Statistica Sinica Preprint No: SS-2025-0480	
Title	Transfer Learning for Covariate-Adaptive Randomized Trials with High-Dimensional Generalized Linear Models
Manuscript ID	SS-2025-0480
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0480
Complete List of Authors	Yujia Gu, Hanzhong Liu and Wei Ma
Corresponding Authors	Wei Ma
E-mails	mawei@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Transfer learning for covariate-adaptive randomized trials with high-dimensional generalized linear models

Yujia Gu[1], Hanzhong Liu[1], Wei Ma[2]

[1]*Tsinghua University and* [2]*Renmin University of China*

Abstract: Covariate-adaptive randomization (CAR) is widely adopted in clinical trials to balance baseline covariates across treatment arms. When outcomes are binary or other types of discrete data, generalized linear models (GLMs) are typically used as working models to improve the efficiency of treatment effect estimation. However, in high-dimensional settings, limited sample sizes may lead to unstable estimation and invalid inference. Moreover, when the underlying data-generating mechanism is misspecified relative to the GLM working model, efficiency gains over the unadjusted estimator are no longer guaranteed. To address these challenges, we first propose a high-dimensional GLM adjustment approach using ℓ_1 regularization, resulting in a GLM-adjusted treatment effect estimator under CAR. When the target trial has a limited sample size, we further develop a transfer learning estimator that incorporates information from a large external trial regarded as source data. As long as the source and target populations are sufficiently similar, the transfer learning estimator requires weaker sparsity conditions and a smaller target sample size than the GLM-adjusted estimator to achieve asymptotic normality. To guarantee efficiency improve-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

ment relative to the unadjusted estimator, we additionally introduce a projection method. With this step, both the GLM-adjusted and transfer learning estimators achieve efficiency gains. Simulation studies and a semi-synthetic data application demonstrate the effectiveness of our transfer learning strategy and the projection method.

Key words and phrases: covariate-adaptive randomization, generalized linear models, projection method, robustness, transfer learning.

1. Introduction

Covariate-adaptive randomization (CAR) is widely used in clinical trials to balance baseline covariates across treatment groups. Common CAR procedures include stratified block randomization (Zelen, 1974) and the minimization method (Taves, 1974; Pocock and Simon, 1975), which were used in approximately 80% of trials surveyed by Ciolino et al. (2019). In this paper, we focus on CAR methods that balance discrete covariates. For reviews and recent developments on other CAR methods, including those that balance continuous covariates, we refer interested readers to Rosenberger and Lachin (2015), Ma et al. (2024), and the references therein.

At the analysis stage, covariate adjustment can potentially improve efficiency by incorporating a working model relating the outcome to baseline covariates. However, concerns about valid inference arise, especially when

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

the treatment assignments have dependence induced by CAR. To address these concerns, Bugni et al. (2018, 2019) proposed robust treatment effect estimators adjusting for stratum variables, which can achieve valid inference. Here, "robust" means achieving valid inference even under model misspecification. Furthermore, linear adjustment for additional baseline covariates has been extensively studied and has been shown to provide valid inference and efficiency gains (Ma et al., 2022; Ye et al., 2023; Gu et al., 2023). In high-dimensional settings, Liu et al. (2023) applied lasso regression for covariate adjustment and proposed lasso-adjusted estimators.

For discrete outcomes, covariate adjustment using generalized linear models (GLMs) is often preferred over linear models, as it can better capture the nonlinear relationship between the outcome and covariates and potentially improve efficiency. However, related work under CAR remains limited. A few model-based approaches have been proposed to ensure valid inference in this setting (Li et al., 2021; Wang and Ma, 2025). Recently, Bannick et al. (2025) studied a GLM-based estimator that incorporated a specific calibration technique to achieve both robustness and efficiency gain under CAR.

Despite the favorable properties of existing GLM-based covariate adjustment methods, they can typically accommodate only a limited number

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

of baseline covariates. In modern clinical trials, high-dimensional covariates such as genomic or behavioral variables are increasingly available. Regularized estimation techniques such as the lasso (Van de Geer, 2008; Negahban et al., 2012) enable valid inference in sparse high-dimensional models. Yet when applied under CAR, limited sample sizes may still lead to invalid inference. Moreover, the sparsity assumption may not hold when the data-generating model is unknown.

A promising approach is to incorporate external trial data via transfer learning, which leverages information from related studies to improve estimation accuracy (Torrey and Shavlik, 2010). In statistics, transfer learning has been shown to accelerate convergence and improve inference in high-dimensional GLMs (Tian and Feng, 2023; Li et al., 2024). However, these results rely on correctly specified models and i.i.d. data, which may not hold under CAR. Recently, Gu et al. (2024) proposed a transfer learning estimator under CAR based on a high-dimensional linear working model, which requires weaker sparsity conditions and a smaller target sample size than the lasso-adjusted estimator of Liu et al. (2023). Their linear approach, as a special case of GLMs, motivates our extension to generalized linear models.

Main contributions: Our main objective is to address the challenge of high-dimensional covariate adjustment under CAR, where limited target

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

sample size and unknown sparsity patterns may hinder the performance of classical GLM methods. To this end, we develop a transfer learning framework that leverages external data to improve estimation and inference in the target population. We first study a GLM-adjusted estimator based on target data alone, serving as a baseline within our framework, and then introduce a projection-based refinement to further improve efficiency. A summary of the theoretical results is provided in Table 1 and our contributions are summarized as follows.

1. **A high-dimensional GLM-adjusted estimator.** In Section 3, we propose a GLM-adjusted treatment effect estimator using ℓ_1-penalized likelihood. We address the complex dependence induced by CAR, establishing the ℓ_1 convergence rate of the coefficient estimator and the asymptotic normality of the GLM-adjusted estimator (Proposition 1).

2. **A transfer learning framework for improved convergence and inference.** In Section 4, we develop a transfer learning approach with the dependent structure under the CAR framework. When source and target populations are sufficiently similar with respect to the population coefficient, the transfer learning estimator achieves a faster convergence rate (Theorem 1) and relaxes the sparsity condition from the target to the source sample size, enabling valid inference with

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

fewer target units (Theorem 2).

3. **A projection method for guaranteed efficiency gain.** In Section 5, to guarantee statistical efficiency gain, we develop a projection method that maps potential outcomes onto the linear space spanned by inverse link functions within treatment groups. In Corollary 1, we show that the projection refinement guarantees the efficiency gain for both the GLM-adjusted and transfer learning estimators.

4. **Practical implications.** Our results suggest that, when similar source data are available, the transfer learning estimator should be preferred over GLM-adjusted estimators, due to its improved convergence and relaxed sparsity requirements. The projection method can be further applied to guarantee the efficiency gain. In the absence of suitable source data, the GLM-adjusted estimator provides a valid and theoretically grounded option for inference under CAR.

2. Framework

We begin by introducing a CAR procedure for n units. Let $a \in \mathcal{A} = \{1, 2, \ldots, A\}$ denote the treatment groups, with $a = 0$ representing the control group. Define $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$ as the assignment set that includes all

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

treatment groups along with the control group. For each unit i, let A_i be the assigned treatment indicator, so that $I(A_i = a)$ indicates unit i is assigned to treatment a. The stratum variable B_i takes values in $\mathcal{K} = \{1, 2, \ldots, K\}$, where K is a fixed number of strata. Denote $p_{[k]} = \Pr(B_i = k) > 0$ as the population proportion of stratum k, and $\pi_{[k]a} = \Pr(A_i = a \mid B_i = k)$ as the target treatment allocation proportion within stratum k. Let $n_{[k]} = \sum_{i=1}^{n} I(B_i = k)$ and $n_{[k]a} = \sum_{i=1}^{n} I(A_i = a, B_i = k)$ be the observed counts of units in stratum k and in treatment group a within stratum k, respectively. The empirical proportion of stratum k is $p_{n[k]} = n_{[k]}/n$.

Each unit has p-dimensional baseline covariates $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top \in \mathbb{R}^p$, which can include both additional covariates not used in the randomization procedure and the original continuous covariates used to construct the strata. We consider a high-dimensional setting where the covariate dimension p increases with the sample size n. Moreover, we assume that the stratum-specific covariance matrix $\Sigma_{[k]XX} = E[\{X_i - E(X_i|B_i = k)\}\{X_i - E(X_i|B_i = k)\}^\top | B_i = k]$, $k = 1, 2, \ldots, K$, is positive definite.

