Statistica Sinica Preprint No: SS-2025-0502	
Title	Cross-Validation Model Averaging Under Covariate-Adaptive Randomization
Manuscript ID	SS-2025-0502
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0502

Complete List of Authors	Fuyi Tu, Jiahui Xin and Wei Ma
Corresponding Authors	Wei Ma
E-mails	mawei@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

CROSS-VALIDATION MODEL AVERAGING UNDER COVARIATE-ADAPTIVE RANDOMIZATION

Fuyi Tu[1], Jiahui Xin[2] and Wei Ma[2]

[1]*Chongqing University of Posts and Telecommunications*

and [2]*Renmin University of China*

Abstract: Covariate-adaptive randomization (CAR) leverages covariates at the design stage to enhance comparability between treatment groups, but typically allows only a limited set of covariates to be incorporated. We therefore study covariate adjustment at the analysis stage to achieve more efficient estimation of the treatment effect under CAR. Existing work on covariate adjustment under CAR typically relies on a single, pre-specified model. Although this complies with regulatory guidance that the adjustment model be specified before the trial, a single-model approach can be vulnerable to misspecification and may be less efficient when the covariate-outcome relationship is uncertain. To address this issue, we propose a model averaging framework for covariate adjustment under CAR that accommodates a collection of pre-specified candidate models. The averaging weights are selected via K-fold cross-validation to minimize the asymptotic variance of the treatment effect estimator. We establish that the resulting estimator is asymptotically normal and achieves asymptotic variance optimality within the candidate model class under mild regularity conditions. Simulation studies demonstrate that the proposed approach yields substantial variance re-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

duction compared with single-model adjustment. Overall, this paper introduces a theoretically grounded and flexible framework for treatment effect estimation under CAR that complies with regulatory pre-specification requirements while effectively handling uncertainty in covariate-outcome relationships.

Key words and phrases: Covariate-adaptive randomization, Covariate adjustment, Cross-validation, Model averaging, Machine learning

1. Introduction

Covariate-adaptive randomization (CAR) is widely used in clinical trials. It incorporates baseline covariates during the randomization process to improve the balance between treatment arms, thus enhancing the statistical efficiency and credibility of treatment effect estimates (ICH E9, 1998; EMA, 2015; Holmberg and Andersen, 2022). Traditional CAR procedures, such as the stratified biased-coin design (Efron, 1971), stratified block randomization (Zelen, 1974) and Pocock and Simon's minimization (Pocock and Simon, 1975), focus on achieving balance within discrete covariate categories or strata. Recent advances have extended designs to balance general covariate features (Ma et al., 2024). However, CAR is constrained to a limited number of covariates, leaving substantial covariate information available for exploitation through adjustment during the analysis stage. Moreover, regulatory guidelines recommend that covariates used for randomization be

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

included in the analysis stage to preserve type I error control and ensure valid inference (FDA, 2023). Consequently, this paper focuses on covariate adjustment for treatment effect estimation in the analysis stage under CAR.

Over the past decades, researchers have explored the efficiency gains achievable through covariate adjustment under complete randomization, especially when the model relating covariates to outcomes is correctly specified (e.g., Robinson and Jewell, 1991; Yang and Tsiatis, 2001; Pocock et al., 2002; Tsiatis et al., 2008). Recent research also focuses on developing robust covariate-adjusted estimation and valid inference for treatment effects under CAR (e.g., Bugni et al., 2018, 2019; Ma et al., 2022; Ye et al., 2022; Tu et al., 2024). However, existing approaches to covariate-adjusted estimation typically rely on a single model, which poses practical challenges because investigators often possess incomplete knowledge of the true relationships between numerous baseline covariates and outcomes at the design stage (Morris et al., 2022). Meanwhile, regulatory guidelines generally require that covariate adjustment models be pre-specified in the statistical analysis plan before outcome data are examined or unblinded (ICH E9 (R1), 2017; FDA, 2023). Although such pre-specification protects trial integrity, prevents data-driven bias, and aids in controlling type I error, the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

chosen model may not adequately capture the underlying relationships in the observed data. Treatment effects can still be estimated validly under model misspecification (Ma et al., 2022; Ye et al., 2023); however, accurately modeling the outcome–covariate relationship can substantially improve estimation efficiency (Rafi, 2023; Tu et al., 2024). This consideration becomes particularly critical in modern clinical trials, where high-dimensional prognostic covariate information is increasingly abundant (Rosenblum and Van Der Laan, 2010; Cerreia-Vioglio et al., 2025). The resulting challenge is to retain the benefits of pre-specification while accommodating uncertainty about the underlying outcome-covariate relationship. To address this issue, we propose a model averaging framework for covariate adjustment in treatment effect estimation. Unlike conventional methods that depend on a single model, our framework combines estimates from a diverse yet predefined set of candidate models. This ensemble approach improves inferential efficiency while fully adhering to regulatory mandates.

Model averaging has emerged as a powerful tool for enhancing prediction and estimation accuracy, with substantial methodological advancements over recent decades (e.g., Yang, 2000; Hansen, 2007; Wan et al., 2010; Hansen and Racine, 2012; Zhang et al., 2015, 2016; Liao et al., 2019; Li et al., 2022; Fang et al., 2022). However, previous research on model averaging

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

has primarily focused on improving predictions, such as in risk assessment, while its application to treatment effect estimation remains relatively limited. Existing contributions in this area include the treatment effect estimation by mixing (TEEM) method of Rolling et al. (2019), which employs a minimax-regret approach but does not target asymptotic optimality; the optimal model averaging estimator of Zhao et al. (2024) for linear models under squared-error loss; and the spline-based additive model approach of Li et al. (2025) for estimating conditional treatment effects. To the best of our knowledge, no prior work has used the asymptotic variance of the treatment effect estimator as the explicit criterion for determining model weights under CAR, nor has it developed valid inference procedures in this setting that accommodate potential model misspecification. We specifically target minimizing the variance of treatment effect estimators because employing different models for covariate adjustment only influences the variance of the oracle estimator. Thus, minimizing this variance is equivalent to enhancing statistical efficiency, which is the primary goal of our study. Our proposed framework improves efficiency, ensures valid inference, and offers a principled yet flexible approach to treatment effect estimation.

