The proof is similar to the proof of Lemma S4 in Liu et al. (2023). We can replace $A_i = 1$ into $A_i = a$ for all $a\in\mathcal{A}_0$ to show our result. Therefore, we omit the proof. ■

Lemma 29 *Under Assumptions 1-5 and $E\{\epsilon_i^2(a)\} < \infty$ for all $a\in\mathcal{A}_0$, the events*

$$E_1 = \left\{\|2S^{(s)}_{[k]X\epsilon(a)}\|_\infty \le \frac{\lambda^{(s)}_{[k]a}}{2}\right\},\quad E_2' = \left\{\|2S_{[k]X\epsilon(a)}\|_\infty \le \frac{\lambda_{[k]a}}{2}\right\}$$

have probability tending to one when the source size $n^{(s)}$ and target size n go to infinity.

Proof The proof is similar to the proof of Lemma S5 in Liu et al. (2023). It's a special case for stratum size $K = 1$. At the same time, we can replace $A_i = 1$ into $A_i = a$ for all $a\in\mathcal{A}_0$ to show our result. Therefore, we omit the proof. ■

Lemma 30 *Under Assumptions 1–5, for any* $u \in V = \{h \in \mathbb{R}^p : \ \|h_{S_{[k]}^C}\|_1 \leq 3\|h_{S_{[k]}}\|_1\}$*, we have*

$$u^{\mathrm{T}} S_{[k]XX(a)} u \geq \kappa \|u\|_2^2$$

holds with probability $1 - c_1 \exp(-c_2 n)$*, for some positive constant* κ *independent of* n*.*

Proof We use a similar argument in Bugni et al. (2018). As $\{Y_i(0), \ldots, Y_i(A), B_i, X_i\}$ are i.i.d. and $A^{(n)}$ is independent of $\{Y_i(0), \ldots, Y_i(A), X_i\}$ conditional on $B^{(n)}$. Therefore, given $A^{(n)}$ and $B^{(n)}$, the distribution of $S_{[k]XX(a)}$ is the same if we order the units first by strata and then by treatments within each strata. Then, independently with $k = 1, 2, \ldots, K$ and $(A^{(n)}, B^{(n)})$, let $\{Y_i^k(0), \ldots, Y_i^k(A), X_i^k\}$ has the same distribution as $\{Y_i(0), \ldots, Y_i(A), X_i\}$ given $B_i = k$. Then $S_{[k]XX(a)} | A^{(n)}, B^{(n)}$ has the same distribution as

$$\frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} \left(X_i^k - \frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \right) \left(X_i^k - \frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \right)^{\mathrm{T}}.$$

By Assumption 1, X_i is uniformly bounded by a constant M. As X_i^k has the same distribution as X_i given $B_i = k$, X_i^k is uniformly bounded by the constant M so that it is a sub-Gaussian random vector. According to Proposition 2 in Negahban et al. (2009) for generalized linear models, it can be applied directly to the standard linear model. For any $N \geq 1$, we have

$$\begin{aligned} & h^{\mathrm{T}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k)(X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k)^{\mathrm{T}} \right\} h \\ & \geq \kappa_1 \|h\|_2 \left\{ \|h\|_2 - \kappa_2 \sqrt{\frac{\log p}{N}} \|h\|_1 \right\}, \quad \text{for all } \|h\|_2 \leq 1 \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 N)$. Here κ_1, κ_2 only depends on the minimum eigenvalue of covariance matrix $\Sigma_{[k]XX}$ and M.

As we can divide $\|h\|_2$ on both sides, we have

$$h^{\mathrm{T}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k)(X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k)^{\mathrm{T}} \right\} h \geq \kappa_1 \|h\|_2 \left\{ \|h\|_2 - \kappa_2 \sqrt{\frac{\log p}{N}} \|h\|_1 \right\}.$$

Next, we consider the vectors in the convex cone

$$V = \{h \in \mathbb{R}^p : \ \|h_{S_{[k]}^C}\|_1 \leq 3\|h_{S_{[k]}}\|_1\}.$$

For any vector $h \in V$,

$$\|h\|_1 = \|h_{S_{[k]}}\|_1 + \|h_{S_{[k]}^C}\|_1 \leq 4\|h_{S_{[k]}}\|_1 \leq 4\sqrt{s_{[k]}} \|h_{S_{[k]}}\|_2.$$

Therefore,

$$\begin{aligned}
&h^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}(X_i^k-\frac{1}{N}\sum_{i=1}^{N}X_i^k)(X_i^k-\frac{1}{N}\sum_{i=1}^{N}X_i^k)^{\mathrm{T}}\right\}h\\
&\geq \kappa_1\|h\|_2\left\{\|h\|_2-\kappa_2\sqrt{\frac{\log p}{N}}\|h\|_1\right\}\\
&\geq \left(1-4\kappa_2\sqrt{\frac{s_{[k]}\log p}{N}}\right)\kappa_1\|h\|_2^2.
\end{aligned}$$

Under the assumption $M_{n[k]}s_{[k]}^2\log^2 p/n\to 0$, we have

$$1-4\kappa_2\sqrt{\frac{s_{[k]}\log p}{n_{[k]}}}>\frac{1}{2}$$

when $n\to\infty$. Therefore,

$$\Pr\left(h^{\mathrm{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}(X_i^k-\frac{1}{N}\sum_{i=1}^{N}X_i^k)(X_i^k-\frac{1}{N}\sum_{i=1}^{N}X_i^k)^{\mathrm{T}}\right\}h\geq\frac{1}{2}\kappa_1\|h\|_2^2\right)\to 1, \tag{1}$$

when $N\to\infty$. Using the almost sure representation theorem, we can construct $\tilde{n}_{[k]a}$ (independent of $\{Y_i^k(0),\ldots,Y_i^k(A),X_i^k\}$) such that $\tilde{n}_{[k]a}/n$ has the same distribution as $n_{[k]a}/n$ and $\tilde{n}_{[k]a}/n\to\pi_{[k]a}p_{[k]}$ almost surely. Therefore, for any $h\in V$,

$$\begin{aligned}
&\Pr\left(h^{\mathrm{T}}\left\{\frac{1}{n_{[k]a}}\sum_{i=1}^{n_{[k]a}}(X_i^k-\frac{1}{n_{[k]a}}\sum_{i=1}^{n_{[k]a}}X_i^k)(X_i^k-\frac{1}{n_{[k]a}}\sum_{i=1}^{n_{[k]a}}X_i^k)^{\mathrm{T}}\right\}h\geq\frac{1}{2}\kappa_1\|h\|_2^2\;\Big|\;A^{(n)},B^{(n)}\right)\\
&=\Pr\left(h^{\mathrm{T}}\left\{\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}}\sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}}(X_i^k-\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}}\sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}}X_i^k)(X_i^k-\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}}\sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}}X_i^k)^{\mathrm{T}}\right\}h\geq\frac{1}{2}\kappa_1\|h\|_2^2\;\Big|\;A^{(n)},B^{(n)}\right)\\
&=E\Bigg[\Pr\left(h^{\mathrm{T}}\left\{\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}}\sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}}(X_i^k-\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}}\sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}}X_i^k)(X_i^k-\frac{1}{n\frac{\tilde{n}_{[k]a}}{n}}\sum_{i=1}^{n\frac{\tilde{n}_{[k]a}}{n}}X_i^k)^{\mathrm{T}}\right\}h\geq\frac{1}{2}\kappa_1\|h\|_2^2\right.\\
&\qquad\left.\;\Big|\;A^{(n)},B^{(n)},\frac{\tilde{n}_{[k]a}}{n}\right)\Bigg]\to 1,
\end{aligned}$$

where the convergence follows from the dominated convergence theorem, $n(\tilde{n}_{[k]a}/n)\to\infty$ a.s., independence of $\tilde{n}_{[k]a}/n$ and $\{Y_i(0)^k,\ldots,Y_i^k(A),X_i^k\}$ and (1). Therefore,

$$\begin{aligned}
&\Pr\left(h^{\mathrm{T}}S_{[k]XX(a)}h\geq\frac{1}{2}\kappa_1\|h\|_2^2\right)\\
&=\Pr\left(h^{\mathrm{T}}\left\{\frac{1}{n_{[k]a}}\sum_{i=1}^{n_{[k]a}}(X_i^k-\frac{1}{n_{[k]a}}\sum_{i=1}^{n_{[k]a}}X_i^k)(X_i^k-\frac{1}{n_{[k]a}}\sum_{i=1}^{n_{[k]a}}X_i^k)^{\mathrm{T}}\right\}h\geq\frac{1}{2}\kappa_1\|h\|_2^2\right)\\
&\to 1.
\end{aligned}$$

■

Lemma 31 *Under Assumptions 1–4 and 10–13, we have*

$$u^{\mathrm{T}} S_{[k]XX(a)} u \geq \kappa_1 \|u\|_2^2 - \kappa_3 \frac{\log p}{n} \|u\|_1^2$$

with probability $1 - c_1 \exp(-c_2 n)$*, where* $\kappa_1, \kappa_3, c_1, c_2$ *are constants independent of* n*.*