We adopt the Neyman-Rubin model to define potential outcomes and treatment effects. Let $\{Y_i(a), a \in \mathcal{A}_0\}$ be the potential outcomes and $Y_i = \sum_{a \in \mathcal{A}_0} I(A_i = a) Y_i(a)$. Let $\{W_i\}_{i=1}^n = \{Y_i(0), \ldots, Y_i(A), B_i, X_i\}_{i=1}^n$ be i.i.d. samples from the population distribution $W = \{Y(0), \ldots, Y(A), B, X\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Our goal is to estimate the treatment effect $\tau_a = E\{Y_i(a) - Y_i(0)\}$ for all $a \in \mathcal{A}$ and to make inference for $\tau = (\tau_1, \tau_2, \ldots, \tau_A)^\top$.

We denote the set of random variables with at least one positive stratum-specific variance as $\mathcal{R}_2 = \{V : \max_{k\in\mathcal{K}} \text{var}(V|B_i = k) > 0\}$. The following assumptions are made for the data generation and CAR procedures.

Assumption 1. *$E\{Y_i^2(a)\} < \infty$ and $Y_i(a) \in \mathcal{R}_2$, for all $a \in \mathcal{A}_0$.*

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 $k \in \mathcal{K}$.*

Assumption 1 requires finite second moments and at least one positive stratum-specific variance for each potential outcome. Assumption 2 describes the CAR mechanism: (1) treatment assignments are independent of potential outcomes and covariates conditional on strata; (2) within-stratum allocation proportions converge to their targets. These assumptions are standard in the CAR literature (Bugni et al., 2019; Gu et al., 2023, 2024) and are satisfied by commonly used randomization procedures.

Next, we introduce some notation. Let $r_i(a)$ for $i = 1, 2, \ldots, n,\ a \in \mathcal{A}_0$ be a transformed outcome. It can be the potential outcome $Y_i(a)$, covariate function $f(X_i)$ or their linear combinations. Denote $\tilde{r}_{i,[k]}(a) = r_i(a) - E\{r_i(a)|B_i = k\}$. The sample means of the transformed outcomes are

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

defined as $\bar{r}_a = (1/n_a) \sum_{i=1}^n I(A_i = a) r_i(a)$ and the sample mean within stratum k is $\bar{r}_{[k]a} = (1/n_{[k]a}) \sum_{i=1}^n I(A_i = a, B_i = k) r_i(a)$. The population variance of $r_i(a)$ 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.

For our transfer learning framework, in addition to the target dataset $\{Y_i, X_i, A_i, B_i\}_{i=1}^n$, we also observe an external trial, which is referred to as the source data. Let the assignment set and the range of stratum labels in the source data be denoted by $\mathcal{A}_0^{(s)}$ and $\mathcal{K}^{(s)}$, respectively. Define $\{W_i^{(s)}\}_{i=1}^{n^{(s)}} = \{Y_i^{(s)}(0), \ldots, Y_i^{(s)}(A), B_i^{(s)}, X_i^{(s)}\}_{i=1}^{n^{(s)}}$ as $n^{(s)}$ i.i.d. samples from the source population $W^{(s)} = \{Y^{(s)}(0), \ldots, Y^{(s)}(A), B^{(s)}, X^{(s)}\}$. Here, the source sample is assumed to be independent of the target sample and the treatment assignments $(A_1, \ldots, A_n)$. Throughout, source data notations are indicated by the superscript (s); for example, $r_i^{(s)}(a)$ denotes the transformed outcomes in the source data. The theoretical assumptions for source data will be introduced in Section 4.

3. Covariate adjustment based on GLMs

Building on the covariate information X, the outcome $Y(a)$ follows a generalized linear model with density function $f\{Y(a)|X\} \sim \exp[\ \{Y(a)X^\top\beta - \psi(X^\top\beta)\}/c(\sigma)]$, where $\psi(\cdot)$ is the cumulant generating function, $\psi'(\cdot)$ is the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

inverse link function, $c(\sigma)$ is the dispersion parameter, and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is an unknown coefficient vector. We consider a generalized linear model as our working model, where $f\{Y(a)|X\}$ can be misspecified. The inverse link function $\psi'(X^\top\beta)$ is then used to adjust for covariates in the estimator. The choice of $\psi(X^\top\beta)$ depends on the problem being considered. For example, for binary responses, $\psi(t) = \log\{1 + \exp(t)\}$ corresponds to the logistic working model; for non-negative integer responses, $\psi(t) = \exp(t)$ corresponds to the Poisson working model.

For each stratum k and treatment a, we define the stratum-specific GLM coefficient $\beta_{[k]\mathrm{glm}}(a)$ of the generalized linear model by

$$\beta_{[k]\mathrm{glm}}(a) = \arg\min_{\beta\in\mathbb{R}^p} E[L\{Y_i(a), X_i, B_i, \beta\}|B_i = k],$$

where the loss function is $L\{Y_i(a), X_i, B_i, \beta\} = \psi(X_i^\top\beta) - Y_i(a)X_i^\top\beta$. Under a high-dimensional setting, we derive the coefficient by minimizing the loss function in the samples with an ℓ_1 penalty, that is

$$\hat{\beta}_{[k]\mathrm{glm}}(a) = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\{\psi(X_i^\top\beta) - Y_i X_i^\top\beta\} + \lambda_{[k]a}\|\beta\|_1,$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1-norm of β and $\lambda_{[k]a}$ is the tuning parameter in each stratum and treatment group. Based on this, the GLM-adjusted

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

treatment effect estimator for each treatment a is given by

$$\hat{\tau}_{\text{glm,a}} = \sum_{k=1}^{K} p_{n[k]} \Big\{ \Big(\bar{Y}_{[k]a} - \big[\bar{\psi}'_{[k]a}\{X_i^\top \hat{\beta}_{[k]\text{glm}}(a)\} - \bar{\psi}'_{[k]}\{X_i^\top \hat{\beta}_{[k]\text{glm}}(a)\} \big] \Big) - \Big(\bar{Y}_{[k]0} - \big[\bar{\psi}'_{[k]0}\{X_i^\top \hat{\beta}_{[k]\text{glm}}(0)\} - \bar{\psi}'_{[k]}\{X_i^\top \hat{\beta}_{[k]\text{glm}}(0)\} \big] \Big) \Big\}$$

and the GLM-adjusted treatment effect estimator is

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

Notice that the formulation of $\hat{\tau}_{\text{glm,a}}$ is equivalent to estimating the effects within each stratum and aggregating them via a weighted sum. Therefore, our proposed estimators inherently adjust for the stratification indicators, which aligns with recent model-robust inference literature under CAR (Ma et al., 2022; Ye et al., 2023).

We define the transformed outcome $r_{i,[k]\text{glm}}(a) = Y_i(a) - \psi'\{X_i^\top \beta_{[k]\text{glm}}(a)\}$, for $B_i = k, A_i = a$. Let $\beta_{j,[k]\text{glm}}(a)$ be the jth element of $\beta_{[k]\text{glm}}(a)$ and $S_{[k]} = \bigcup_{a \in \mathcal{A}_0} \{j \in \{1, \ldots, p\} : \beta_{j,[k]\text{glm}}(a) \neq 0\}$ be the union of the support of $\beta_{[k]\text{glm}}(a)$ for all $a \in \mathcal{A}_0$. Also, let $s_{[k]} = |S_{[k]}|$ be the number of relevant covariates in stratum k. The following assumptions on the inverse link function are required to derive the ℓ_1 convergence rate of $\hat{\beta}_{[k]\text{glm}}(a)$ and the asymptotic behavior of $\hat{\tau}_{\text{glm}}$. Let $||v||_\infty$ denote the ℓ_∞-norm of a vector v.