To select the optimal model weights, we employ a K-fold cross-validation (CV) procedure that determines data-driven weights by minimizing a crite-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

rion serving as a proxy for the asymptotic variance of the treatment effect estimator. This process operates entirely within a pre-defined set of candidate models, rigorously adhering to the pre-specification principles to ensure regulatory compliance. Under mild regularity conditions, we establish that the selected model weights are asymptotically optimal in the sense of achieving the minimal possible variance among all feasible weight combinations within the candidate set. We further prove that the proposed treatment effect estimator is asymptotically normal and provide a consistent variance estimator to support the construction of valid confidence intervals and hypothesis tests. Notably, these theoretical results remain valid across various covariate-adaptive randomization procedures.

The remainder of this paper is organized as follows. Section 2 introduces the framework and notation. We detail the model averaging estimator and the CV-based weight selection procedure in Section 3, and derive the asymptotic optimality and normality of the proposed treatment effect estimator in Sections 4 and 5. Section 6 reports the simulation results and demonstrates the practical performance of the model averaging estimator based on a clinical trial example; Section 7 provides conclusions.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

2. Framework and notation

This paper considers a two-arm randomized trial involving n subjects. For subject $i = 1, 2, \ldots, n$, let A_i be the treatment assignment indicator, where $A_i = 1$ denotes assignment to the treatment arm and $A_i = 0$ to the control arm. The target proportion of subjects assigned to treatment is $\pi = P(A_i = 1) \in (0, 1)$. We adopt the Neyman–Rubin potential outcomes framework, in which the observed outcome for subject i is

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0),$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes under the treatment and control groups, respectively. Before randomization, we observe p-dimensional baseline covariates $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ for each subject. The subjects are then stratified into S strata based on some or all components of $\boldsymbol{X}_i$ via a deterministic function that maps the support of $\boldsymbol{X}_i$ to $\{1, \ldots, S\}$. We denote the resulting stratum label by B_i. Since the labels are obtained through a deterministic function of the independent covariates $\boldsymbol{X}_i$, the B_i are independent across subjects. To rule out the degenerate case, we assume that subject assignment probabilities are strictly positive across all strata, i.e., $P(B_i = s) > 0$ for $s \in \{1, \ldots, S\}$. Subscripts 1 and 0 are used to denote quantities relating to the treatment and control groups, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Let $n_1 = \sum_{i=1}^n A_i$ be the number of treated subjects and $n_0 = \sum_{i=1}^n (1-A_i)$ the number of subjects assigned to the control group. Within each stratum s, we use the subscript $[s]$ to index statistics calculated using information from $\mathcal{J}_{[s]}$, where $\mathcal{J}_{[s]}$ denotes the set of subjects in stratum s. Specifically, we define $n_{[s]} = \sum_{i\in\mathcal{J}_{[s]}} 1$ as the total number of subjects in stratum s, $n_{[s]1} = \sum_{i\in\mathcal{J}_{[s]}} A_i$ as the number of treated subjects in stratum s, and $n_{[s]0} = \sum_{i\in\mathcal{J}_{[s]}} (1-A_i)$ as the number of control subjects in stratum s. The proportion of the total sample and the proportion treated within stratum s are denoted by $p_{n[s]} = n_{[s]}/n$ and $\pi_{n[s]} = n_{[s]1}/n_{[s]}$, respectively. The parameter of interest is the average treatment effect (ATE)

$$\tau = E\{Y_i(1) - Y_i(0)\}.$$

Define the L_2 norm of a random variable V as $||V||_{L_2} = \{E(V^2)\}^{1/2}$. Applying this definition to $V = f(\boldsymbol{X})$ gives $||f(\boldsymbol{X})||_{L_2} = \left[E\{f(\boldsymbol{X})^2\}\right]^{1/2}$. Let $\mathcal{R}_2 = \{V : \max_{s=1,\ldots,S} \mathrm{Var}(V|B_i = s) > 0\}$ be the set of random variables that have positive variance in at least one stratum. The following assumptions are imposed on the data-generating process and the randomization procedure.

Assumption 1. $\{Y_i(1), Y_i(0), \boldsymbol{X}_i\}_{i=1}^n$ is an independent and identically distributed (i.i.d.) sample from the population distribution of $\{Y(1), Y(0), \boldsymbol{X}\}$. Moreover, $E\{Y_i^2(a)\} < \infty$ and $Y_i(a) \in \mathcal{R}_2$ for $a = 0, 1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Assumption 2. Conditional on $\boldsymbol{B}^{(n)} = \{B_1, \ldots, B_n\}$, treatment assignments $\boldsymbol{A}^{(n)} = \{A_1, \ldots, A_n\}$ and $\{Y_i(1), Y_i(0), \boldsymbol{X}_i\}_{i=1}^n$ are independent.

Assumption 3. In each stratum $s \in \{1, \ldots, S\}$, where the total number of strata S is finite, the proportion of treated units $\pi_{n[s]}$ converges in probability to π.

Remark 1. Assumption 1 ensures that the sample is representative of the population and prevents issues with extreme values dominating the results, thus allowing for valid inference. Importantly, the treatment indicators $\{A_i\}_{i=1}^n$ are allowed to be dependent across subjects, as naturally occurs under covariate-adaptive randomization. When applying different estimation methods to baseline covariates, additional assumptions are imposed on $\boldsymbol{X}_i$. Assumption 2 formalizes the independence between the treatment assignment mechanism and the potential outcomes, conditional on the stratification variables. It holds under simple and restricted randomization (Rosenberger and Lachin, 2015), and is widely satisfied by existing covariate-adaptive randomization methods, including the stratified biased-coin design (Efron, 1971), stratified block design (Zelen, 1974), stratified adaptive biased-coin design (Wei, 1978), Pocock and Simon's minimization (Pocock and Simon, 1975), and Hu and Hu's method (Hu and Hu, 2012). Assumption 3 ensures the asymptotic balance in treatment allocation across

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

strata, which prevents systematic imbalances that may inflate variances.