Proof Similar to the proof of Lemma 30, we construct $\{Y_i^k(0), \ldots, Y_i^k(A), X_i^k\}$ and $S_{[k]XX(a)}$ has the same distribution as

$$\frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} \left(X_i^k - \frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \right) \left(X_i^k - \frac{1}{n_{[k]a}} \sum_{i=1}^{n_{[k]a}} X_i^k \right)^{\mathrm{T}}.$$

Also, by the same argument in Negahban et al. (2009), when $N \to \infty$,

$$h^{\mathrm{T}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k)(X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k)^{\mathrm{T}} \right\} h \geq \kappa_1 \|h\|_2 \left\{ \|h\|_2 - \kappa_2 \sqrt{\frac{\log p}{N}} \|h\|_1 \right\}.$$

By a general inequality, we have

$$\kappa_1 \kappa_2 \sqrt{\frac{\log p}{N}} \|h\|_1 \|h\|_2 \leq \frac{\alpha}{2} \|h\|_2^2 + \frac{(\kappa_1 \kappa_2)^2}{2\alpha} \frac{\log p}{N} \|h\|_1^2,$$

for some positive constant α. We can choose $\alpha = \kappa_1$, then we have

$$h^{\mathrm{T}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k \right) \left(X_i^k - \frac{1}{N} \sum_{i=1}^{N} X_i^k \right)^{\mathrm{T}} \right\} h \geq \frac{\kappa_1}{2} \|h\|_2^2 - \kappa_3 \frac{\log p}{N} \|h\|_1^2,$$

for some positive constant κ_1 and κ_3. Then by the similar arguments in the proof of Lemma 30, we have

$$\Pr \left(h^{\mathrm{T}} S_{[k]XX(a)} h \geq \frac{\kappa_1}{2} \|h\|_2^2 - \kappa_3 \frac{\log p}{N} \|h\|_1^2 \right) \to 1.$$

■

Lemma 32 *Suppose that the conditions in Proposition 26 hold, and that there exist* $\hat{\beta}_{[k]}(a)$ *and* $\beta_{[k]}(a)$*, such that*

$$\frac{1}{n_{[k]}} \sum_{i \in [k]} [X_i^{\mathrm{T}} \{ \hat{\beta}_{[k]}(a) - \beta_{[k]}(a) \}]^2 = o_p(1),$$

for all $k \in \mathcal{K}, a \in \mathcal{A}_0$*. Let* $\hat{\sigma}^2_{\mathrm{gen,bc}}$ *be the estimator of* $\sigma^2_{\mathrm{gen,bc}}$*, such that*

$$\sigma^2_{\mathrm{gen,bc}} = e_{bc}^{\mathrm{T}} V_{\mathrm{gen}} e_{bc} = V_{\mathrm{gen}}(b,b) - 2V_{\mathrm{gen}}(b,c) + V_{\mathrm{gen}}(c,c).$$

Then we have

$$\begin{aligned}\hat{\sigma}^2_{\text{gen,bc}} = \hat{\sigma}^2_{r_{\text{gen}}(b)} + \hat{\sigma}^2_{r_{\text{gen}}(c)} + \hat{\sigma}^2_{bc,HY} \\ - \sum_{k\in\mathcal{K}} p_{n[k]} \big\{\hat{\beta}_{[k]}(b) - \hat{\beta}_{[k]}(c)\big\}^{\mathrm{T}} S_{[k]XX} \big\{\hat{\beta}_{[k]}(b) - \hat{\beta}_{[k]}(c)\big\} \\ + 2\sum_{k\in\mathcal{K}} p_{n[k]} \big\{\hat{\beta}_{[k]}(b) - \hat{\beta}_{[k]}(c)\big\}^{\mathrm{T}} \big\{(S_{[k]XY(b)} - S_{[k]XY(c)})\big\}\end{aligned}$$

is a consistent estimator for $\sigma^2_{\text{gen,bc}}$.

Proof The proof is similar to that of Theorem 3 in Gu et al. (2023). The proof for $\hat{\sigma}^2_{bc,HY}$ stays the same. We will focus on $\sigma^2_{r_{\text{gen}}(b)}$ and the other two terms.

The consistency of the estimator can be implied by

$$\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(\hat{r}_i - \bar{\hat{r}}_{[k]a})^2 \xrightarrow{P} \text{var}(r_i | B_i = k)$$

for $r_i = Y_i - X_i^{\mathrm{T}}\beta_{[k]}(a)$ and $X_i^{\mathrm{T}}\{\beta_{[k]}(b) - \beta_{[k]}(c)\}$, and

$$\begin{aligned}\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{\beta_{[k]}(b) - \beta_{[k]}(c)\}(X_i - \bar{X}_{[k]a})^{\mathrm{T}}(Y_i - \bar{Y}_{[k]a}) \\ \xrightarrow{P} \{\beta_{[k]}(b) - \beta_{[k]}(c)\}\Sigma_{[k]XY(a)}.\end{aligned}$$

By Lemma 33 below, we have

$$\begin{aligned}\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(Y_i - \bar{Y}_{[k]a})^2 \xrightarrow{P} \text{var}\{Y_i(a) | B_i = k\}, \\ \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(r_i - \bar{r}_{[k]a})^2 \xrightarrow{P} \text{var}\{r_i(a) | B_i = k\}.\end{aligned}$$

Then we have

$$\begin{aligned}&\frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(\hat{r}_i - \bar{\hat{r}}_{[k]a})^2 \\ &= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(r_i - \bar{r}_{[k]a})^2 \\ &- \frac{2}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(r_i - \bar{r}_{[k]a})(X_i - \bar{X}_{[k]a})^{\mathrm{T}}(\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)) \\ &+ \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\big[(X_i - \bar{X}_{[k]a})^{\mathrm{T}}\{\hat{\beta}_{[k]}(a) - \beta_{[k]}(a)\}\big]^2 \\ &= \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)(r_i - \bar{r}_{[k]a})^2 + o_p(1) \\ &\xrightarrow{P} \text{var}(Y_i(a) | B_i = k).\end{aligned}$$

The third term converges to zero due to the condition that

$$\frac{1}{n_{[k]}}\sum_{i\in[k]}[X_i^{\mathrm{T}}\{\hat{\beta}_{[k]}(a)-\beta_{[k]}(a)\}]^2=o_p(1),$$

and the second term converges to zero due to the Cauchy-Schwarz inequality. Similarly, in the last term of the variance estimator, we can only deal with

$$\left\{\hat{\beta}_{[k]}(b)-\hat{\beta}_{[k]}(c)\right\}^{\mathrm{T}}\left\{S_{[k]XY(b)}-S_{[k]XY(c)}\right\}.$$

Also, it is sufficient to show that for any treatments b and c, we have

$$\{\hat{\beta}_{[k]}(b)\}^{\mathrm{T}}S_{[k]XY(c)}\xrightarrow{P}\beta_{[k]}^{\mathrm{T}}(b)\Sigma_{[k]XY(c)}.$$

The left-hand side can be expressed as

$$\begin{aligned}\{\hat{\beta}_{[k]}(b)\}^{\mathrm{T}}S_{[k]XY(c)}=\beta_{[k]}^{\mathrm{T}}(b)\frac{1}{n_{[k]c}}\sum_{i=1}^{n}I(A_i=c,B_i=k)(X_i-\bar{X}_{[k]c})^{\mathrm{T}}(Y_i-\bar{Y}_{[k]c})\\+\frac{1}{n_{[k]c}}\{\hat{\beta}_{[k]}(b)-\beta_{[k]}(b)\}^{\mathrm{T}}\sum_{i=1}^{n}I(A_i=c,B_i=k)(X_i-\bar{X}_{[k]c})(Y_i-\bar{Y}_{[k]c}).\end{aligned}$$

As

$$\begin{aligned}&\frac{1}{n_{[k]c}}\sum_{i=1}^{n}I(A_i=c,B_i=k)[(X_i-\bar{X}_{[k]c})^{\mathrm{T}}\{\hat{\beta}_{[k]}(b)-\beta_{[k]}(b)\}]^2\\&\qquad\le\frac{1}{n_{[k]c}}\sum_{i=1}^{n}I(B_i=k)[(X_i-\bar{X}_{[k]c})^{\mathrm{T}}\{\hat{\beta}_{[k]}(b)-\beta_{[k]}(b)\}]^2\\&\qquad\le\frac{1}{n_{[k]}}\sum_{i=1}^{n}I(B_i=k)[X_i^{\mathrm{T}}\{\hat{\beta}_{[k]}(b)-\beta_{[k]}(b)\}]^2\\&\qquad=o_p(1),\end{aligned}$$

and the second term is $o_p(1)$ due to the Cauchy-Schwarz inequality, we show the convergence of the last term in the estimator.