Assumption 3. *$\{X_i\}_{i=1}^n$ are uniformly bounded, which means that there exists a constant M independent of n, such that $\max_{i=1,\ldots,n} ||X_i||_\infty \leq M$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Assumption 4. *ψ is infinitely differentiable and strongly convex. Meanwhile, ψ'' satisfies the Lipschitz condition such that $|\psi''(u+v)-\psi''(v)| \leq L|u|$ for all $u, v \in \mathbb{R}$ and some constant L. One of the following conditions should be satisfied for some constant M_ψ: (1) ψ'' is uniformly bounded such that $\psi'' \leq M_\psi$; (2) $\sup_{|\Delta| \leq c} \psi''(X_i^\top \beta + \Delta) \leq M_\psi < \infty$ for some constant c.*

Assumption 3 requires that X_i is uniformly bounded, which may be a stringent assumption about X_i. However, this assumption helps to relax conditions on the transformed outcome $r_{i,[k]\mathrm{glm}}(a)$, allowing our GLM-adjusted estimator to remain valid when the working model is misspecified. Assumption 4 requires that ψ is infinitely differentiable and strongly convex, meaning that there exists a constant $c_\psi > 0$ such that $\psi''(t) \geq c_\psi$ for all relevant t (e.g., $t = X_i^\top \beta$). This condition is satisfied by commonly used generalized linear models such as logistic and Poisson regression models with uniformly bounded X_i. Furthermore, it is an essential requirement for the Hessian matrix to be positive definite, which is a mild assumption when X_i is sub-Gaussian (Li et al., 2024). Another requirement on ψ is similar to Assumption (GLM2) in the full-length version of Negahban et al. (2012), which includes the aforementioned models: the logistic model satisfies Assumption 4(1), and the Poisson model satisfies Assumption 4(2). To establish our theoretical guarantees, we make the following assumptions

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

regarding the population covariance and second moment matrices, as well as the tuning parameter $\lambda_{[k]a}$.

Assumption 5. *Let* $Q_{[k]XX} = E[X_i X_i^\top | B_i = k]$. *There exist constants* κ_l *and* κ_u *independent of* n *and* p, *satisfying* $0 < \kappa_l \leq \kappa_u < \infty$, *such that* $\lambda_{\min}(\Sigma_{[k]XX}) \geq \kappa_l, \lambda_{\max}(Q_{[k]XX}) \leq \kappa_u$ *for all* $k = 1, 2, \ldots, K$.

Assumption 6. *There exist constants* $c_{\lambda[k]} > 0$ *and* $c_M > 1$, *which are independent of* n *and determined by the upper bound* M *in Assumption 3 and the second moment* $E\{r^2_{i,[k]\mathrm{glm}}(a)\}$. *Furthermore, there exists a sequence* $M_{n[k]} \to \infty$ *with* n *in each stratum* k, *such that* $M^2_{n[k]} s^2_{[k]} (\log p)^2 / n \to 0$, $k = 1, 2, \ldots, K$. *The tuning parameter* $\lambda_{[k]a}$ *belongs to the interval*

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

Assumption 5 ensures that the minimum eigenvalue of the population-level stratum-specific covariance matrix $\Sigma_{[k]XX}$ is strictly positive, and the maximum eigenvalue of the second moment matrix $Q_{[k]XX}$ is bounded by a finite constant, preventing perfect collinearity, as well as explosive variances and means. In clinical trials, this condition can be supported by domain-driven pre-selection of distinct covariates, coupled with screening for redundant variables. Assumption 6 requires that the tuning parameter $\lambda_{[k]a}$ in each treatment a and stratum k be of order $O\{(M_{n[k]} \log p / n)^{1/2}\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Here the sequence $M_{n[k]}$ can diverge very slowly, e.g., $M_{n[k]} = \log\log n$, so that $\lambda_{[k]a}$ is approximately of order $(\log p/n)^{1/2}$, which is the same as the requirements in high-dimensional GLM studies (Van de Geer, 2008; Negahban et al., 2012).

The following proposition can be established for the ℓ_1 convergence rate of $\hat{\beta}_{[k]\mathrm{glm}}(a)$ and the asymptotic normality for $\hat{\tau}_{\mathrm{glm}}$. Let 1_A denote an A-dimensional vector with ones and $m_{[k]}(a) = E\{Y_i(a)|B_i = k\} - E\{Y_i(a)\}$. Throughout, we use $(\cdot : a \in \mathcal{A})$ to denote an A-dimensional column vector, and $\{\cdot : (a, a') \in \mathcal{A} \times \mathcal{A}\}$ to denote a matrix with entries indexed by (a, a'). Also, let V_{glm} be the asymptotic variance of $\hat{\tau}_{\mathrm{glm}}$, whose detailed expression is given below.

Proposition 1. *Under Assumptions 1–6, if* $r_{i,[k]\mathrm{glm}}(a) \in \mathcal{R}_2$, $E\{r^2_{i,[k]\mathrm{glm}}(a)\} < \infty$ *for* $a \in \mathcal{A}_0$, *then* $\|\hat{\beta}_{[k]\mathrm{glm}}(a) - \beta_{[k]\mathrm{glm}}(a)\|_1 = O_p[s_{[k]}\{M_{n[k]}(\log p)/n\}^{1/2}]$, for $k \in \mathcal{K}$ and $a \in \mathcal{A}_0$. *Moreover,* $\sqrt{n}\ V_{\mathrm{glm}}^{-1/2}(\hat{\tau}_{\mathrm{glm}} - \tau) \xrightarrow{d} \mathrm{N}(0, I_A)$, *where* I_A *is the* $A \times A$ *identity matrix,* $V_{\mathrm{glm}} = V_{H\tilde{Y}} + V_{\tilde{r}_{\mathrm{glm}}} + V_{\tilde{\psi}_{a0}} + V_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} + V^{\top}_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}}$, *and*

$$V_{H\tilde{Y}} = \sum_{k\in\mathcal{K}} p_{[k]} \left(m_{[k]}(a) - m_{[k]}(0) : a \in \mathcal{A}\right) \times \left(m_{[k]}(a') - m_{[k]}(0) : a' \in \mathcal{A}\right)^{\top},$$

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

$$V_{\tilde{\psi}_{a0}} = \Big\{ \sum_{k\in\mathcal{K}} p_{[k]} \mathrm{Cov}[\psi'\{X_i^{\top}\beta_{[k]\mathrm{glm}}(a)\} - \psi'\{X_i^{\top}\beta_{[k]\mathrm{glm}}(0)\},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

$$\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a')\} - \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(0)\}] : (a, a') \in \mathcal{A} \times \mathcal{A}\Big\},$$

$$V_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} = \Big\{ \sum_{k\in\mathcal{K}} p_{[k]} \mathrm{Cov}\Big(\{Y_i(a) - Y_i(0)\} - \big[\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a)\} - \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(0)\}\big],$$

$$\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a')\} - \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(0)\}\Big) : (a, a') \in \mathcal{A} \times \mathcal{A}\Big\}.$$

Remark 1. Recent work has studied the use of generalized linear models for treatment effect estimation in randomized trials. Guo and Basse (2023) used the logistic and Poisson regression models as examples to illustrate their proposed Oaxaca–Blinder estimator. Bannick et al. (2025) applied their general theory to GLMs. Both papers consider the case where the dimension of the covariates is fixed. We consider a more general situation to handle high-dimensional covariates.

One reason to use a covariate-adjusted treatment effect estimator is its potential efficiency gain over the benchmark estimator $\hat{\tau}_{\mathrm{ben}}$, defined as $\hat{\tau}_{\mathrm{ben}} = (\hat{\tau}_{\mathrm{ben},1}, \ldots, \hat{\tau}_{\mathrm{ben,A}})^\top$ with $\hat{\tau}_{\mathrm{ben,a}} = \sum_{k=1}^{K} p_{n[k]}(\bar{Y}_{[k]a} - \bar{Y}_{[k]0})$ (Bugni et al., 2019). When the GLM is correctly specified, i.e., $E\{Y_i(a)|X_i, B_i = k\} = \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a)\}$, we have $V_{\tilde{r}_{\mathrm{glm}}\tilde{\psi}_{a0}} = 0$ and $\hat{\tau}_{\mathrm{glm}}$ achieves efficiency gain over $\hat{\tau}_{\mathrm{ben}}$. However, under model misspecification, $\hat{\tau}_{\mathrm{glm}}$ may be less efficient than $\hat{\tau}_{\mathrm{ben}}$ (Guo and Basse, 2023; Bannick et al., 2025). To guarantee efficiency gain regardless of model specification, we propose a projection method in Section 5.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

4. Transfer learning

4.1 Context and algorithm

When some strata are small, $\hat{\beta}_{[k]\text{glm}}(a)$ becomes unreliable. Transfer learning addresses this by incorporating external source data (Gu et al., 2024). We impose the following assumptions to construct a feasible transfer learning method for the GLM-adjusted estimator.