The mean and variance of a random variable V are denoted by $\mu_V = E(V)$ and $\sigma_V^2 = \mathrm{Var}(V)$, respectively. Let $\tilde{V} = V - E(V|B)$ denote V centered at its stratum-specific conditional mean. Here $B \in \{1, \ldots, S\}$ denotes the stratum label, indicating which stratum the subject belongs to. For random variables $r_i(a), i = 1, \ldots, n$, and $a \in \{0, 1\}$, the sample means under treatment and control are defined as $\bar{r}_1 = n_1^{-1} \sum_{i=1}^n A_i r_i(1)$ and $\bar{r}_0 = n_0^{-1} \sum_{i=1}^n (1 - A_i) r_i(0)$, respectively. The corresponding stratum-specific sample means in stratum s are $\bar{r}_{[s]1} = n_{[s]1}^{-1} \sum_{i \in \mathcal{J}_{[s]}} A_i r_i(1)$ and $\bar{r}_{[s]0} = n_{[s]0}^{-1} \sum_{i \in \mathcal{J}_{[s]}} (1 - A_i) r_i(0)$, respectively. Building on the results of Tu et al. (2024), we first establish the asymptotic normality of the treatment effect estimator constructed using the oracle transformed outcomes:

Proposition 1. *Suppose that the oracle covariate-adjusted outcomes* $r_i(a) = Y_i(a) - h(\boldsymbol{X}_i)$ *belong to* $\mathcal{R}_2$ *and have finite second moments for* $a = 0, 1$ *and* $i = 1, \ldots, n$*, where* h *is a fixed measurable function of the covariates. Then, under Assumptions 1–3,*

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\big(0, \varsigma_r^2(\pi) + \varsigma_{Hr}^2\big), \ \hat{\varsigma}_r^2(\pi) + \hat{\varsigma}_{Hr}^2 \xrightarrow{P} \varsigma_r^2(\pi) + \varsigma_{Hr}^2,$$

where $\hat{\tau} = \sum_{s=1}^S p_{n[s]} \{\bar{r}_{[s]1} - \bar{r}_{[s]0}\}$ *denotes the stratum-weighted difference-in-means treatment effect estimator based on the oracle transformed outcomes*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

across treatment arms.

The asymptotic variance decomposes into two intuitive components: $\varsigma_r^2(\pi)$, which captures the within-stratum variability in the transformed outcomes, and ς_{Hr}^2, which quantifies the between-stratum heterogeneity in treatment effects. Formally,

$$\varsigma_r^2(\pi) = \frac{1}{\pi}\sigma^2_{r_i(1)-E\{r_i(1)|B_i\}} + \frac{1}{1-\pi}\sigma^2_{r_i(0)-E\{r_i(0)|B_i\}},$$

$$\varsigma_{Hr}^2 = \sum_{s=1}^{S} p_s\Big(\big[E\{r_i(1)|B_i=s\} - E\{r_i(1)\}\big] - \big[E\{r_i(0)|B_i=s\} - E\{r_i(0)\}\big]\Big)^2.$$

To facilitate inference, we construct plug-in estimators for these variance components, expressed as sample analogs that leverage observed or estimated transformed outcomes $\hat{r}_i(a)$:

$$\begin{aligned}
\hat{\varsigma}_r^2(\pi) &= \frac{1}{\pi}\sum_{s=1}^{S} p_{n[s]}\left[\frac{1}{n_{[s]1}}\sum_{i\in\mathcal{J}_{[s]}} A_i\Big\{\hat{r}_i(1) - \frac{1}{n_{[s]1}}\sum_{j\in\mathcal{J}_{[s]}} A_j\hat{r}_j(1)\Big\}^2\right] \\
&\quad + \frac{1}{1-\pi}\sum_{s=1}^{S} p_{n[s]}\left[\frac{1}{n_{[s]0}}\sum_{i\in\mathcal{J}_{[s]}} (1-A_i)\Big\{\hat{r}_i(0) - \frac{1}{n_{[s]0}}\sum_{j\in\mathcal{J}_{[s]}} (1-A_j)\hat{r}_j(0)\Big\}^2\right], \\
\hat{\varsigma}_{Hr}^2 &= \sum_{s=1}^{S} p_{n[s]}\left[\Big\{\frac{1}{n_{[s]1}}\sum_{j\in\mathcal{J}_{[s]}} A_j\hat{r}_j(1) - \frac{1}{n_1}\sum_{i=1}^{n} A_i\hat{r}_i(1)\Big\}\right. \\
&\quad \left. - \Big\{\frac{1}{n_{[s]0}}\sum_{j\in\mathcal{J}_{[s]}} (1-A_j)\hat{r}_j(0) - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\hat{r}_i(0)\Big\}\right]^2.
\end{aligned}$$

3. Model averaging estimator and weight selection procedure

Under complex data-generating processes, traditional single-model covariate-adjusted methods for treatment effect estimation are susceptible to model

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

misspecification and may lead to inefficient estimators. For example, a linear model may fail to adequately capture the nonlinear relationships between covariates and outcomes. To address this limitation and enhance efficiency, we adopt a model averaging approach that integrates projections from a set of candidate models into a weighted linear combination, where the weights are constrained to be nonnegative and to sum to 1. This approach mitigates the limitations of individual models while aiming to minimize the asymptotic variance of the ATE estimator.

The performance of the model averaging estimator hinges critically on the selection of weights assigned to candidate models. Although various weight selection criteria exist, such as those based on information criteria including AIC or BIC (Singh et al., 2010), their theoretical properties under the complex dependence structures induced by CAR are not well understood. Moreover, these criteria may not align well with the objective of precise treatment effect estimation. As established by Tu et al. (2024), the covariate adjustment methods in the oracle treatment effect estimator influence only its asymptotic variance, without affecting bias. Consequently, our goal is to enhance efficiency by directly minimizing the asymptotic variance $\hat{\varsigma}_r^2(\pi) + \hat{\varsigma}_{Hr}^2$ of the treatment effect estimator. Notably, for any transformed outcomes of the form $r_i(a) = Y_i(a) - h(\boldsymbol{X}_i), a = 0, 1$, the difference between

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

these transformed outcomes equals that between the original outcomes, implying $\varsigma_{Hr}^2 = \varsigma_{HY}^2$, where ς_{HY}^2 is obtained by replacing the transformed outcome r with the original outcome Y in the statistic ς_{Hr}^2. Consequently, we only need to minimize $\varsigma_r^2(\pi)$. The proposed procedure proceeds as follows:

I. Partition the data set into K folds, where each of the first $K-1$ folds contains $J = \lfloor n/K \rfloor$ observations and the final fold contains the remaining $n - (K-1) \times \lfloor n/K \rfloor$ observations.