Finally, we need to show that

$$\left\{\hat{\beta}_{[k]}(b)-\hat{\beta}_{[k]}(c)\right\}^{\mathrm{T}}S_{[k]XX}\left\{\hat{\beta}_{[k]}(b)-\hat{\beta}_{[k]}(c)\right\}\xrightarrow{P}\left\{\beta_{[k]}(b)-\beta_{[k]}(c)\right\}^{\mathrm{T}}\Sigma_{[k]XX}\left\{\beta_{[k]}(b)-\beta_{[k]}(c)\right\}.$$

We can write the LHS as

$$\frac{1}{n_{[k]a}}\sum_{i=1}^{n}I(B_i=k)\left[(X_i-\bar{X}_{[k]})^{\mathrm{T}}\left\{\hat{\beta}_{[k]}(b)-\hat{\beta}_{[k]}(c)\right\}\right]^2$$

and split the term into three parts:

$$\begin{aligned}&\frac{1}{n_{[k]a}}\sum_{i=1}^{n}I(B_i=k)\left[(X_i-\bar{X}_{[k]})^{\mathrm{T}}\left\{\beta_{[k]}(b)-\beta_{[k]}(c)\right\}\right]^2,\\&\frac{1}{n_{[k]a}}\sum_{i=1}^{n}I(B_i=k)\left[(X_i-\bar{X}_{[k]})^{\mathrm{T}}\left\{\hat{\beta}_{[k]}(b)-\beta_{[k]}(b)-(\hat{\beta}_{[k]}(c)-\beta_{[k]}(c))\right\}\right]^2,\end{aligned}$$

and their production. Similar to the former arguments, the first term converges to the required result. The second and third terms converge to zero due to the condition and the Cauchy-Schwarz inequality. ■

Lemma 33 *For $f_i = f(Y_i(0), \ldots, Y_i(A), B_i, X_i)$ that satisfies $E[f_i^2] < \infty$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} f_i I(A_i = a) \xrightarrow{P} \sum_{k\in\mathcal{K}} p_{[k]}\pi_{[k]a} E(f_i|B_i = k),$$
$$\frac{1}{n}\sum_{i=1}^{n} f_i I(A_i = a, B_i = k) \xrightarrow{P} p_{[k]}\pi_{[k]a} E(f_i|B_i = k).$$

Proof It is a generalized version of Lemma C.4 from Bugni et al. (2019) and Lemma 1 from Ma et al. (2022a). They proved their lemmas by setting $f_i = f(Y_i(0), \ldots, Y_i(A), B_i)$ and $f_i = f(Y_i(0), Y_i(1), X_i)$. The proof can be easily generalized to $f_i = f(Y_i(0), \ldots, Y_i(A), B_i, X_i)$ so we omit it. ■

Appendix D. Additional simulations

D.1 Unequal allocation ratios within each stratum in Section 6

Figures 5–7 represent the simulation results when the allocation ratios are unequal within each stratum. The allocation ratios are 1:1:1 in the first strata and 2:2:1 in the other strata, and the block size for the stratified block randomization is 15. Other settings are the same as the settings in Section 6 in the main text. The simulation results are similar to the results in the main text.

D.2 Additional illustration of the l_1 norm bound in Theorem 16

In this subsection, we present additional simulations to illustrate the l_1 norm results described in Theorem 16, which are closely related to the difference between the target and source coefficients, denoted by h. Specifically, we compare $\|\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)\|_1$ with $\|\hat{\beta}_{[k]\text{lasso}}(a) - \beta_{[k]\text{proj}}(a)\|_1$ to show how increasing h affects $\|\hat{\beta}_{[k]\text{tl}}(a) - \beta_{[k]\text{proj}}(a)\|_1$.

D.2.1 Theoretical bound

We begin by numerically evaluating the theoretical bounds established in Proposition 6 and Theorem 16. Figure 8 displays the upper bound in Theorem 16 as a function of the heterogeneity parameter h, together with the corresponding l_1-norm bound for $\hat{\beta}_{[k]\text{lasso}}(a)$ given in Proposition 6. This comparison clarifies how the tightness and informativeness of the transfer learning bound depend on the similarity between the source and target projection coefficients. As h increases, the bound in Theorem 16 grows monotonically and eventually exceeds the lasso bound, indicating that the theoretical advantage of transfer learning diminishes when the source and target models become sufficiently dissimilar. Conversely, when h is small to moderate, the transfer learning bound is tighter, which supports

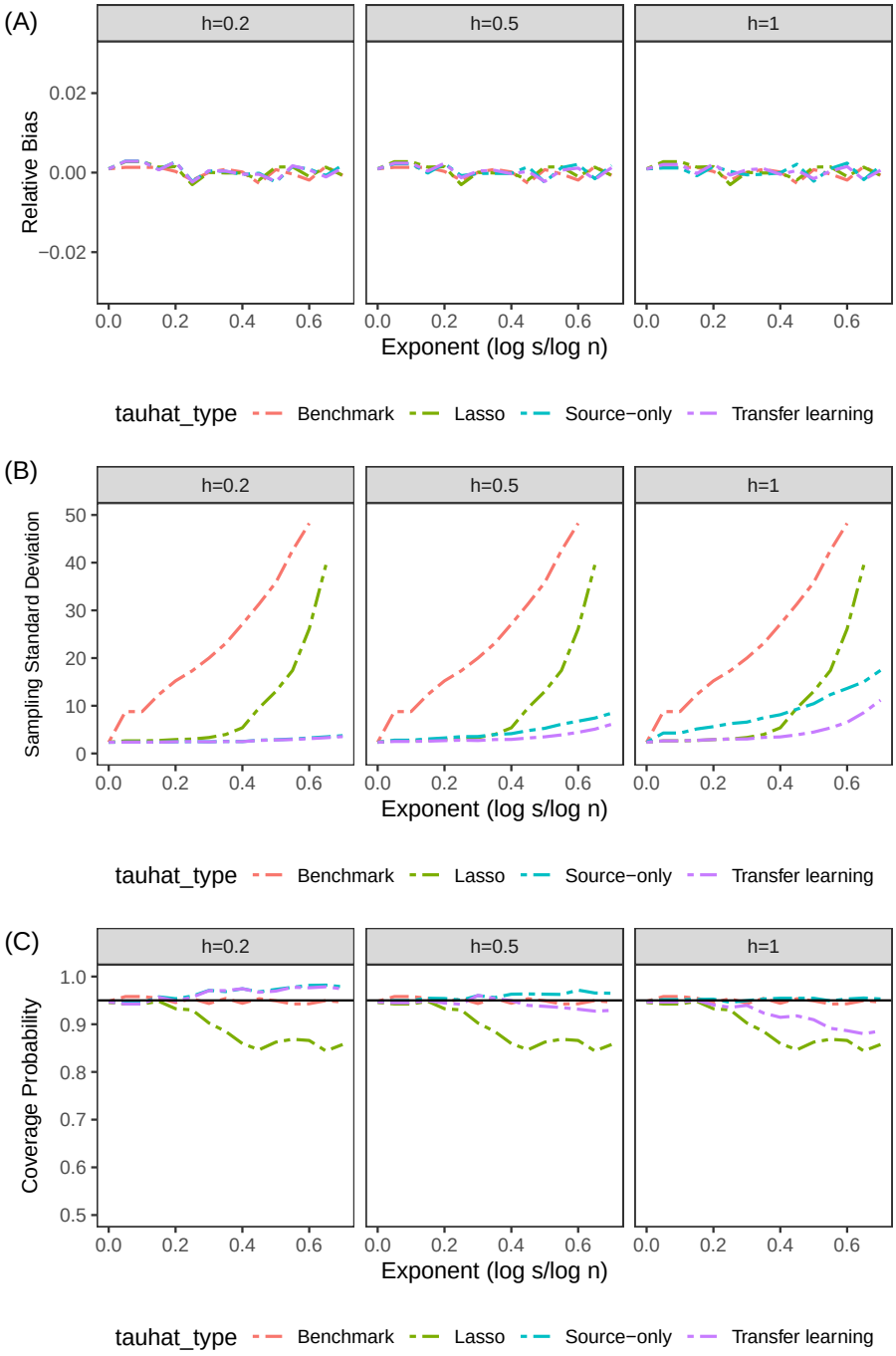

Figure 5: Simulations in Case 1 under stratified block randomization with unequal allocation ratio. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