Assumption 7. *(1)* $X_i^{(s)}$ *is a p-dimensional vector.* $\mathcal{A}_0^{(s)} = \mathcal{A}_0$ *and* $\mathcal{K}^{(s)} = \mathcal{K}$*; (2)* $E\{(Y_i^{(s)})^2(a)\} < \infty$ *and* $Y_i^{(s)}(a) \in \mathcal{R}_2$*, for all* $a \in \mathcal{A}_0^{(s)}$*. Additionally,* $\{X_i^{(s)}\}_{i=1}^{n^{(s)}}$ *are uniformly bounded by a constant, which means that there exists a constant* M *independent of* $n^{(s)}$ *such that* $\max_{i=1,\ldots,n^{(s)}} ||X_i^{(s)}|| \leq M$*; (3) Conditional on* $\{B_1^{(s)}, \ldots, B_n^{(s)}\}$, $\{A_1^{(s)}, \ldots, A_n^{(s)}\}$ *is independent of* $\{Y_i^{(s)}(0), \ldots, Y_i^{(s)}(A), X_i^{(s)}\}_{i=1}^{n^{(s)}}$*; and (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 7(1) requires that the treatments and the stratum labels are identical in both the source and target data. Using the same treatments in two clinical trials means that the specific drugs, biomarkers, or treatment procedures are the same. Similarly, when two trials have identical stratum labels, it indicates that they rely on the same covariates, such as gender or age groups, for the randomization process. This assumption

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

is mainly imposed to simplify the technical development, by ensuring that information can be appropriately transferred across corresponding treatment groups and strata in the source and target data. In practice, it may be relaxed to allow for approximate alignment between source and target populations; see Remark 3 for further discussion. Assumption 7(2) mirrors Assumption 1, with a uniform bound M across both datasets. The CAR conditions in Assumptions 7(3)–(4) allow the randomization procedures to differ between trials.

Define the stratum-specific projection coefficient for generalized linear models in the source data as

$$\beta^{(s)}_{[k]\mathrm{glm}}(a) = \arg\min_{\beta\in\mathbb{R}^p} E[L\{Y_i^{(s)}(a), X_i^{(s)}, B_i^{(s)}, \beta\}|B_i^{(s)} = k].$$

The working model for source data is the same as the target data, which implies that the loss function is identical across source and target data in the same treatment and stratum. We use the difference between the GLM coefficients as the similarity measure between source and target data in each treatment a and stratum k, which is $\delta_{[k]}(a) = \beta_{[k]\mathrm{glm}}(a) - \beta^{(s)}_{[k]\mathrm{glm}}(a)$. Our transfer learning algorithm shares a similar spirit with the method of Tian and Feng (2023). In each treatment a and stratum k, we first estimate

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

$\beta^{(s)}_{[k]\mathrm{glm}}(a)$ in source data as

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

Then we correct the bias $\delta_{[k]}(a)$ in target data as

$$\hat{\delta}_{[k]}(a) = \arg\min_{\delta\in\mathbb{R}^p} \frac{1}{n_{[k]a}} \sum_{i=1}^{n} I(A_i = a, B_i = k)\Bigg(\psi\Big[X_i^\top\{\hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a) + \delta\}\Big] - Y_i X_i^\top\{\hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a) + \delta\}\Bigg) + \lambda_{[k]a}\|\delta\|_1.$$

Then the transfer learning coefficient $\hat{\beta}_{[k]\mathrm{tl}}(a)$ is obtained by

$$\hat{\beta}_{[k]\mathrm{tl}}(a) = \hat{\delta}_{[k]}(a) + \hat{\beta}^{(s)}_{[k]\mathrm{glm}}(a).$$

Finally, we plug in the transfer learning coefficient to derive the transfer learning estimator in each treatment a, which is

$$\hat{\tau}_{\mathrm{tl,a}} = \sum_{k=1}^{K} p_{n[k]}\Bigg\{\Big(\bar{Y}_{[k]a} - \big[\bar{\psi}'_{[k]a}\{X_i^\top\hat{\beta}_{[k]\mathrm{tl}}(a)\} - \bar{\psi}'_{[k]}\{X_i^\top\hat{\beta}_{[k]\mathrm{tl}}(a)\}\big]\Big) - \Big(\bar{Y}_{[k]0} - \big[\bar{\psi}'_{[k]0}\{X_i^\top\hat{\beta}_{[k]\mathrm{tl}}(0)\} - \bar{\psi}'_{[k]}\{X_i^\top\hat{\beta}_{[k]\mathrm{tl}}(0)\}\big]\Big)\Bigg\},$$

and the transfer learning estimator for treatment effect is

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

4.2 Theoretical results

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

This characterizes the degree of similarity between the source and target data. Next, we introduce two assumptions for the source data that are analogous to those imposed on the target data.

Assumption 8. *In each stratum k, the minimum eigenvalue of the stratum-specific covariance matrix $\Sigma^{(s)}_{[k]XX}$ is denoted by $\Lambda^{(s)}_{\min,[k]}$, and it satisfies $\Lambda^{(s)}_{\min,[k]} \geq \kappa_l > 0$.*

Assumption 9. *There exist constants $c^{(s)}_{\lambda[k]} > 0$, $c^{(s)}_M > 1$, and a sequence $M^{(s)}_{n[k]} \to \infty$ with $n^{(s)}$ (taking the same functional form as $M_{n[k]}$) in each stratum k, such that $\{M^{(s)}_{n[k]}\}^2 s^2_{[k]} (\log p)^2 / n^{(s)} \to 0$, $k = 1, 2, \ldots, K$, and the tuning parameter $\lambda^{(s)}_{[k]a}$ lies in the interval*

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

These assumptions are analogous to Assumptions 5 and 6 for the target data. One difference is that Assumption 8 only requires the minimum eigenvalue of the source covariance matrix $\Sigma^{(s)}_{[k]XX}$ to be bounded. Intuitively, because we use a generalized linear model anchored to the target X_i, we can separate the influence of X_i from the coefficient estimation of $\beta_{[k]\mathrm{glm}}(a)$. Therefore, the source data is only involved in estimating $\beta_{[k]\mathrm{glm}}(a)$ (where a bounded minimum eigenvalue is sufficient), while the maximum eigenvalue condition on $Q_{[k]XX}$ remains strictly necessary for the target

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

data. To advance from the results in Proposition 1 to the transfer learning improvements in Theorem 1, we require the aforementioned additional assumptions on the source data (Assumptions 7–9) and replace Assumption 6 by a relaxation on sequence $M_{n[k]}$. Then, we establish the ℓ_1 convergence rate of $\hat{\beta}_{[k]\mathrm{tl}}(a)$, which is faster than we obtained for $\hat{\beta}_{[k]\mathrm{glm}}(a)$ in Proposition 1 as long as the similarity conditions hold with large enough source data.

Theorem 1. *Under Assumptions 1–5 and 7–9, suppose* $h < s_{[k]}\sqrt{\log p/n}$, $n^{(s)} \gg n$ *and* $h(\log p/n^{(s)})^{1/2} = o(1)$, *and the tuning parameter* $\lambda_{[k]a}$ *belongs to the interval*

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

where $M_{n[k]} \to \infty$ *with* $M_{n[k]}/n \to 0$. *If* $r_{i,[k]\mathrm{glm}}(a) \in \mathcal{R}_2$ *and* $E\{r_{i,[k]\mathrm{glm}}(a)\}^2 < \infty$ *for* $a \in \mathcal{A}_0$, *then* $\|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\|_1 = O_p\{s_{[k]}\sqrt{(M^{(s)}_{n[k]}\log p)/(n+n^{(s)})} + h\}$.

With the convergence rate of $\|\hat{\beta}_{[k]\mathrm{tl}}(a) - \beta_{[k]\mathrm{glm}}(a)\|_1$, we can derive the asymptotic normality of $\hat{\tau}_{\mathrm{tl}}$, which has the same asymptotic covariance matrix as $\hat{\tau}_{\mathrm{glm}}$ but requires weaker sparsity conditions.