II. For $k = 1, \ldots, K$,

   (a) Exclude the k-th fold from the data set and use the remaining observations to fit $m = 1, \ldots, M$ models for covariate adjustment. Here, M is assumed to be fixed. Denote the estimated covariate adjustment functions for stratum s and treatment group a by $\hat{h}_{m[s]}^{(-k)}(\cdot, a)$ for $m = 1, \ldots, M, s = 1, \ldots, S$ and $a = 0, 1$.

   (b) Compute the predictions for observations within the k-th fold using the M fitted models for each treatment group.

III. Compute the transformed outcomes for each model combination as follows: for the i-th observation in fold k and stratum s,

$$\hat{r}_i(\boldsymbol{w}, a) = Y_i(a) - \left\{ (1-\pi) \sum_{m=1}^{M} w_{(m)1} \hat{h}_{m[s]}^{(-k)}(\boldsymbol{X}_i, 1) + \pi \sum_{m=1}^{M} w_{(m)0} \hat{h}_{m[s]}^{(-k)}(\boldsymbol{X}_i, 0) \right\},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

where $\boldsymbol{w} = (w_{(1)1}, w_{(2)1}, \ldots, w_{(M)1}, w_{(1)0}, w_{(2)0}, \ldots, w_{(M)0})^{\mathrm{T}}$ is a $2M$-dimensional vector, and $w_{(m)1}$ and $w_{(m)0}$ can be different for model m. In the following statements, we use the subscript m to denote the statistics corresponding to model m, and the subscript k to denote the statistics computed within fold k. Then, for each fold, we construct a variance estimator $\hat{\varsigma}^2_{\hat{r}(\boldsymbol{w})k}$, and the K-fold cross-validation criterion is

$$CV_K(\boldsymbol{w}) = \frac{1}{K}\sum_{k=1}^{K} \hat{\varsigma}^2_{\hat{r}(\boldsymbol{w})k}.$$

Here $\hat{\varsigma}^2_{\hat{r}(\boldsymbol{w})k}$ is the plug-in within-stratum variance estimator computed from the transformed outcomes in the held-out fold I_k, using nuisance functions trained on I_k^c.

IV. Select the model weights by minimizing the K-fold cross-validation criterion:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} CV_K(\boldsymbol{w}),$$

where $\mathcal{W} = \{\boldsymbol{w} : \sum_{m=1}^{M} w_{(m)a} = 1, w_{(m)a} \geq 0, a = 0, 1\}$. The final treatment effect estimator is

$$\begin{aligned}\hat{\tau} \;=\; \sum_{s=1}^{S} p_{n[s]} \Bigg[&\left\{ \bar{Y}_{[s]1} - \frac{1}{n_{[s]1}} \sum_{i\in\mathcal{J}_{[s]}} (A_i - \pi_{n[s]}) \sum_{m=1}^{M} \hat{w}_{(m)1} \sum_{k=1}^{K} \mathbb{1}_{[i\in I_k]} \hat{h}^{(-k)}_{m[s]}(\boldsymbol{X}_i, 1) \right\} \\ &- \left\{ \bar{Y}_{[s]0} + \frac{1}{n_{[s]0}} \sum_{i\in\mathcal{J}_{[s]}} (A_i - \pi_{n[s]}) \sum_{m=1}^{M} \hat{w}_{(m)0} \sum_{k=1}^{K} \mathbb{1}_{[i\in I_k]} \hat{h}^{(-k)}_{m[s]}(\boldsymbol{X}_i, 0) \right\} \Bigg].\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Remark 2. The transformed outcomes used in the proposed treatment effect estimator are designed to improve estimation precision by subtracting the variation in potential outcomes that can be explained by baseline covariates. In this sense, the proposed estimator can be conceptualized as a model-averaged and cross-fitted extension of the covariate-adjusted estimators developed by Liu et al. (2023) and Tu et al. (2024). In addition, we retain the out-of-fold predictions in the final estimator to preserve the fold-level separation between nuisance fitting and treatment effect estimation. This is analogous to the role of cross-fitting in double machine learning (Chernozhukov et al., 2018). A full-sample refit of $h_{m[s]}$ may improve prediction accuracy, but it defines a different estimator that requires further conditions for valid inference.

Remark 3. The performance of the cross-validation procedure can be sensitive to the choice of K, reflecting a trade-off between nuisance fitting and validation precision. A larger K uses more observations for fitting the candidate models, but leaves fewer observations in each validation fold, which may make the foldwise variance estimates less stable, particularly when some stratum-treatment cells are small. A smaller K provides larger validation folds but uses fewer observations for model fitting. In practice, values around 5 are commonly adopted in the statistical literature.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

4. Optimality of the selected weights

In this section, we establish the optimality properties of the cross-validated weights proposed in Section 3 under standard regularity conditions. The decomposition in the Supplementary Material shows that, under the conditions stated below, the proposed treatment effect estimator is asymptotically equivalent at the root-n scale to applying a stratified difference-in-means estimator to a transformed outcome of the form $Y_i(a) - \{(1-\pi)h_{[s]}(\boldsymbol{X}_i, 1) + \pi h_{[s]}(\boldsymbol{X}_i, 0)\}$. As shown in Liu et al. (2023), the asymptotic variance of the treatment effect estimator based on this transformed outcome is minimized when $h_{[s]}(\boldsymbol{X}_i, a) = E\{Y_i(a)|\boldsymbol{X}_i, B_i = s\}, s = 1, \ldots, S, a = 0, 1$. This result directly motivates our choice of the risk function. Specifically, the asymptotic variance of the ATE estimator depends critically on how accurately the weighted combination of candidate model predictions approximates the true conditional means. A more accurate approximation leads to a smaller asymptotic variance. Accordingly, within each stratum s, we define the risk as the variance of the discrepancy between the weighted

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

prediction and its optimal conditional mean target:

$$
\begin{aligned}
& R(\boldsymbol{w}, s) \\
= \; & \mathrm{Var}\Bigg(\Bigg\{ (1-\pi) \sum_{m=1}^{M} w_{(m)1} \sum_{k=1}^{K} \mathbb{1}_{[i \in I_k]} \hat{h}_{m[s]}^{(-k)}(\boldsymbol{X}_i, 1) + \pi \sum_{m=1}^{M} w_{(m)0} \sum_{k=1}^{K} \mathbb{1}_{[i \in I_k]} \hat{h}_{m[s]}^{(-k)}(\boldsymbol{X}_i, 0) \Bigg\} \\
& - \Big[(1-\pi) E\{Y_i(1) | \boldsymbol{X}_i, B_i = s\} + \pi E\{Y_i(0) | \boldsymbol{X}_i, B_i = s\} \Big] \Bigg),
\end{aligned}
$$

where the variance is taken with respect to $\boldsymbol{X}_i$. By minimizing this risk function over the weights, we find the weighted combination of candidate models that best approximates the true conditional means, which directly reduces the asymptotic variance of the ATE estimator and improves the efficiency of inference on the target estimand. The overall risk function is then scaled and aggregated across strata: $R(\boldsymbol{w}) = \{\pi(1-\pi)\}^{-1} \sum_{s=1}^{S} p_s R(\boldsymbol{w}, s)$, with p_s denoting the expected proportion of units in stratum s.