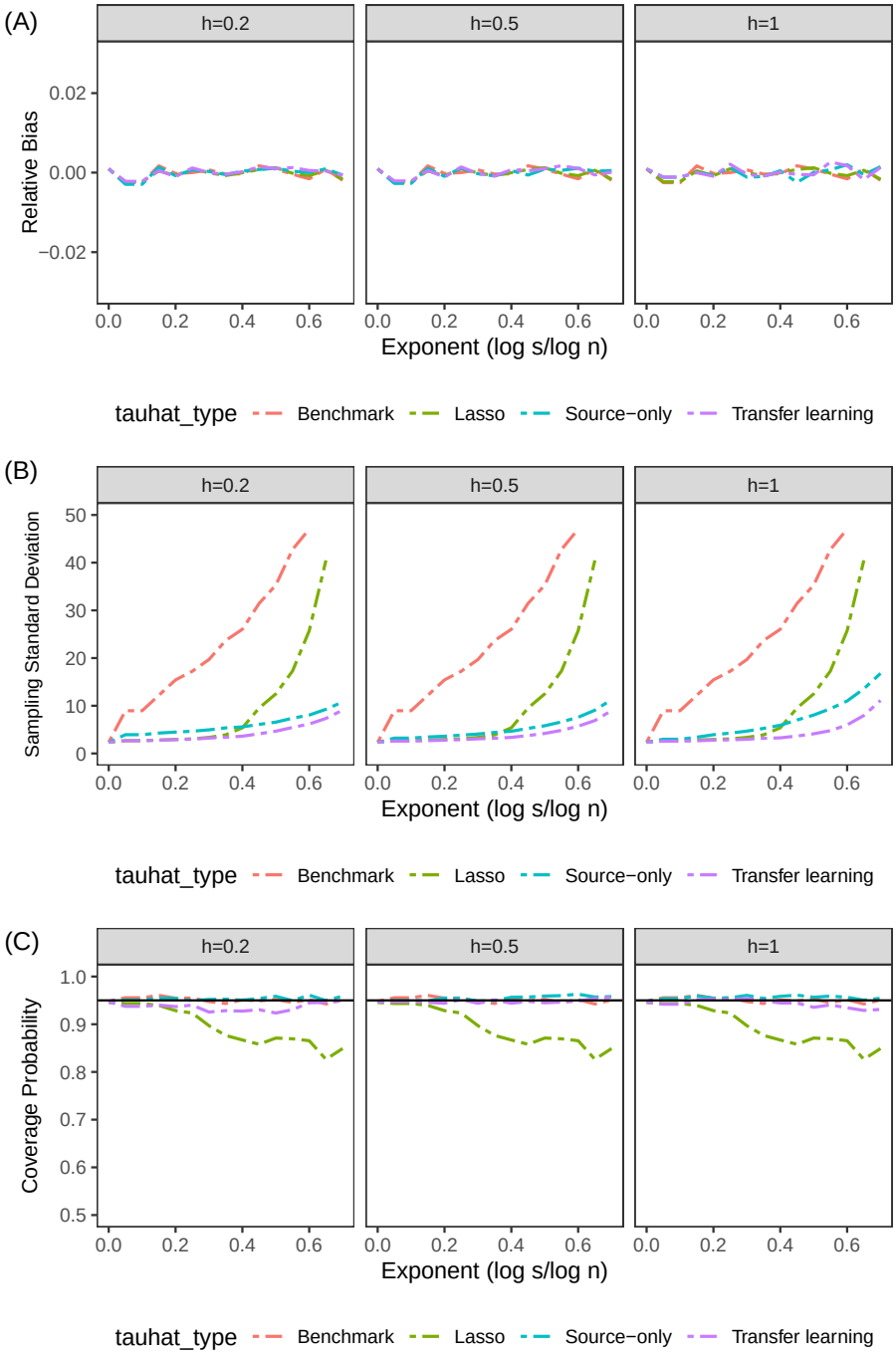

Figure 6: Simulations in Case 2 under stratified block randomization with unequal allocation ratio. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

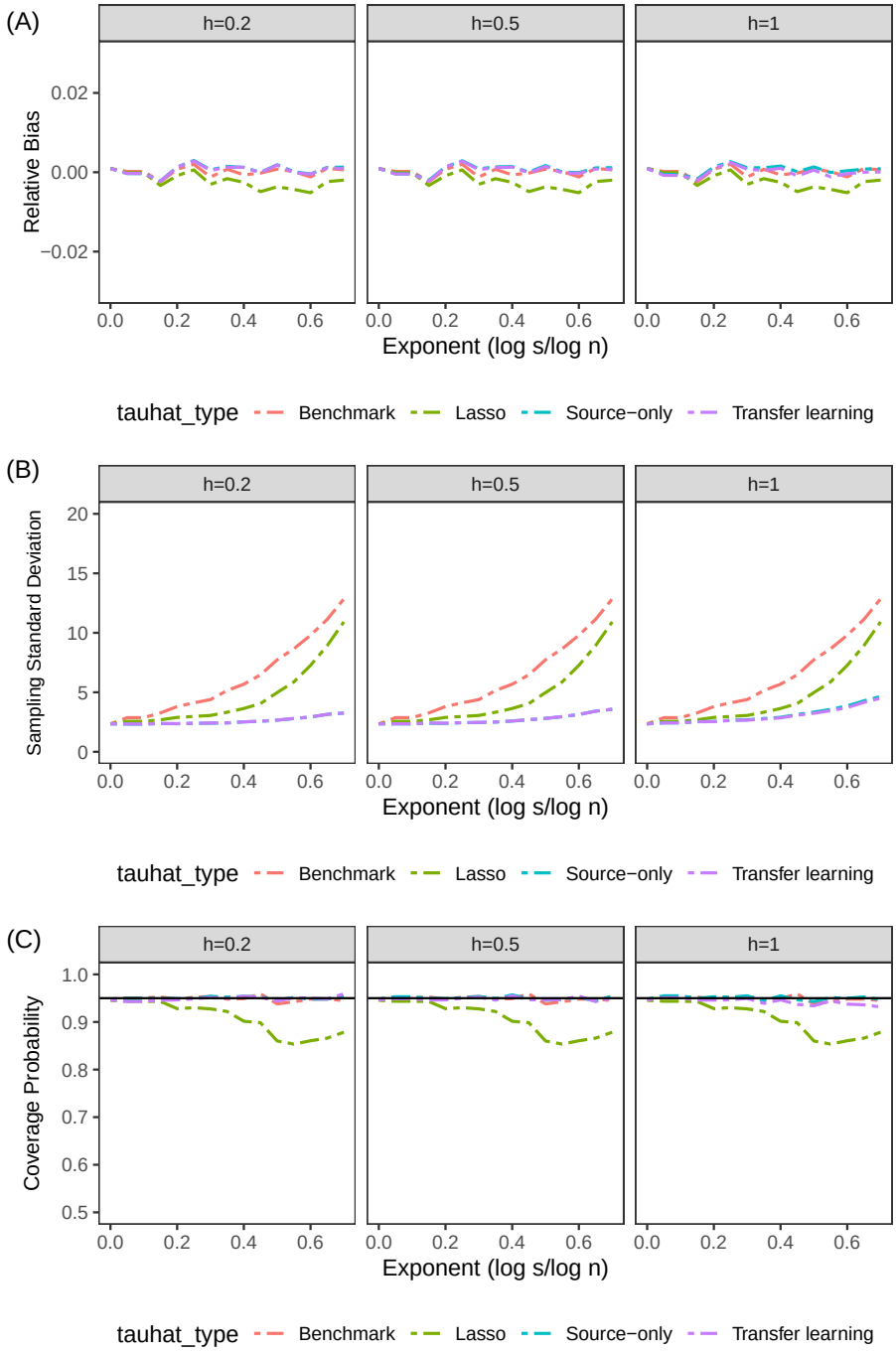

Figure 7: Simulations in Case 3 under stratified block randomization with unequal allocation ratio. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

the conditions stated in the main text under which the bound is informative. In the next section, we further investigate whether this theoretical behavior is reflected in finite-sample simulations of estimation error.

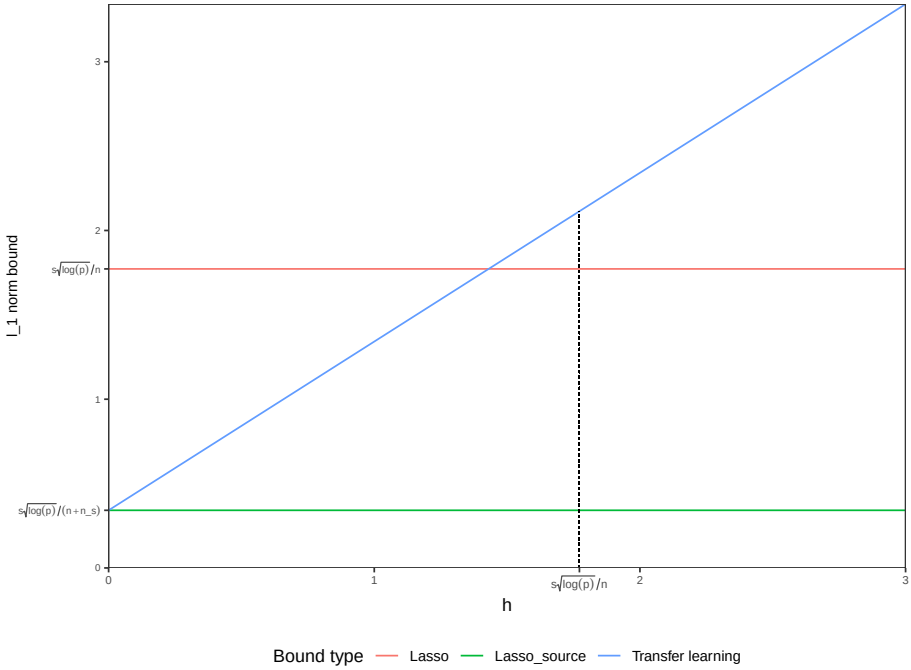

Figure 8: Theoretical l_1 norm error bounds in Proposition 6 and Theorem 15. In this simulation, target sample size $n = 50$, source sample size $n^{(s)} = 2000$, $M_n = \log\log n$, $s = 5$ and $p = 100$.

D.2.2 Simulation

In this section, we conduct additional simulation studies to examine how the theoretical behavior described in Theorem 16 is reflected in finite-sample performance. In particular, we investigate how the relative improvement of the proposed transfer learning estimator over applying the lasso to the target data alone varies with the discrepancy h.