Theorem 2. *Under Assumptions 1–5 and 7–9, suppose* $h < s_{[k]}\sqrt{\log p/n}$, $n^{(s)} \gg n$ *and* $h(M_{n[k]}\log p)^{1/2} = o(1)$, *and the tuning parameter* $\lambda_{[k]a}$ *satisfies the condition in Theorem 1. When* $r_{i,[k]\mathrm{glm}}(a) \in \mathcal{R}_2$ *and* $E\{r_{i,[k]\mathrm{glm}}(a)\}^2 <$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

∞ *for* $a \in \mathcal{A}_0$, *we have* $\sqrt{n}\, V_{\text{glm}}^{-1/2}(\hat{\tau}_{\text{tl}} - \tau) \xrightarrow{d} \mathrm{N}(0, I_A)$, *where* V_{glm} *is as defined in Proposition 1.*

Remark 2. Theorem 2 has a stricter requirement on the similarity condition such that $h(M_{n[k]} \log p)^{1/2} = o(1)$ compared to Theorem 1. Intuitively, establishing asymptotic normality requires scaling the remainder terms by an extra factor of $\sqrt{n}$, which demands tighter control on the bias term related to h. Compared with Proposition 1, $\hat{\beta}_{[k]\text{tl}}(a)$ achieves a faster convergence rate when $n^{(s)} \gg n$ and $h < s_{[k]}\sqrt{\log p/n}$, thereby reducing the target sample size required for asymptotic normality. This rate result parallels those in Tian and Feng (2023) and Li et al. (2024), except for the presence of the diverging sequences $M_{n[k]}$ and $M_{n[k]}^{(s)}$. The introduction of these two sequences enlarges the tuning parameters to accommodate the relaxed conditions on the residuals, which allows for possible misspecification of the generalized linear model in the target and source datasets, respectively.

Remark 3. In the current formulation, Assumption 7 requires exact alignment of the treatment space and stratification structure. While this facilitates a clean theoretical presentation, it can be practically relaxed. For instance, when the source data involves a finer stratification or a broader set of treatments ($\mathcal{A} \subset \mathcal{A}^{(s)}$ and $\mathcal{K} \subset \mathcal{K}^{(s)}$), our transfer learning procedure

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

remains applicable to the overlapping strata, and the established theoretical guarantees continue to hold. Furthermore, when facing reverse inclusion ($\mathcal{A}^{(s)} \subset \mathcal{A}$ and $\mathcal{K}^{(s)} \subset \mathcal{K}$) or merely partial intersections ($\mathcal{A}^{(s)} \cap \mathcal{A} \neq \emptyset$ and $\mathcal{K}^{(s)} \cap \mathcal{K} \neq \emptyset$), we can construct a composite estimator. Such an estimator can leverage transfer learning within the shared strata and treatment arms, while employing the target-only estimator for the remainder. A rigorous theoretical treatment of this composite approach is left for future research.

Remark 4. Theorems 1 and 2 rely on a coefficient-level similarity condition, expressed through the boundedness of $\delta_{[k]}(a)$. This requirement might not hold under a weaker condition where only the nuisance functions $E\{Y(a)|X\}$ are similar but the covariate distributions differ substantially. However, our transfer learning framework remains practically motivated. For example, one could select specific source subjects or explicitly match the source covariate distribution to the target trial. Although data selection methods designed for independent data, such as the cross-validation approach in (Tian and Feng, 2023), may perform adequately in practice, they lack theoretical guarantees in our setting. This is because adaptive randomization introduces complex dependency among responses, which complicates the sample-splitting procedures required for standard

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

cross-validation. From an applied perspective, this similarity condition can be evaluated qualitatively based on domain expertise. Specifically, historical trial data have a strong potential to be selected as valid source data if they share highly comparable clinical characteristics with the target data, such as targeting the same biomarkers or involving treatments with similar active pharmaceutical ingredients.

Remark 5. Gu et al. (2024) proposed a transfer learning method for clinical trials using a high-dimensional linear model as the working model. Building on this work, we study transfer learning procedures for GLMs, thereby accommodating a broader class of outcomes and link functions. Consistent with their findings, our results show that transfer learning improves the convergence rate of the coefficient estimators. However, because we characterize the ℓ_1 convergence rate with respect to $\beta_{[k]\mathrm{glm}}(a)$, the asymptotic variance remains V_{glm}.

5. Variance reduction via projection

While transfer learning improves coefficient estimation, it does not alter the asymptotic variance, so it cannot guarantee efficiency gains over the benchmark. To achieve this, we introduce a projection-based refinement. Building on insights from linear adjustment (Gu et al., 2023), we directly project

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

$Y_i(a)$ onto the linear span of the mean function $\{\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(a)\}, a \in \mathcal{A}_0\}$ within each stratum k.

Let $\Psi_{[k]}(X_i) = (\psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(0)\}, \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(1)\}, \ldots, \psi'\{X_i^\top \beta_{[k]\mathrm{glm}}(A)\})^\top$ be a $(A+1)$-dimensional vector, $\Sigma_{[k]\Psi\Psi} = \mathrm{Cov}\{\Psi_{[k]}(X_i) \mid B_i = k\}$ denote its stratum-specific covariance matrix. We estimate $\Psi_{[k]}(X_i)$ by

$$\hat{\Psi}_{[k],\dagger}(X_i) = \Big(\psi'\{X_i^\top \hat{\beta}_{[k]\dagger}(0)\}, \psi'\{X_i^\top \hat{\beta}_{[k]\dagger}(1)\}, \ldots, \psi'\{X_i^\top \hat{\beta}_{[k]\dagger}(A)\}\Big)^\top,$$

where $\dagger = \mathrm{glm}, \mathrm{tl}$. Define $\beta_{[k]}^{\mathrm{proj}}(a) = \arg\min_{\beta \in \mathbb{R}^{A+1}} E\{([Y_i(a) - E\{Y_i(a)|B_i = k\}] - [\Psi_{[k]}(X_i) - E\{\Psi_{[k]}(X_i)|B_i = k\}]^\top \beta)^2 | B_i = k\}$. We can estimate $\beta_{[k]}^{\mathrm{proj}}(a)$ by linear regression in each treatment a and stratum k, that is

$$Y_i \sim \alpha + \hat{\Psi}_{[k],\dagger}^\top(X_i)\beta_{[k]}^{\mathrm{proj}}(a), \ A_i = a, \ B_i = k.$$

Let $\hat{\beta}_{[k]}^{\mathrm{proj}}(a)$ denote the estimator of $\beta_{[k]}^{\mathrm{proj}}(a)$ and $i \in [k]$ denote the unit i in stratum k. The GLM-projected treatment effect estimator (GLM-P estimator) and the transfer learning-projected treatment effect estimator (TL-P estimator) for treatment a are obtained as

$$\hat{\tau}_{\dagger,a}^{\mathrm{proj}} = \sum_{k \in \mathcal{K}} p_{n[k]} \Bigg(\Big[\bar{Y}_{[k]a} - \frac{1}{n_{[k]a}} \sum_{i \in [k]} \big\{I(A_i = a) - \frac{n_{[k]a}}{n_{[k]}}\big\} \hat{\Psi}_{[k],\dagger}^\top(X_i) \hat{\beta}_{[k]}^{\mathrm{proj}}(a)\Big] - \Big[\bar{Y}_{[k]0} - \frac{1}{n_{[k]0}} \sum_{i \in [k]} \big\{I(A_i = 0) - \frac{n_{[k]0}}{n_{[k]}}\big\} \hat{\Psi}_{[k],\dagger}^\top(X_i) \hat{\beta}_{[k]}^{\mathrm{proj}}(0)\Big] \Bigg),$$

and $\hat{\tau}_\dagger^{\mathrm{proj}} = (\hat{\tau}_{\dagger,1}^{\mathrm{proj}}, \hat{\tau}_{\dagger,2}^{\mathrm{proj}}, \ldots, \hat{\tau}_{\dagger,A}^{\mathrm{proj}})^\top$, for $\dagger = \mathrm{glm}, \mathrm{tl}$. A key challenge remains in obtaining consistent estimation of $\Psi_{[k]}(X_i)$. The following results estab-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

lish that the GLM-P and TL-P estimators are asymptotically normal under the same conditions as in Proposition 1 and Theorem 2, respectively. Define the projected transformed outcome $r_{i,[k]\text{proj}}(a) = Y_i(a) - \Psi_{[k]}(X_i)^\top \beta^{\text{proj}}_{[k]}(a)$, for $B_i = k$, $A_i = a$.