Ideally, one would select weights $\boldsymbol{w} \in \mathcal{W}$ to minimize $R(\boldsymbol{w})$. However, this objective is infeasible in practice, as it depends on the unknown conditional expectations. Instead, we adopt a data-driven approach and select weights by minimizing a K-fold cross-validation criterion. We show that the resulting empirical weights asymptotically minimize $R(\boldsymbol{w})$. This result fills an important gap in the existing literature. To the best of our knowledge, previous studies have not systematically investigated weight selection among multiple candidate models under covariate-adaptive randomization.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Establishing this result is non-trivial, as the dependence among observations induced by CAR requires careful theoretical treatment. Moreover, we demonstrate that the proposed method is not only practically implementable but also asymptotically optimal in large samples.

For each model $m \in \{1, \ldots, M\}$, stratum $s \in \{1, \ldots, S\}$, and treatment arm $a \in \{0, 1\}$, let $h^*_{m[s]}(\cdot, a)$ be the limiting function of $\hat{h}^{(-k)}_{m[s]}(\cdot, a)$. Suppose that subject $i \in I_k$ lies in stratum s, and let $\tilde{\boldsymbol{X}}_i \sim \boldsymbol{X}_i | B_i = s$ be an independent copy of $\boldsymbol{X}_i$ given $B_i = s$ that is also independent of the out-of-fold observations $\{Y_j, \boldsymbol{X}_j, A_j, B_j\}_{j \in I_k^c}$. We next state the assumptions required for establishing these asymptotic results.

Assumption 4. Conditional on the training-fold data used to construct $\hat{h}^{(-k)}_{m[s]}(\cdot, a)$, both $\hat{h}^{(-k)}_{m[s]}(\tilde{\boldsymbol{X}}_i, a)$ and $h^*_{m[s]}(\tilde{\boldsymbol{X}}_i, a)$ have bounded second moments for all $m \in \{1, \ldots, M\}, s \in \{1, \ldots, S\}$, and $a = 0, 1$.

Assumption 5. Let $(I_k)_{k=1}^K$ denote a random K-fold partition of the index set $\{1, \ldots, n\}$, and let I_k^c denote the complement of I_k. Then the following condition holds uniformly over m, s and a:

$$E\Big[\big\{\hat{h}^{(-k)}_{m[s]}(\tilde{\boldsymbol{X}}_i, a) - h^*_{m[s]}(\tilde{\boldsymbol{X}}_i, a)\big\}^2 \mid B_i = s, \{Y_j, \boldsymbol{X}_j, A_j, B_j\}_{j \in I_k^c}\Big] = o_P(1), \quad (4.1)$$

where the conditional expectation is taken with respect to $\tilde{\boldsymbol{X}}_i$. The estimated function $\hat{h}^{(-k)}_{m[s]}(\cdot, a)$ is treated as fixed given $\{Y_j, \boldsymbol{X}_j, A_j, B_j\}_{j \in I_k^c}$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

the stratum indicator $B_i = s$. In addition, the conditional squared-error risks in (4.1) are uniformly integrable.

Remark 4. Assumption 5 imposes uniform L_2 consistency of the out-of-fold projection estimators $\hat{h}^{(-k)}_{m[s]}(\cdot, a)$ toward their population limits $h^*_{m[s]}(\cdot, a)$. This condition controls the estimation error of the cross-validated predictions and ensures that the resulting risk estimates are stable across folds, which is essential for valid weight selection in model averaging. This assumption is satisfied by a broad class of estimators commonly used in practice. With low-dimensional covariates, one may take $\hat{h}$ to be the ordinary least-squares or maximum-likelihood fitted regression function; under standard regularity conditions, the squared error is $O_p(n^{-1})$ (Van der Vaart, 2000). With high-dimensional covariates, if the regression function is assumed to be sparse and approximately linear, one may take $\hat{h}$ to be the Lasso-type fitted function. Under a restricted eigenvalue condition, the squared error is then $O_p(s \log p/n)$ (Bickel et al., 2009). If the covariate-outcome relationship is assumed to be possibly nonlinear, one may take $\hat{h}$ to be a function estimated using a random forest, and the squared error is $o_p(1)$ under suitable regularity conditions (Chi et al., 2022). Since each training sample I_k^c has size $n(1 - 1/K)$ and K is fixed, these rates have the same order as their full-sample counterparts. Moreover, because the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

numbers of candidate models, strata, and treatment arms are finite, taking the supremum over (m, s, a) does not change the convergence order.

For any $\boldsymbol{w} \in \mathcal{W}$ and $B_i = s$, define the common oracle adjustment and the corresponding oracle transformed outcomes by

$$q^*_{\boldsymbol{w}}(\boldsymbol{X}_i, s) = (1-\pi)\sum_{m=1}^{M} w_{(m)1} h^*_{m[s]}(\boldsymbol{X}_i, 1) + \pi \sum_{m=1}^{M} w_{(m)0} h^*_{m[s]}(\boldsymbol{X}_i, 0),$$

$$r^*_i(\boldsymbol{w}, a) = Y_i(a) - q^*_{\boldsymbol{w}}(\boldsymbol{X}_i, B_i), \qquad a = 0, 1.$$

The two variance components in Proposition 1, evaluated at these outcomes, are denoted by $\varsigma^2_{r^*(\boldsymbol{w})}(\pi)$ and $\varsigma^2_{Hr^*(\boldsymbol{w})}$, respectively.