We consider a stratified block randomization with three treatment arms. Within each stratum, units are assigned with an allocation ratio of $1:1:1$ and a block size of 6. Outcomes in both the target and source datasets are generated from the following linear model:

$$Y_i(a) = \tau_a + X_{i1}\beta_0 + X_{i1}X_{i2}^\top\beta + \epsilon_i(a).$$

Here, τ_a denotes the treatment effect. $X_{i1} \in \{1, 2\}$ is a stratum variable with $\Pr(X_{i1} = 1) = 0.4$, and $X_{i2} \in \mathbb{R}^s$ is a covariate vector with independent entries, each drawn from a Uniform$[-2, 2]$ distribution, where $s \in \{2, 4, 6\}$ in our settings. To consider a high-dimensional setting, in both the lasso and transfer learning procedures, additional covariates beyond X_{i2} are generated to reach a total dimension of $p = 100$. These additional covariates

are independent of X_{i1} and X_{i2}, and are drawn from a multivariate normal distribution with mean zero and a diagonal covariance matrix, where all diagonal elements are equal to 2. The error term $\epsilon_i(a) \sim \mathcal{N}(0, \sigma_\beta^2)$, where σ_β^2 is chosen to ensure a signal-to-noise ratio of SNR = 2. The coefficient $\beta_0 = 1$ is fixed and shared across the target and source models.

In the target data, the sample size is $n = 300$, and the coefficient vector β is s-dimensional with each element equal to $2/s$, so that $\|\beta\|_1 = 2$ for all values of s. The source data have the same sample size $n_{\text{source}} = 1200$, using coefficient vector β_{source}, whose elements are all equal to $(2 + h)/s$. This construction ensures $\|\beta_{\text{source}} - \beta\|_1 = h$, where h ranges from 0 to 3 and directly controls the degree of heterogeneity between the source and target models.

As we derive $\hat{\beta}_{[k]\text{lasso}}(a)$ and $\hat{\beta}_{[k]\text{tl}}(a)$ separately within each stratum and treatment group, we only report results for stratum $X_{i1} = 1$ and treatment $a = 0$, since the patterns are consistent across strata and treatment arms. We evaluate the ratio of ℓ_1-norm error based on $R = 1000$ replicates,

$$\frac{\|\hat{\beta}_{[1],*}(0) - \beta\|_1}{\|\hat{\beta}_{[1]\text{lasso}}(0) - \beta\|_1},$$

where $\hat{\beta}_{[1],*}(0)$ denotes either the lasso vector $\hat{\beta}_{[1],\text{lasso}}(0)$ or the proposed transfer learning vector $\hat{\beta}_{[1]\text{tl}}(0)$. The case $h = 0$, corresponding to ideally identical source and target models, is included as an additional benchmark. The simulation results are presented in Fig 9.

The simulation results reveal three main patterns. First, for a fixed s, as h increases, the ratio of l_1-norm errors for the transfer learning vector gradually increases: it starts below 1, indicating that transfer learning outperforms the lasso on the target data alone, and eventually exceeds 1, showing that the advantage diminishes when the source and target data become too different. Second, the value of h at which the ratio equals 1 increases with s. This indicates that in less sparse settings, transfer learning remains beneficial even when the source and target models differ more substantially. Third, for $h = 0$, the baseline ratio decreases as s increases, indicating that transfer learning performs increasingly better relative to the lasso when the signal vector becomes less sparse, as source and target models are identical.

D.3 Additional illustration of valid inference against negative transfer

In this section, we conduct a simulation study to demonstrate that our transfer learning method provides valid statistical inference and is robust against negative transfer, particularly when the l_1-norm error for $\hat{\beta}_{[k]\text{tl}}(a)$ becomes worse relative to $\hat{\beta}_{[k]\text{lasso}}(a)$ as the difference h grows. Following the experimental setup described in Appendix D.2.2, we evaluate the performance using relative bias, sampling standard deviation, and empirical coverage probability across $R = 2000$ independent replicates. The results are reported in Figure 10. Although the l_1-norm estimation error worsens with larger h (as shown in Figure 9), our inference method maintains robust performance. In Figure 10(B), the proposed transfer learning estimator $\hat{\tau}_{\text{tl}}$ consistently exhibits a smaller or comparable sampling standard deviation compared to the lasso-adjusted estimator $\hat{\tau}_{\text{lasso}}$ and the source-only estimator. As the sparsity s increases, the variance reduction achieved by our transfer learning method becomes even more pronounced, while the benchmark estimator consistently exhibits the

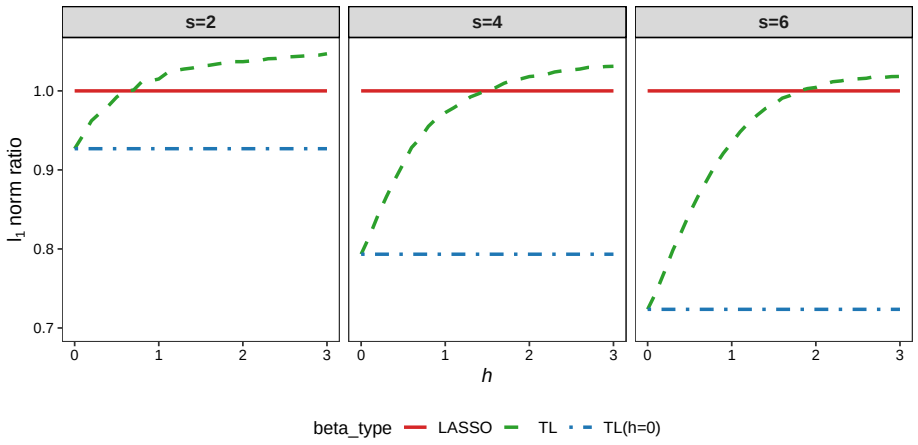

Figure 9: l_1-norm ratios of different estimators for $\beta_{[k]\text{proj}}(a)$. Each panel corresponds to a different sparsity level s.

highest variance. In Figure 10 (C), When s is small ($s = 2, 4$), the coverage probability of $\hat{\tau}_{\text{tl}}$ closely targets the nominal 95% level, demonstrating its resilience to negative transfer. When s increases to 6, $\hat{\tau}_{\text{tl}}$ becomes slightly conservative (with coverage hovering around 95%–96%), whereas the lasso-adjusted estimator $\hat{\tau}_{\text{lasso}}$ undercovers significantly, dropping to approximately 92%.

D.4 Additional illustration of the restricted strong convexity condition

To empirically validate the restricted strong convexity (RSC) condition in high-dimensional settings, we conduct a simulation experiment to investigate the lower bound of the empirical prediction norm $\|X^\top \beta\|_2/\sqrt{n}$. We consider a fixed sample size $n = 400$ and a covariate dimension $p = 1000$. The design matrix $X \in \mathbb{R}^{p\times n}$ is generated from a Beta distribution, where each entry X_{ij} is independently sampled from $\text{Beta}(\alpha, \alpha)$ with $\alpha = 0.1$, and subsequently centered and scaled to have zero mean and unit variance. Since the Beta distribution has bounded support on $[0, 1]$, the centered and scaled covariates remain uniformly bounded, thereby satisfying the uniform boundedness condition in Assumption 1. To explore the impact of sparsity, we vary the support size $s = \|\beta\|_0$ across 30 logarithmically spaced values up to a maximum sparsity of $s_{\max} = 120$. For each sparsity level s, we perform 100 Monte Carlo replicates. In each replicate, a sparse coefficient vector $\beta \in \mathbb{R}^p$ is constructed by randomly selecting s indices with independent standard normal entries ($\beta_i \sim \text{N}(0, 1)$), followed by an l_2-normalization to enforce $\|\beta\|_2 = 1$.

We follow the relationship between the empirical prediction norm $\|X^\top \beta\|_2/\sqrt{n}$ and the scaled norm $\sqrt{\log p/n}\|\beta\|_1$. As illustrated in Figure 11, the blue dots represent the individual results collected in all 3,000 independent simulation runs. To formalize the empirical lower bound, we extract the minimum value of $\|X^\top \beta\|_2/\sqrt{n}$ within each sparsity level and fit a linear envelope using ordinary least squares. The resulting empirical RSC