Theorem 3. *Under conditions of Proposition 1 for* $\hat{\tau}_{\text{glm}}$ *or Theorem 2 for* $\hat{\tau}_{\text{tl}}$, *suppose* $\Sigma_{[k]\Psi\Psi}$ *is positive definite in each stratum* k. *Then for* $\dagger = \text{glm}, \text{ tl}$, *we have* $\sqrt{n}(V^{\text{proj}}_{\text{glm}})^{-1/2}(\hat{\tau}^{\text{proj}}_{\dagger} - \tau) \xrightarrow{d} \text{N}(0, I_A)$, *where* $V^{\text{proj}}_{\text{glm}} = V_{\tilde{r}_{\text{proj}}} + V_{\psi_{\text{proj}}} + V_{H\tilde{Y}}$,

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

$$V_{\psi_{\text{proj}}} = \Big\{ \sum_{k\in\mathcal{K}} p_{[k]} \{\beta^{\text{proj}}_{[k]}(a) - \beta^{\text{proj}}_{[k]}(0)\}^\top \Sigma_{[k]\Psi\Psi} \{\beta^{\text{proj}}_{[k]}(a') - \beta^{\text{proj}}_{[k]}(0)\} : (a, a') \in \mathcal{A}\times\mathcal{A} \Big\},$$

and $V_{H\tilde{Y}}$ *is the same as that defined in Proposition 1.*

Next, we show that $\hat{\tau}^{\text{proj}}_{\dagger}$ is more efficient than $\hat{\tau}_{\text{ben}}$ in the following corollary. Let $V \preceq 0$ denote that the matrix V is negative semi-definite. Also, let V_{ben} denote the asymptotic covariance matrix of $\hat{\tau}_{\text{ben}}$, whose explicit form is given in (Bugni et al., 2019).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Corollary 1. *Under Assumptions 1–4, we have*

$$V_{\mathrm{glm}}^{\mathrm{proj}} - V_{\mathrm{ben}} = \sum_{k\in\mathcal{K}} p_{[k]} \Bigg[-\frac{1}{\pi_{[k]0}} \beta_{[k]}^{\mathrm{proj}}(0)^\top \Sigma_{[k]\Psi\Psi} \beta_{[k]}^{\mathrm{proj}}(0) 1_A 1_A^\top$$
$$- \mathrm{diag}\Big\{ \frac{1}{\pi_{[k]a}} \beta_{[k]}^{\mathrm{proj}}(a)^\top \Sigma_{[k]\Psi\Psi} \beta_{[k]}^{\mathrm{proj}}(a) : a \in \mathcal{A} \Big\}$$
$$+ \Big\{ \{\beta_{[k]}^{\mathrm{proj}}(a) - \beta_{[k]}^{\mathrm{proj}}(0)\}^\top \Sigma_{[k]\Psi\Psi} \{\beta_{[k]}^{\mathrm{proj}}(a') - \beta_{[k]}^{\mathrm{proj}}(0)\} : (a, a') \in \mathcal{A} \times \mathcal{A} \Big\} \Bigg] \preceq 0.$$

Remark 6. The form of our projection estimator is similar to the estimator proposed in Cohen and Fogarty (2024). They applied a specific calibration method to generalized Oaxaca–Blinder estimators to address the non-inferior variance problem identified in Guo and Basse (2023). Bannick et al. (2025) used the same technique for their generalized treatment effect estimator and applied it to generalized linear models with a fixed dimension p for the additional baseline covariates. Using a projection viewpoint, we provide a unified explanation of these methods and clarify their underlying connections.

In summary, we propose a transfer learning enhancement and a projection refinement to the GLM-adjusted estimator. In particular, transfer learning improves coefficient estimation and relaxes sparsity conditions, while the efficiency gain in treatment effect estimation is achieved through the projection method. We summarize the contributions of each result in Table 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Table 1: Summary of estimators and their theoretical properties

Estimator	Theoretical result	Faster convergence rate	Guarantee efficiency gain
$\hat{\tau}_{\text{glm}}$	Proposition 1	–	–
$\hat{\tau}_{\text{tl}}$	Theorems 1–2	✓	×
$\hat{\tau}_{\text{glm}}^{\text{proj}}$	Theorem 3 & Corollary 1	×	✓
$\hat{\tau}_{\text{tl}}^{\text{proj}}$	Theorem 3 & Corollary 1	✓	✓

To obtain valid inference, we construct plug-in estimators for the asymptotic variances of the proposed estimators, focusing on contrasts between the potential outcomes under any two treatments. The asymptotic variances can be expressed in terms of population quantities such as GLM coefficients, projection coefficients, and covariance matrices, all of which are replaced by their corresponding estimators or natural sample analogues. For example, the variance estimator of $\hat{\tau}_{\text{glm}}^{\text{proj}}$ involves $\hat{\beta}_{[k]\text{glm}}(a)$, $\hat{\beta}_{[k]}^{\text{proj}}(a)$, the sample variance of $Y_i(a)$, and the sample covariance matrix of $\Psi_{[k]}(X_i)$. The resulting variance estimators are consistent. See the Supplementary Material for explicit expressions and proofs.

6. Simulation study

In this section, we compare the performance of our proposed estimators with a benchmark estimator $\hat{\tau}_{\text{ben}}$ across various scenarios. We consider

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

logistic regression and Poisson regression models to justify our theorems for the GLM-adjusted estimator, the transfer learning estimator and their corresponding projection estimators.

We consider a high-dimensional setting where the covariate dimension is $p = 100$. Let X_i be an s-dimensional covariate such that X_{i1} is a stratified random variable and $\{X_{ij}\}_{j=2}^{s}$ are regarded as additional baseline covariates. X_{i1} takes values in $\{1, 2\}$ with probabilities $\{0.4, 0.6\}$. The distribution of $\{X_{ij}\}_{j=2}^{s}$ will be specified in each model below. Independent of X_i, covariates other than X_i are generated by a multivariate normal distribution with mean zero and a diagonal covariance matrix where all diagonal elements are 2. For the randomization procedure, we consider stratified block randomization for 3 treatments and use X_{i1} as the stratum variable, resulting in 2 strata. The allocation ratio is 1:1:1 among treatments, with a block size of 6. The data generation models for target data and source data are detailed below:

Model 1 (logistic model): X_i follows Unif$[-5, 5]$ for $j = 2, 3, \ldots, s$ and they are independent of each other. Let $\mu_0 = 0.5$ and $\mu_1 = \mu_2 = 0.2$. As for the coefficients in the target data, $\beta_j(0) = 1$, $\beta_j(1) = 0.5$ and $\beta_j(2) = 0.5$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

for $j = 1, 2, \ldots, s$. The data generation models are as follows:

$$\Pr\{Y_i(a) = 1|X_i\} = \text{expit}\Big\{\mu_a + X_{i1}\beta_1(a) + \sum_{j=2}^{s} X_{ij}\beta_j(a)\Big\}, \quad a = 0, 1, 2.$$

Model 2 (Poisson model): X_i follows Unif$[-0.5, 0.5]$ for $j = 2, 3, \ldots, s$ and they are independent of each other. Let $\mu_0 = 0.5$ and $\mu_1 = \mu_2 = 0.2$. As for the coefficients in the target data, $\beta_j(0) = 0.6$, $\beta_j(1) = 0.4$ and $\beta_j(2) = 0.5$ for $j = 1, 2, \ldots, s$. The data generation models are as follows:

$$\Pr\{Y_i(a) = k|X_i\} = \frac{\lambda_a^k(X_i)e^{-\lambda_a(X_i)}}{k!}, \ \lambda_a(X_i) = \exp\Big\{\mu_a + X_{i1}\beta_1(a) + \sum_{j=2}^{s} X_{ij}\beta_j(a)\Big\},$$

Both the target data and the source data follow either Model 1 or Model 2 with different coefficients. While both models are correctly specified here, an additional simulation setup examining the impact of model misspecification can be found in Section S5.3 of the Supplementary Material. We have $n = 600$ units in the target data and $n^{(s)} = 2400$ units in the source data. Let $s = \lceil n^r \rceil$ where $r = 0.04, 0.08, \ldots, 0.48, 0.5$. In source data, $\beta_j^{(s)}(a) = \beta_j(a) + hj/s$ for $j = 1, 2, \ldots, s$, $h = 0.1, 0.3, 0.5$, and $a = 0, 1, 2$, for both Models 1 and 2. This setup leads to $\|\beta(a) - \beta^{(s)}(a)\|_1 = h(s+1)/2$.