Denote

$$\begin{aligned} R^*(\boldsymbol{w}, s) \;=\; & \mathrm{Var}\Bigg(\Big\{(1-\pi)\sum_{m=1}^{M} w_{(m)1} h^*_{m[s]}(\boldsymbol{X}_i, 1) + \pi \sum_{m=1}^{M} w_{(m)0} h^*_{m[s]}(\boldsymbol{X}_i, 0)\Big\} \\ & - \Big[(1-\pi)E\{Y_i(1)|\boldsymbol{X}_i, B_i = s\} + \pi E\{Y_i(0)|\boldsymbol{X}_i, B_i = s\}\Big]\Bigg), \end{aligned}$$

and define $R^*(\boldsymbol{w}) = \{\pi(1-\pi)\}^{-1}\sum_{s=1}^{S} p_s R^*(\boldsymbol{w}, s)$ as the population risk function obtained by replacing the estimated projection functions with their corresponding limiting functions.

Let $\xi = \inf_{\boldsymbol{w}\in\mathcal{W}} R^*(\boldsymbol{w})$ denote the minimum risk over the candidate weight set $\mathcal{W}$. A direct expansion shows that, for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$,

$$\left\{\varsigma^2_{r^*(\boldsymbol{w})}(\pi) + \varsigma^2_{Hr^*(\boldsymbol{w})}\right\} - \left\{\varsigma^2_{r^*(\boldsymbol{w}')}(\pi) + \varsigma^2_{Hr^*(\boldsymbol{w}')}\right\} = R^*(\boldsymbol{w}) - R^*(\boldsymbol{w}').$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Thus, minimizing $R^*(\boldsymbol{w})$ over $\mathcal{W}$ is equivalent to minimizing the asymptotic variance over the same candidate weight set.

We impose the following assumption, which rules out the degenerate case where the optimal conditional expectation is exactly represented by the candidate model class.

Assumption 6. The minimum population risk ξ is strictly positive.

Under the above assumptions, we can establish the following theorem, which shows that the K-fold cross-validation weights asymptotically minimize the population risk function.

Theorem 1. *Under Assumptions 1–6, it holds that*

$$\frac{R(\hat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w})} \xrightarrow{P} 1.$$

Remark 5. The proof of Theorem 1 relies on several technical ingredients. In particular, Lemma S3 establishes the uniform integrability of convex combinations of the projection functions $h(\cdot)$. Building on this result, Lemma S4 derives a uniform law of large numbers. These two lemmas together allow us to verify the conditions of Lemma 1 in Gao et al. (2019), which requires uniform convergence with respect to the weight vector $\boldsymbol{w}$.

Next, we consider the special case where $\xi = 0$. In this setting, there ex-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

ists a convex combination of candidate models that achieves the optimal covariate adjustment, in the sense that $\inf_{\boldsymbol{w}\in\mathcal{W}} R^*(\boldsymbol{w}, s) = 0$ for every stratum s.

We treat this as a supplementary case and establish the following result.

Theorem 2. *Under Assumptions 1–5, if $\xi = 0$, then $R(\hat{\boldsymbol{w}}) = o_P(1)$.*

Theorem 2 shows that, when the optimal covariate adjustment lies in the convex hull of the candidate models, the cross-validated weights consistently recover such an optimal combination, leading to vanishing risk.

5. Inference with the model averaging estimator

After establishing the asymptotic optimality of our CV-selected weights, we now turn to statistical inference for the proposed estimator. Valid inference for model averaging estimators under covariate-adaptive randomization is complicated by the data-dependent CV-selected weights and the dependence induced by CAR. We address these challenges by establishing asymptotic normality of the proposed estimator, which supports asymptotically valid confidence intervals and hypothesis tests.

In the theorem below, $r_i^*(\boldsymbol{w}^*, a)$ denotes the oracle transformed outcome induced by an oracle risk minimizer $\boldsymbol{w}^*$, and $\varsigma^2_{r^*(\boldsymbol{w}^*)}(\pi)$ and $\varsigma^2_{Hr^*(\boldsymbol{w}^*)}$ are the corresponding variance components that account for this weight-dependent transformed outcome.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Theorem 3. *Let $\boldsymbol{w}^*$ be the optimal weight vector and $\hat{\boldsymbol{w}}$ be the estimated weights obtained from the proposed model averaging procedure. Suppose that $r_i^*(\boldsymbol{w}^*, a) \in \mathcal{R}_2$. Under Assumptions 1–5, we have*

$$\sqrt{n}\{\hat{\tau}(\hat{\boldsymbol{w}})-\tau\} \xrightarrow{d} \mathcal{N}\big(0, \varsigma^2_{r^*(\boldsymbol{w}^*)}(\pi)+\varsigma^2_{Hr^*(\boldsymbol{w}^*)}\big),\ \hat{\varsigma}^2_{r(\hat{\boldsymbol{w}})}(\pi)+\hat{\varsigma}^2_{Hr(\hat{\boldsymbol{w}})} \xrightarrow{P} \varsigma^2_{r^*(\boldsymbol{w}^*)}(\pi)+\varsigma^2_{Hr^*(\boldsymbol{w}^*)}.$$

Remark 6. The estimation error of $\hat{h}^{(-k)}_{m[s]}$ is absorbed through Assumption 5. In the proof of Theorem 3 in the Supplementary Material, the first two terms of Eq. (S3.1) are considered together as a paired CAR imbalance contrast of the fitted adjustment, conditional on the training folds. The accuracy of $\hat{h}^{(-k)}$ is used when the leading plug-in transformed outcome is replaced by its oracle version based on $h^*_{m[s]}$. Under Assumption 5, candidate learners may converge at different rates, and the cross-validation criterion selects the asymptotically optimal convex combination for the risk objective among the candidate adjustments.

Remark 7. When the candidate set contains only one model, the proposed estimator reduces to the standard single-model covariate-adjusted estimator, making Proposition 1 a special case of Theorem 3 with the same limiting distribution. This single-model constraint differs from model selection, where multiple candidates are available but the final estimator is still restricted to a single selected model. By contrast, model averaging allows data-driven convex combinations of the candidate models, which

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

can be asymptotically preferable (Le and Clarke, 2022). The model averaging weights directly affect the first-order asymptotic variance and are selected to identify the optimal combination over the entire weight space, whereas model selection is restricted to its vertices. Theorem 1 establishes the asymptotic optimality of the cross-validation weights over this weight space. Consequently, the proposed model averaging estimator is asymptotically no less efficient than model selection and can be strictly more efficient when a combination of candidate models approximates the optimal conditional mean functions more accurately than any individual model.