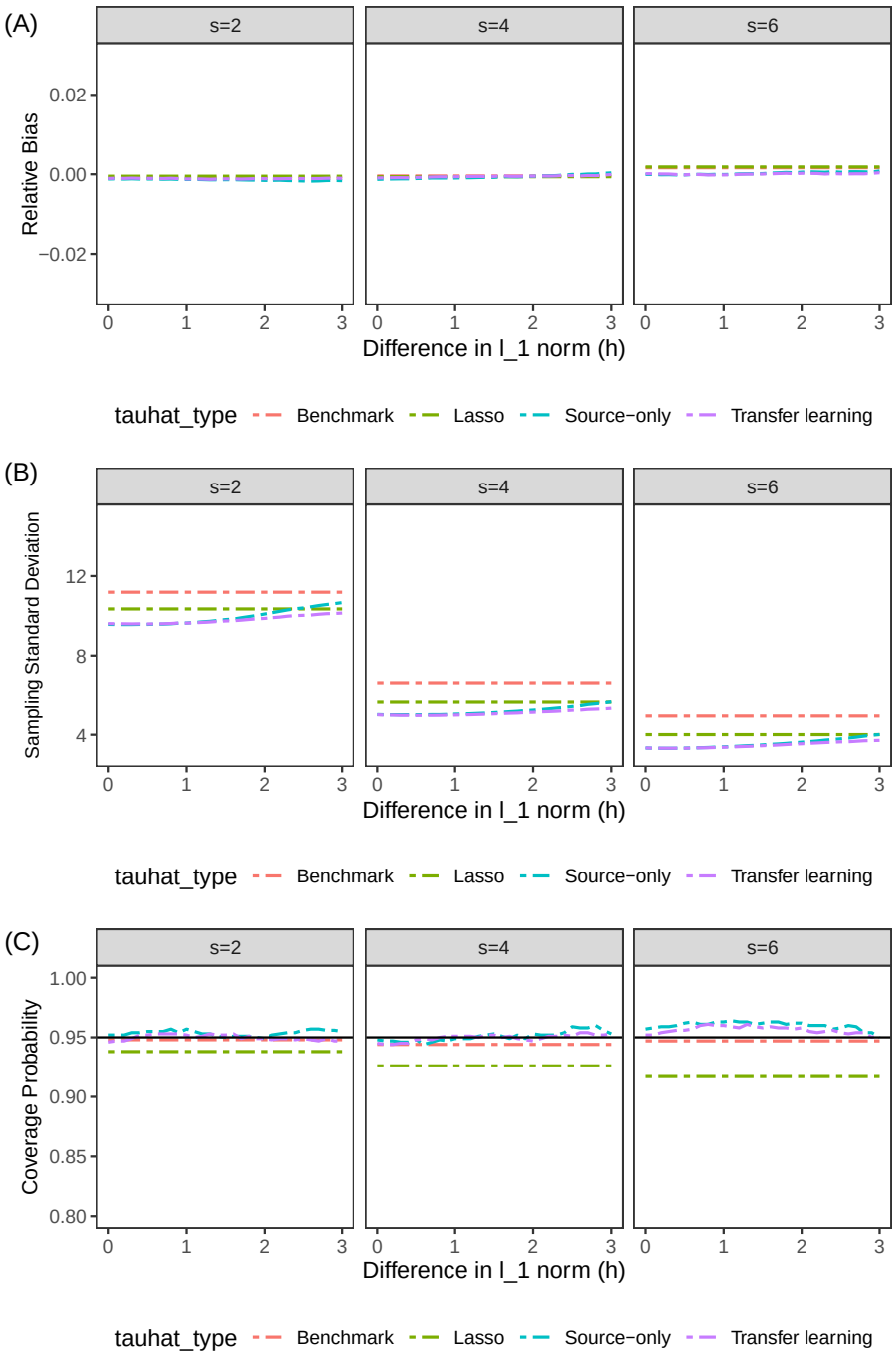

Figure 10: Simulations under stratified block randomization. Each column represents different s in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$, $\hat{\tau}_{\text{so}}$ and $\hat{\tau}_{\text{tl}}$.

lower bound is depicted by the orange dashed line, which satisfies:

$$\frac{\|X^\top \beta\|_2}{\sqrt{n}} \geq \kappa_1 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\beta\|_1.$$

The simulation results verify that the RSC condition holds with high probability under our experimental configuration, as the empirical lower bound remains well above zero throughout the entire range of sparsity.

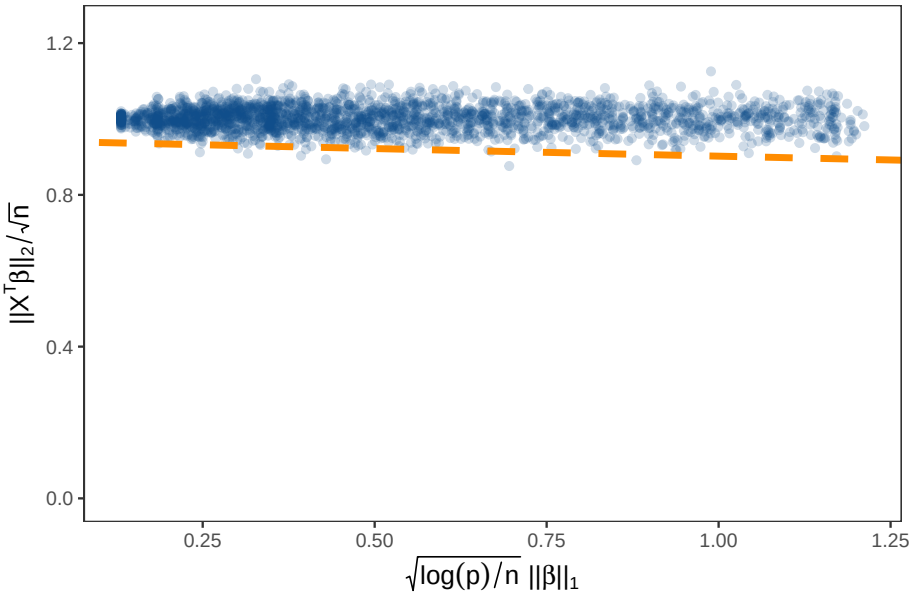

Figure 11: Empirical verification of the restricted strong convexity condition under bounded covariates and positive definite covariance matrix. The blue dots represent individual simulation trials across varying sparsities, while the dashed orange line denotes the fitted empirical lower bound.

D.5 Numerical illustration of the transfer learning framework under response-adaptive randomization

In this section, we investigate the performance of the transfer learning framework for the lasso-adjusted estimator under randomizations other than covariate-adaptive randomization. Specifically, we consider the doubly adaptive biased coin design (DBCD), a response-adaptive randomization procedure that sequentially adjusts treatment allocation probabilities based on both accumulating efficacy data and current assignment proportions. By incorporating observed outcomes into the randomization procedure, DBCD violates Assumption 2 in the main text. Consequently, the l_1-norm error and the asymptotic normality derived in Theorems 15 and 16 may no longer hold, and the variance estimator in Theorem 21 may become inconsistent. Therefore, we conduct numerical simulations in this section to explore whether the transfer learning estimator $\hat{\tau}_{\text{tl}}$ still maintains an advantage over the

lasso-adjusted estimator $\hat{\tau}_{\text{lasso}}$ under DBCD. For the variance estimation of $\hat{\tau}_{\text{lasso}}$, we adopt the approach proposed in Tu and Ma (2025). Meanwhile, the variance estimator for the transfer learning estimator is constructed by replacing the lasso coefficient vector $\hat{\beta}_{\text{lasso}}(a)$ in the variance estimator of $\hat{\tau}_{\text{lasso}}$ with $\hat{\beta}_{\text{tl}}(a)$.

We now introduce the details of the simulation settings. The following model is used to generate the potential outcomes for both the source and target data:

$$Y_i(a) = \mu_a + X_i^\top \beta + \epsilon_{a,i}, \quad \text{for } a \in \mathcal{A}_0.$$

We consider an s-dimensional covariate vector X_i, where each component X_{ij} is independently generated from Unif$[0, 2]$. We focus on the treatment setting with $|\mathcal{A}_0| = 2$, representing one treatment group and one control group. For each model, $(X_i, \epsilon_{0,i}, \epsilon_{1,i})$ are independent and identically distributed (i.i.d.), and the error term $\epsilon_{a,i}$ follows a t-distribution with 3 degrees of freedom, scaled to have mean zero and unit variance for $a = 0, 1$. To mimic a high-dimensional setting, additional noise covariates are also generated from Unif$[0, 2]$ to bring the total dimension to $p = 100$. We examine the results as the difference between the source and target data increases. Let $s = \lceil n^r \rceil$ with $r \in \{0, 0.05, \ldots, 0.7\}$. In the target data, we set $\beta_j = 1$ for $j = 1, 2, \ldots, s$. In the source data, the coefficients are specified as $\beta_j^{(s)} = 1 + \frac{hj}{s}$ for $j = 1, 2, \ldots, s$, where $h \in \{0.2, 0.5, 1\}$. This setup results in the l_1 norm difference of $\|\beta - \beta^{(s)}\|_1 = h(s + 1)/2$.

We present the simulation results under DBCD for three treatment effect estimators: the benchmark difference-in-means estimator $\hat{\tau}_{\text{ben}}$, the lasso-adjusted estimator $\hat{\tau}_{\text{lasso}}$ from Tu and Ma (2025), and the proposed transfer learning estimator $\hat{\tau}_{\text{tl}}$. The sample sizes for the target and source data are set to $n = 100$ and $n^{(s)} = 400$, respectively. The target allocation function for DBCD is adopted from Zhang and Rosenberger (2006). All estimators are evaluated over $R = 2000$ independent replicates. Performance is evaluated using relative bias, standard deviation, and the coverage probability of the 95% confidence interval, following the same definitions provided in Section 6. The simulation results are summarized in Figure 12. Overall, the numerical patterns remain highly consistent with our previous findings in Section 6. Specifically, the proposed transfer learning estimator $\hat{\tau}_{\text{tl}}$ consistently outperforms both $\hat{\tau}_{\text{ben}}$ and $\hat{\tau}_{\text{lasso}}$ in terms of estimation efficiency, with coverage probabilities remaining close to the nominal 95% level in most settings as the sparsity parameter s increases. Although $\hat{\tau}_{\text{tl}}$ exhibits a slight undercoverage when facing a large source-target discrepancy combined with a dense model setting (i.e., $h = 1$ and a large s), its coverage probability decreases at a visibly slower rate compared to that of $\hat{\tau}_{\text{lasso}}$, demonstrating its robustness under response-adaptive randomization.