We present the simulation results for five treatment effect estimators in each model: the benchmark estimator $\hat{\tau}_{\text{ben}}$, the GLM-adjusted estimator $\hat{\tau}_{\text{glm}}$, the transfer learning estimator $\hat{\tau}_{\text{tl}}$, and their corresponding projection estimators $\hat{\tau}_{\text{glm}}^{\text{proj}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$. For all high-dimensional GLM estimations

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

in both the target and source data, the tuning parameters are selected via 10-fold cross-validation using the R package `glmnet` (Friedman et al., 2010) with the '1se' rule, which selects the largest tuning parameter λ such that the cross-validation error remains within one standard error of the minimum. Our estimators are calculated over $R = 1000$ replicates. Performance is measured by relative bias, sampling standard deviation, and coverage probability of 95% confidence intervals. The relative bias is defined as $\sqrt{n}R^{-1}\sum_{r=1}^{R}(\hat{\tau}_r-\tau)/\hat{\sigma}$, where $\hat{\tau}_r$ 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. To further isolate the contributions of the proposed methods, we provide an additional comparative perspective where each pair of adjacent estimators is directly compared in terms of their sampling standard deviations and nominal coverage probabilities. These detailed pairwise comparisons can be found in Section S5.2 of the Supplementary Material. The main results are summarized as follows:

(1) In Figure 1(A), under logistic regression, the relative biases of all estimators are close to zero, which is expected as all estimators are consistent estimators for the treatment effect. In Figure 2(A), under Poisson regression, when the sparsity of the coefficients decreases, $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{glm}}^{\text{proj}}$ exhibit noticeable relative biases. After applying transfer learning, if the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

source data and target data are similar ($h = 0.1$, $h = 0.3$), $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ have relative biases close to zero.

(2) In Figures 1(B) and 2(B), $\hat{\tau}_{\text{glm}}^{\text{proj}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ consistently obtain smaller standard deviations than their non-projection counterparts, which is consistent with Corollary 1. While Figure 2(B) shows that GLM-adjusted estimators alone may not guarantee efficiency gains over the benchmark, the transfer learning estimators $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ consistently reduce variance. Notably, $\hat{\tau}_{\text{tl}}^{\text{proj}}$ demonstrates efficiency improvements across all scenarios, even with large source-target discrepancies ($h = 0.5$).

(3) In Figures 1(C) and 2(C), as s increases, $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{glm}}^{\text{proj}}$ suffer substantial coverage drops, which may be due to overfitting when the condition $s \log p < \sqrt{n}$ is violated. In contrast, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ maintain the nominal 95% coverage under sufficient similarity, which agrees with Theorem 2. Even when the similarity condition is violated, their coverage declines much more slowly than that of the target-only estimators.

7. A semi-synthetic clinical trial data analysis

In this section, we analyze a semi-synthetic clinical trial dataset based on a study of nefazodone and the Cognitive Behavioral-Analysis System of Psychotherapy (CBASP) (Keller et al., 2000), which compared nefazodone,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

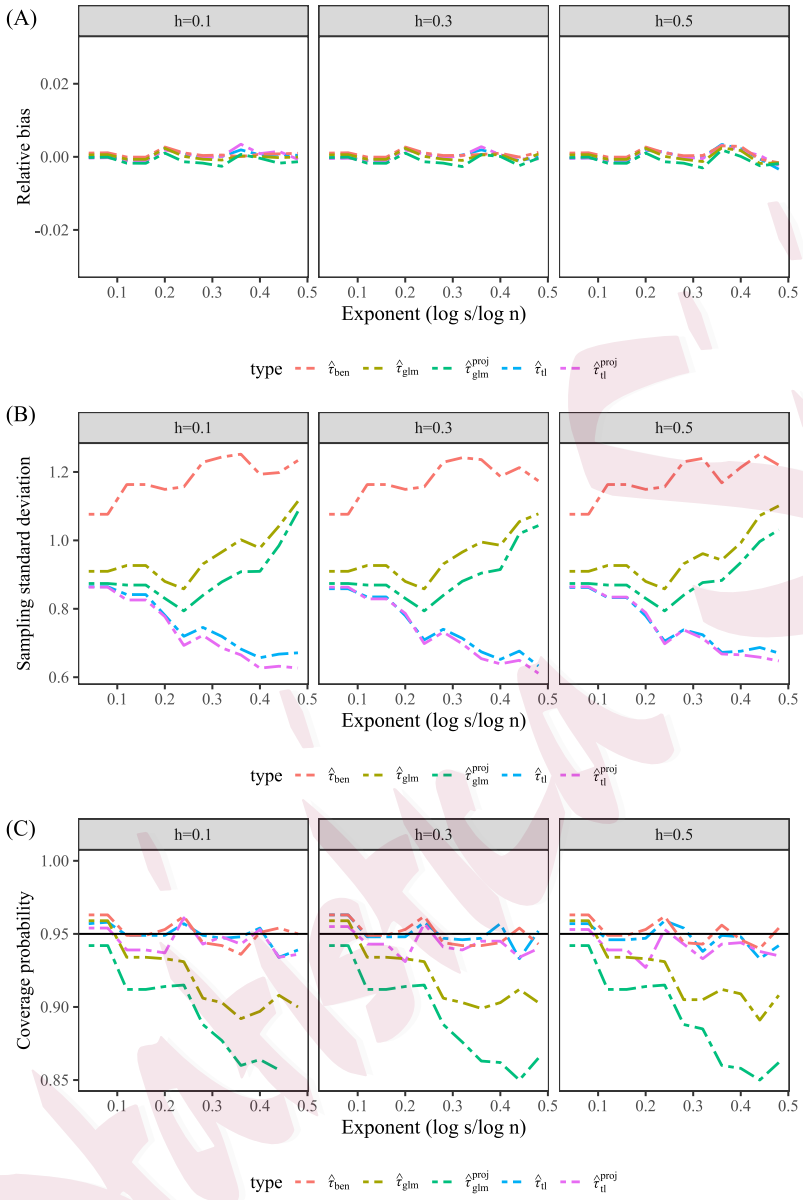


Figure 1: Simulation results for Case 1. Each column represents different h in the simulation settings. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