Based on the above asymptotic normality, we derive a model averaging framework for valid statistical inference on the ATE τ that permits hypothesis testing and the construction of confidence intervals at the nominal level $\alpha \in (0,1)$. To test the null hypothesis $H_0 : \tau = \tau_0$ against the two-sided alternative $H_1 : \tau \neq \tau_0$, we consider a Wald-type test based on the standardized estimator:

$$\phi(Z^n) = \mathbb{1}_{\left[\left|\frac{\hat{\tau}(\hat{\boldsymbol{w}})-\tau_0}{\sqrt{\{\hat{\varsigma}^2_{r(\hat{\boldsymbol{w}})}(\pi)+\hat{\varsigma}^2_{Hr(\hat{\boldsymbol{w}})}\}/n}}\right| > z_{1-\frac{\alpha}{2}}\right]},$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of a standard normal random variable, and $Z^n = \frac{\hat{\tau}(\hat{\boldsymbol{w}})-\tau_0}{\sqrt{\{\hat{\varsigma}^2_{r(\hat{\boldsymbol{w}})}(\pi)+\hat{\varsigma}^2_{Hr(\hat{\boldsymbol{w}})}\}/n}}$ is the test statistic. While this procedure relies on large-sample approximations, simulations in the next section demonstrate

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

good finite-sample performance.

6. Numerical experiments

6.1 Simulation studies

In this section, we examine the empirical performance of estimators with different candidate models. The detailed data-generating process can be found in the Supplementary Material. For covariate adjustment, we employ four candidate models: linear regression (linear), local linear kernel regression (kernel), natural spline (ns), and random forest (rf). We also perform model averaging (ma) and model selection (ms) based on these four models. We conducted simulation studies under various randomization schemes, including complete randomization, stratified block randomization, Pocock and Simon's minimization, and Hu and Hu's method. For brevity, we present here only the results obtained under Pocock and Simon's minimization with a biased-coin probability of 0.75 and equal marginal weights. Additional results for other randomization procedures are provided in the Supplementary Material. They suggest that the asymptotic properties of the treatment effect estimators are insensitive to the specific choice of randomization scheme. The relative bias, standard deviation (SD) of the treatment effect estimators, and empirical coverage probability (CP) of the 95%

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

confidence interval are computed using $1,000$ replications. To determine the number of folds in the subsequent simulations, we first examine changes in the standard deviations of the treatment effect estimators for $K = 2, 5, 10$, with the sample size fixed at $n = 500$.

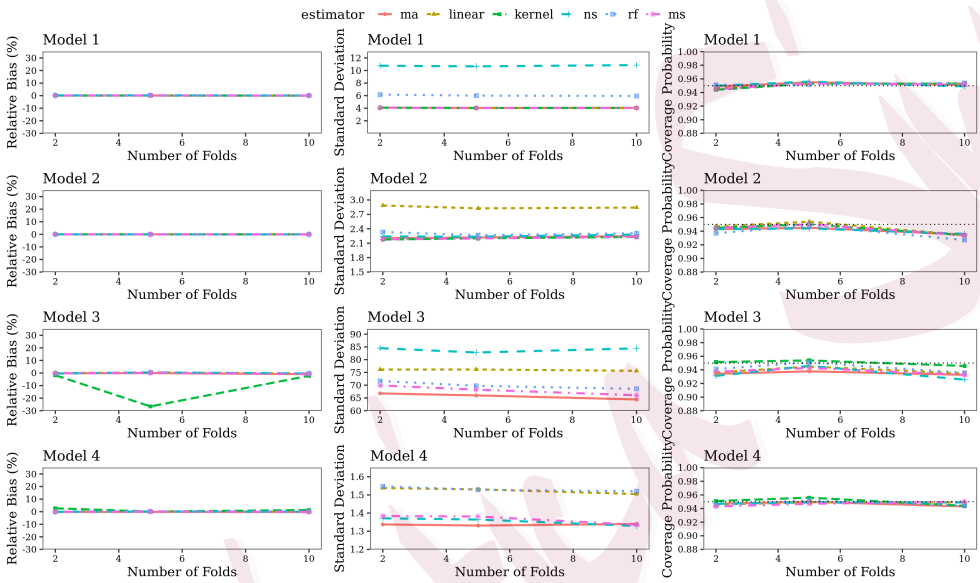

Figure 1: Standard deviations of the treatment effect estimators with different numbers of folds under each model.

As illustrated in Figure 1, the performance of the treatment effect estimators is generally stable across the considered values of K. Increasing K from 2 to 5 produces modest reductions in standard deviation in some settings, whereas further increasing K to 10 yields little systematic improvement and occasionally leads to slight increases. The relative biases

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

and coverage probabilities are also broadly comparable across values of K, except for the instability of the local linear kernel-adjusted estimator under Model 3. In view of its generally stable performance and moderate computational cost, we use $K = 5$ in the subsequent simulation studies.

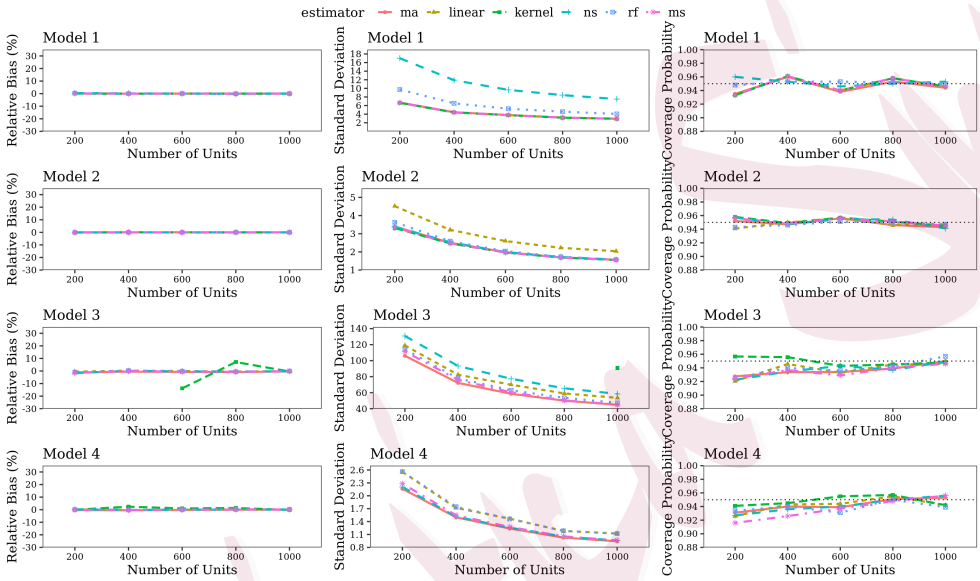

Figure 2: Relative biases, standard deviations, and coverage probabilities of the treatment effect estimators with $K = 5$ and different numbers of units under each model.