Appendix E. Real data results

E.1 Treatment effect estimation and inference

In this section, we apply the proposed transfer learning algorithm to the CBASP dataset analyzed in Section 7, without synthetically generating any unobserved outcomes. Specifically, the target and source populations are divided by the clinical status of patients, with HAMD ≤ 17 and HAMD ≥ 18 representing the target and source population, respectively. The additional covariates are constructed following the exact same procedure as described

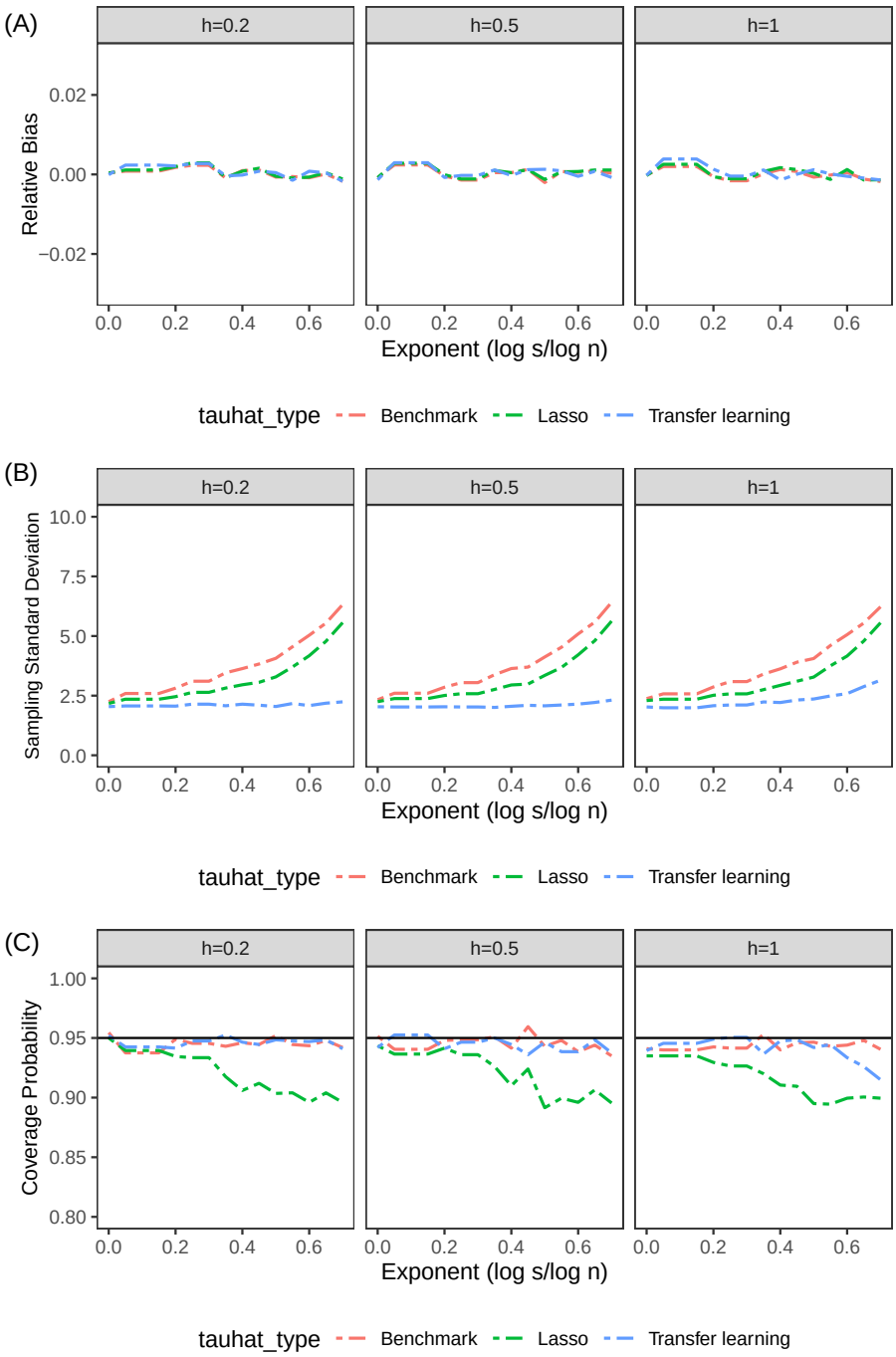

Figure 12: Simulations under the response-adaptive randomization. Each column represents different h in the simulation settings, resulting in different l_1 norms of the bias. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{lasso}}$ and $\hat{\tau}_{\text{tl}}$.

in Section 7. As summarized in Table 4, the results of the real-data analysis remain highly consistent with our previous findings presented in Table 2.

Table 4: Treatment effect estimates, 95% confidence intervals and variance reductions for nefazodone CBASP trial data.

	τ_1			τ_2		
Estimator	Estimate	95%CI	VR(%)	Estimate	95%CI	VR(%)
$\hat{\tau}_{\text{ben}}$	3.83	(0.76, 6.91)	–	4.28	(1.12, 7.43)	–
$\hat{\tau}_{\text{so}}$	3.89	(0.82, 6.95)	0.84	4.54	(1.34, 7.74)	-2.82
$\hat{\tau}_{\text{lasso}}$	3.79	(0.76, 6.83)	2.78	4.31	(1.22, 7.40)	4.08
$\hat{\tau}_{\text{tl}}$	4.39	(1.43, 7.35)	7.59	4.86	(1.84, 7.89)	8.10

Note: τ_1, the treatment effect of nefazodone versus combination therapy; τ_2, the treatment effect of CBASP versus combination therapy; Estimates: treatment effect estimates; CI, confidence interval; VR, variance reduction.

E.2 Interpretability study for lasso adjustment

In this section, we conduct an exploratory interpretability study to show that the lasso adjustment, especially when enhanced by our transfer learning method, can select relevant variables to improve efficiency. We use the CBASP data in the same way as in the last section and form $p = 135$ additional covariates as well. Here, we examine the selection frequency and the average absolute coefficient estimates for the eight baseline covariates in Table 1. In detail, the selection frequency is calculated as the proportion of times a baseline covariate group is selected across the six projection vectors (representing three treatment arms and two strata). In this calculation, a baseline group is counted as selected if at least one of its associated expanded covariates has a non-zero coefficient. The coefficient is computed by taking the average of the absolute projection coefficients across all expanded covariates derived from the same baseline variable. Finally, the summary statistics presented in Table 5 are obtained by averaging these metrics over 200 replicates.

Table 5 summarizes the variable selection behavior and the estimated projection coefficients of the proposed transfer learning method. As shown in the table, AGE (0.57), Mstatus2 (0.51), and HAMD17 (0.50) have the highest selection frequencies across the projection vectors. In terms of the coefficients, HAMD17 and TreatPD have the largest average absolute coefficients (3.1 and 2.6, respectively, on the 10^{-2} scale). These results demonstrate that our method can select relevant baseline variables, which lead to efficiency gain as shown in Table 4.

Table 5: Variable selection frequencies and absolute estimated projection coefficients using the proposed transfer learning method.

Variable	Selection frequency	Absolute coefficients ($\times 10^{-2}$)
Mstatus2	0.51	1.9
TreatPD	0.45	2.6
AGE	0.57	1.8
HAMD17	0.50	3.1
HAMD24	0.37	1.6
HAMD_COGNID	0.48	1.6
NDE	0.38	1.2
HAMA_SOMATI	0.42	1.9

References

Susan Athey, Raj Chetty, and Guido Imbens. Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676, 2020.

Anthony C. Atkinson. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika, 69:61–67, 1982.

Yuehao Bai, Liang Jiang, Joseph P Romano, Azeem M Shaikh, and Yichong Zhang. Covariate adjustment in experiments with matched pairs. Journal of Econometrics, 241(1): 105740, 2024.

Marlena S Bannick, Jun Shao, Jingyi Liu, Yu Du, Yanyao Yi, and Ting Ye. A general form of covariate adjustment in clinical trials under covariate-adaptive randomization. Biometrika, in press, 2025.

Colin B. Begg and Boris Iglewicz. A treatment allocation procedure for sequential clinical trials. Biometrics, 36(1):81–90, 1980.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics, 37(4):1705–1732, 2009.

Jeffrey A Boatman, David M Vock, and Joseph S Koopmeiners. Borrowing from supplemental sources to estimate causal effects from a primary data source. Statistics in Medicine, 40(24):5115–5130, 2021.

Federico A Bugni, Ivan A Canay, and Azeem M Shaikh. Inference under covariate-adaptive randomization. Journal of the American Statistical Association, 113(524):1784–1796, 2018.

Federico A Bugni, Ivan A Canay, and Azeem M Shaikh. Inference under covariate-adaptive randomization with multiple treatments. Quantitative Economics, 10(4):1747–1785, 2019.

Peter Bühlmann and Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin, 2011.

Changxiao Cai, T Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. The Annals of Statistics, 52(1):207–232, 2024.

T Tony Cai and Hongming Pu. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. arXiv preprint arXiv:2401.12272, 2024.

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. The Annals of Statistics, 49(1):100–128, 2021.