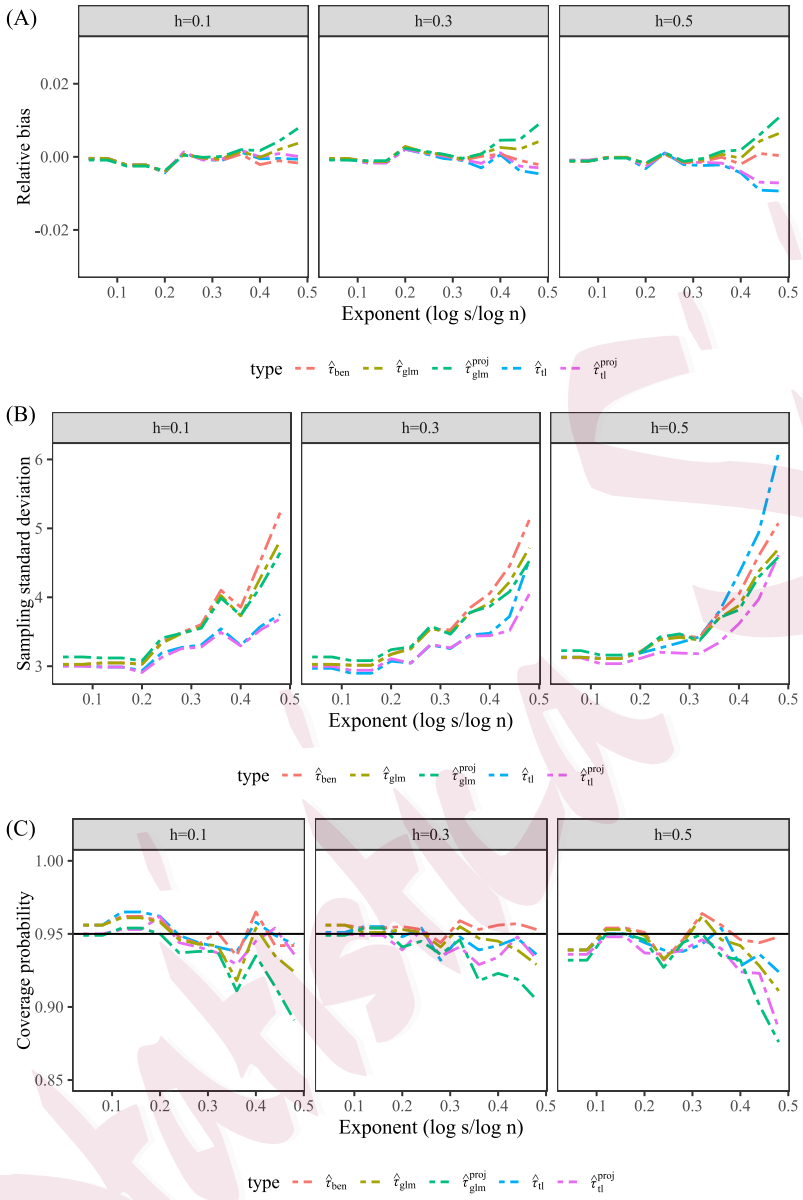


Figure 2: Simulation results for Case 2. Each column represents different h in the simulation settings. (A) Relative bias for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}^{\text{proj}}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}^{\text{proj}}_{\text{tl}}$. (B) Sampling standard deviation for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}^{\text{proj}}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}^{\text{proj}}_{\text{tl}}$. (C) Coverage probability for $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}^{\text{proj}}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}^{\text{proj}}_{\text{tl}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

CBASP, and their combination in chronic depression. We use the CBASP group as the treatment group, indexed by $A_i = 1$, and the combination group as the control group, indexed by $A_i = 0$. The binary outcome is the dichotomized FinalHAMD score (24-item Hamilton Rating Scale for Depression), with data stratified by GENDER. Additional covariates are generated from eight baseline variables, resulting in $p = 135$ covariates. The target and source populations are defined by HAMD17 < 20 and HAMD17 ≥ 20, respectively. In each iteration, we generate semi-synthetic data by sampling 200 units from the target population and 1200 units from the source population. Further details on the semi-synthetic data generation are provided in the Supplementary Material. We compare five estimators $\hat{\tau}_{\text{ben}}$, $\hat{\tau}_{\text{glm}}$, $\hat{\tau}_{\text{tl}}$, $\hat{\tau}_{\text{glm}}^{\text{proj}}$, and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ using 1000 Monte Carlo iterations. As shown in Table 2, all estimators produce similar treatment effect estimates. Regarding the variance, $\hat{\tau}_{\text{glm}}$ has the same variance as $\hat{\tau}_{\text{ben}}$, and the variance of $\hat{\tau}_{\text{tl}}$ is slightly smaller than theirs. After projection, the variances decrease, with $\hat{\tau}_{\text{tl}}^{\text{proj}}$ attaining the smallest variance. In terms of coverage probability, $\hat{\tau}_{\text{glm}}$ and $\hat{\tau}_{\text{glm}}^{\text{proj}}$ have coverage slightly below 95%, whereas after transfer learning, $\hat{\tau}_{\text{tl}}$ and $\hat{\tau}_{\text{tl}}^{\text{proj}}$ achieve approximately nominal 95% coverage.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

Table 2: Results of the semi-synthetic clinical trial data

Estimator	Estimate ($\times 10^2$)	Variance ($\times 10^2$)	Coverage probability ($\times 10^2$)
$\hat{\tau}_{\text{ben}}$	-0.939	0.471	94.6
$\hat{\tau}_{\text{glm}}$	-0.932	0.471	94.3
$\hat{\tau}_{\text{tl}}$	-0.908	0.466	94.6
$\hat{\tau}_{\text{glm}}^{\text{proj}}$	-0.789	0.466	93.1
$\hat{\tau}_{\text{tl}}^{\text{proj}}$	-0.991	0.448	94.6

8. Discussion

In this paper, we have proposed a transfer learning framework for treatment effect estimation under CAR with high-dimensional GLMs. By integrating external source data and a projection refinement, our approach achieves valid inference under relaxed sparsity conditions while guaranteeing efficiency gains over the benchmark estimator.

Despite its advantages, our transfer learning framework has some limitations regarding source-target comparability and the scope of the alignment assumptions. First, the transfer learning approach requires coefficient-level similarity between the source and target populations; see Remark 4 for a qualitative assessment strategy and practical considerations. Second, as we have discussed in Remark 3, the current theoretical framework assumes

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

exact alignment of the treatment space and stratification structure across the source and target populations. While practical implementations can address minor structural mismatches, formally extending the theory to accommodate partially overlapping strata or differing stratification criteria remains an open challenge. These two directions represent important avenues for future research. Additionally, although the projection method achieves efficiency gains under model misspecification, developing nonparametric approaches to transfer knowledge across trials is another promising direction.

Supplementary Material

The online supplementary material contains details on the variance estimators, the generation of the semi-synthetic clinical trial data, and proofs.

Acknowledgments

We thank the Associate Editor and the reviewers for their valuable comments and suggestions. 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 Shenzhen Science and Technology Program (grant No.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

AI2026036). Wei Ma is the corresponding author of this paper.

References

Bannick, M. S., J. Shao, J. Liu, Y. Du, Y. Yi, and T. Ye (2025). A general form of covariate adjustment in clinical trials under covariate-adaptive randomization. *Biometrika*, in press.

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

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics 10*(4), 1747–1785.

Ciolino, J. D., H. L. Palac, A. Yang, M. Vaca, and H. M. Belli (2019). Ideal vs. real: A systematic review on handling covariates in randomized controlled trials. *BMC Medical Research Methodology 19*, 136–146.

Cohen, P. L. and C. B. Fogarty (2024). No-harm calibration for generalized oaxaca–blinder estimators. *Biometrika 111*(1), 331–338.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Gu, Y., H. Liu, and W. Ma (2023). Regression-based multiple treatment effect estimation under covariate-adaptive randomization. *Biometrics 79*(4), 2869–2880.

Gu, Y., H. Liu, and W. Ma (2024). Incorporating external data for analyzing randomized clinical trials: A transfer learning approach. *arXiv preprint arXiv:2409.04126*.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

REFERENCES

Guo, K. and G. Basse (2023). The generalized oaxaca-blinder estimator. *Journal of the American Statistical Association 118*(541), 524–536.

Keller, M. B., J. P. McCullough, D. N. Klein, B. Arnow, D. L. Dunner, A. J. Gelenberg, et al. (2000). 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.

Li, S., L. Zhang, T. T. Cai, and H. Li (2024). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association 119*(546), 1274–1285.

Li, Y., W. Ma, Y. Qin, and F. Hu (2021). Testing for treatment effect in covariate-adaptive randomized trials with generalized linear models and omitted covariates. *Statistical Methods in Medical Research 30*(9), 2148–2164.

Liu, H., F. Tu, and W. Ma (2023). Lasso-adjusted treatment effect estimation under covariate-adaptive randomization. *Biometrika 110*(2), 431–447.

Ma, W., P. Li, L.-X. Zhang, and F. Hu (2024). A new and unified family of covariate adaptive randomization procedures and their properties. *Journal of the American Statistical Association 119*(545), 151–162.

Ma, W., F. Tu, and H. Liu (2022). Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective. *Statistics in Medicine 41*(29), 5645–5661.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

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

Rosenberger, W. F. and J. M. Lachin (2015). *Randomization in Clinical Trials: Theory and Practice* (2nd ed.). John Wiley & Sons.

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

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

Torrey, L. and J. Shavlik (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–264. IGI global.

Van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*(2), 614–645.

Wang, T. and W. Ma (2025). The impact of misclassification on covariate-adaptive randomized clinical trials with generalized linear models. *Journal of Statistical Planning and Inference 234*, 106209.

Ye, T., J. Shao, Y. Yi, and Q. Zhao (2023). Toward better practice of covariate adjust-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0480

ment in analyzing randomized clinical trials. *Journal of the American Statistical Association 118*(544), 2370–2382.

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

Yujia Gu

Department of Statistics and Data Science, Tsinghua University, Beijing, 100084, China

E-mail: guyj@mail.tsinghua.edu.cn

Hanzhong Liu

Department of Statistics and Data Science, Tsinghua University, Beijing, 100084, China

E-mail: lhz2016@tsinghua.edu.cn

Wei Ma

Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, China

E-mail: mawei@ruc.edu.cn