Sylvie Chevret, Jean-François Timsit, and Lucie Biard. Challenges of using external data in clinical trials- an illustration in patients with COVID-19. BMC Medical Research Methodology, 22(1):321, 2022.

Jianing Chu, Wenbin Lu, and Shu Yang. Targeted optimal treatment regime learning using summary statistics. Biometrika, 110(4):913–931, 2023.

Jody D Ciolino, Hannah L Palac, Amy Yang, Mireya Vaca, and Hayley M Belli. Ideal vs. real: A systematic review on handling covariates in randomized controlled trials. BMC Medical Research Methodology, 19:136–146, 2019.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: A review. Statistical Science, 39(1):165–191, 2024.

Issa J Dahabreh and Kirsten Bibbins-Domingo. Causal inference about the effects of interventions from observational studies in medical journals. Journal of the American Medical Association, 331(21):1845–1853, 2024.

Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. Statistics in Medicine, 39(14):1999–2014, 2020.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In International Conference on Machine Learning, pages 647–655. PMLR, 2014.

Bradley Efron. Forcing a sequential experiment to be balanced. Biometrika, 58(3):403–417, 1971.

U.S. Food and Drug Administration. Rare diseases: Natural history studies for drug development: Draft guidance for industry, 2022. URL `https://www.fda.gov/media/122425/download`.

Yujia Gu, Hanzhong Liu, and Wei Ma. Regression-based multiple treatment effect estimation under covariate-adaptive randomization. Biometrics, 79(4):2869–2880, 2023.

Brian P Hobbs, Bradley P Carlin, Sumithra J Mandrekar, and Daniel J Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics, 67(3):1047–1056, 2011.

Feifang Hu, Xiaoqing Ye, and Li-Xin Zhang. Multi-arm covariate-adaptive randomization. Science China Mathematics, 66(1):163–190, 2023.

Yanqing Hu and Feifang Hu. Asymptotic properties of covariate-adaptive randomization. The Annals of Statistics, 40(3):1794–1815, 2012.

Francis L Huang. Using instrumental variable estimation to evaluate randomized experiments with imperfect compliance. Practical Assessment, Research & Evaluation, 23(2): n2, 2018.

Jian Huang, Joel L. Horowitz, and Shuangge Ma. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica, 19(4):1603–1618, 2009.

Jiayu Huang, Mingqiu Wang, and Yuanshan Wu. Estimation and inference for transfer learning with high-dimensional quantile regression. arXiv preprint arXiv:2211.14578, 2022.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. Statistical Science, 15:46–60, 2000.

Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. Statistics in Medicine, 34:3724–3749, 2015.

Jun Jin, Jun Yan, Robert H Aseltine, and Kun Chen. Transfer learning with large-scale quantile regression. Technometrics, pages 1–30, 2024.

Yonghan Jung and Alexis Bellot. Efficient policy evaluation across multiple different experimental datasets. Advances in Neural Information Processing Systems, 37:136361–136392, 2024.

Yonghan Jung, Iván Díaz, Jin Tian, and Elias Bareinboim. Estimating causal effects identifiable from a combination of observations and experiments. Advances in Neural Information Processing Systems, 36:46446–46490, 2023.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. Advances in Neural Information Processing Systems, 31, 2018.

Martin B Keller, James P McCullough, Daniel N Klein, Bruce Arnow, David L Dunner, Alan J Gelenberg, John C Markowitz, Charles B Nemeroff, James M Russell, Michael E Thase, et al. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. New England Journal of Medicine, 342(20):1462–1470, 2000.

Dasom Lee, Shu Yang, Lin Dong, Xiaofei Wang, Donglin Zeng, and Jianwen Cai. Improving trial generalizability using observational studies. Biometrics, 79(2):1213–1225, 2023.

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

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning in large-scale gaussian graphical models with false discovery rate control. Journal of the American Statistical Association, 118(543):2171–2183, 2023.

Sai Li, Linjun Zhang, T Tony Cai, and Hongzhe Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. Journal of the American Statistical Association, 119(546):1274–1285, 2024.

Hanzhong Liu, Fuyi Tu, and Wei Ma. Lasso-adjusted treatment effect estimation under covariate-adaptive randomization. Biometrika, 110(2):431–447, 2023.

Jing Liu, Jeffrey S Barrett, Efthimia T Leonardi, Lucy Lee, Satrajit Roychoudhury, Yong Chen, and Panayiota Trifillis. Natural history and real-world data in rare diseases: applications, limitations, and future perspectives. The Journal of Clinical Pharmacology, 62:S38–S55, 2022.

Wei Ma, Feifang Hu, and Lixin Zhang. Testing hypotheses of covariate-adaptive randomized clinical trials. Journal of the American Statistical Association, 110(510):669–680, 2015.

Wei Ma, Fuyi Tu, and Hanzhong Liu. Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective. Statistics in Medicine, 41(29): 5645–5661, 2022a.

Wei Ma, Mengxi Wang, and Hongjian Zhu. Seamless phase II/III clinical trials with covariate adaptive randomization. Statistica Sinica, 32(2):1079–1098, 2022b.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. The Annals of Statistics, 37(1):246–270, 2009.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Advances in Neural Information Processing Systems, 22, 2009.

Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. Clinical Trials, 7(1):5–18, 2010.

J. S. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. Annals of Agricultural Sciences, 10:1–51, 1923.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2009.

Raphael Petegrosso, Sunho Park, Tae Hyun Hwang, and Rui Kuang. Transfer learning across ontologies for phenome–genome association prediction. Bioinformatics, 33(4):529–536, 2017.

Stuart J Pocock. The combination of randomized and historical controls in clinical trials. Journal of Chronic Diseases, 29(3):175–188, 1976.

Stuart J. Pocock and Richard Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics, 31(1):103–115, 1975.

Ahnaf Rafi. Efficient semiparametric estimation of average treatment effects under covariate adaptive randomization. arXiv preprint arXiv:2305.08340, 2023.

William F Rosenberger and John M Lachin. Randomization in Clinical Trials: Theory and Practice. John Wiley & Sons, 2nd edition, 2015.

Donald Rubin. Estimating causal effects of treatments in experimental and observational studies. Ets Research Bulletin Series, 1972(2):i–31, 1972.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688, 1974.

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, pages 15–18, 2019.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In European Conference on Computer Vision, pages 213–226. Springer, 2010.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics, 70(4):1023–1032, 2014.

Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, and Bernhard Schölkopf. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. Advances in Neural Information Processing Systems, 21, 2008.

Jun Shao, Xinxin Yu, and Bob Zhong. A theory for testing hypotheses under covariate-adaptive randomization. Biometrika, 97(2):347–360, 2010.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society Series A, 174(2):369–386, 2011.

Jeremy B Sussman and Rodney A Hayward. An IV for the RCT: Using instrumental variables to adjust for treatment contamination in randomised controlled trials. British Medical Journal, 340, 2010.

Donald R. Taves. Minimization: A new method of assigning patients to treatment and control groups. Clinical Pharmacology & Therapeutics, 15(5):443–453, 1974.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544):2684–2697, 2023.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58(1):267–288, 1996.

Lisa Torrey and Jude Shavlik. Transfer learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pages 242–264. IGI global, 2010.

Fuyi Tu and Wei Ma. Covariate-adjusted inference for doubly adaptive biased coin design. Statistical Methods in Medical Research, 34(9):1795–1820, 2025.

Fuyi Tu, Wei Ma, and Hanzhong Liu. A unified framework for covariate adjustment under stratified randomisation. Stat, 13(4):e70016, 2024.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics, 42(3):1166–1202, 2014.

Bingkai Wang, Ryoko Susukida, Ramin Mojtabai, Masoumeh Amin-Esmaeili, and Michael Rosenblum. Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. Journal of the American Statistical Association, 118(542):1152–1163, 2023.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. Journal of Big Data, 3(1):1–40, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, 2020.

Lili Wu and Shu Yang. Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In Conference on Causal Learning and Reasoning, pages 904–926. PMLR, 2022.

Lili Wu and Shu Yang. Transfer learning of individualized treatment rules from experimental to real-world data. Journal of Computational and Graphical Statistics, 32(3):1036–1045, 2023.

Fan Yang, Hongyang R Zhang, Sen Wu, Christopher Ré, and Weijie J Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. Journal of Machine Learning Research, 26(113):1–88, 2025.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. Journal of the Royal Statistical Society Series B, 85(3):575–596, 2023.

Ting Ye, Yanyao Yi, and Jun Shao. Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. Biometrika, 109:33–47, 2022.

Ting Ye, Jun Shao, Yanyao Yi, and Qingyuan Zhao. Toward better practice of covariate adjustment in analyzing randomized clinical trials. Journal of the American Statistical Association, 118(544):2370–2382, 2023.

Marvin Zelen. The randomization and stratification of patients to clinical trials. Journal of Chronic Diseases, 27(7):365–375, 1974.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36(4):1567–1594, 2008.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society Series B, 76 (1):217–242, 2014.

Lanju Zhang and William F Rosenberger. Response-adaptive randomization for clinical trials with continuous outcomes. Biometrics, 62(2):562–569, 2006.

Ke Zhu, Hanzhong Liu, and Yuehan Yang. Design-based theory for lasso adjustment in randomized block experiments and rerandomized experiments. Journal of Business & Economic Statistics, 43(3):544–555, 2025.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1):43–76, 2020.