Figure 2 shows that the relative biases of most treatment effect estimators are negligible across the scenarios considered. The main exception is the local linear kernel-adjusted estimator under Model 3, which exhibits unstable relative bias with small sample sizes because sparse observations

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

make it difficult to estimate the highly nonlinear relationship involving X_1 in the control group. The proposed model averaging estimator generally has the smallest standard deviations among the estimators considered, with its efficiency advantage being most pronounced under Model 3, where no single candidate model is correctly specified. Across all settings, the standard deviations decrease as the sample size increases, while the coverage probabilities remain reasonably close to the nominal 95% level.

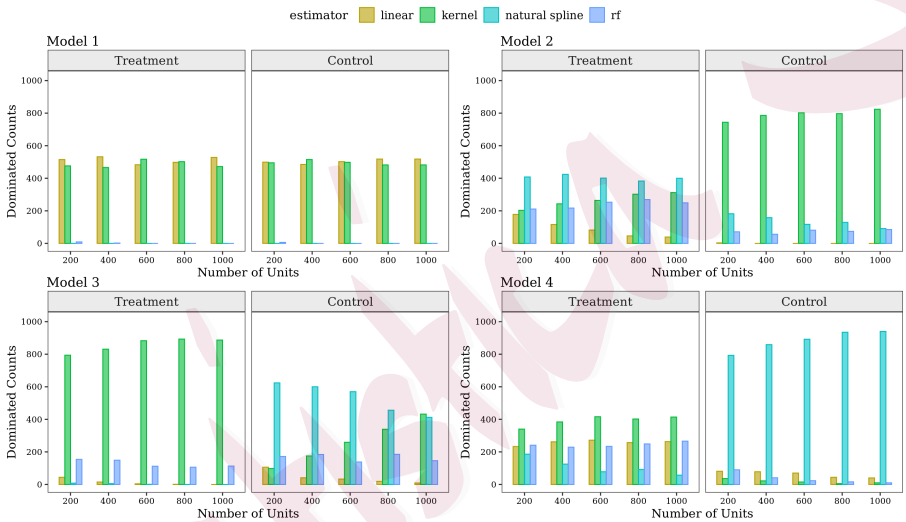

Figure 3: Dominance counts for each adjustment method across 1000 simulation replications with different numbers of units under each model.

Figure 3 summarizes how often each candidate adjustment method dominates across simulation replications. The dominant method varies substantially across data-generating models, treatment arms, and sample

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

sizes, indicating that no single adjustment method is uniformly optimal. In Model 1, the linear and kernel adjustments dominate most frequently in both treatment and control groups, consistent with the relatively simple outcome structure. In Model 2, the dominant method differs sharply by treatment arm: natural spline, kernel, and random forest all contribute in the treatment group, whereas the kernel method overwhelmingly dominates in the control group. Model 3 shows an even stronger arm-specific pattern, with kernel adjustment dominating in the treatment group, while natural spline dominates the control group for smaller sample sizes and kernel becomes increasingly competitive as the sample size grows. In Model 4, the natural spline dominates in the control group, whereas the treatment group exhibits a more mixed pattern. These results highlight substantial model uncertainty and treatment-arm heterogeneity and provide empirical support for the proposed model averaging approach, which efficiently combines candidate adjustments rather than relying on a single model.

6.2 Clinical trial example

In this section, we consider a clinical trial example based on synthetic data generated from a chronic depression trial, which was conducted to compare the efficacy of three treatments (Keller et al., 2000). We focus on two of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

the treatments, Nefazodone alone and the combination of Nefazodone and the cognitive behavioral-analysis system of psychotherapy (CBASP). The total number of subjects is 440. To generate the synthetic data, we first fit a nonparametric spline with the function `bigssa` in the R package `bigspline` and the selected covariates detailed in Table S17 of the Supplementary Material. The response of interest is the total Hamilton anxiety score (HAMA score), with a lower score indicating less severe anxiety.

Then, we use gender for stratification and generate the treatment assignments using different randomization procedures. Once the treatment assignments are produced, we generate subject responses through the fitted model. For the data analysis, each model uses a different set of covariates for treatment effect adjustment; details are provided in the Supplementary Material, Table S18. Here, we present the treatment effect estimates and variance reductions of the same covariate-adjusted estimators as those used in the simulation study under stratified block randomization, using the stratified difference-in-means estimator as the reference.

First, all estimators considered show that patients receiving the combination of Nefazodone and CBASP have lower HAMA scores than those receiving Nefazodone alone, suggesting a benefit of the combined therapy. Moreover, all covariate-adjusted treatment effect estimators achieve sub-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Table 1: Performance of covariate-adjusted estimators in the synthetic depression trial data: treatment effect estimates and variance reductions.

Allocation	Statistics	sdim	linear	kernel	ns	rf	ma
Equal ($\pi = \frac{1}{2}$)	Estimates	-3.55	-3.33	-3.46	-3.35	-3.41	-3.39
	Variance Reduction	–	7.60%	9.08%	8.45%	8.94%	12.29%
Unequal ($\pi = \frac{2}{3}$)	Estimates	-3.57	-3.53	-3.53	-3.51	-3.73	-3.58
	Variance Reduction	–	6.07%	4.10%	4.75%	9.40%	11.02%

Abbreviation: sdim, stratified difference-in-means; ns, natural spline; rf, random forest; ma, model averaging.

stantial efficiency gains relative to the unadjusted stratified difference-in-means estimator. Under equal allocation, the kernel-adjusted estimator achieves the largest variance reduction among the single-model approaches, whereas under unequal allocation ($\pi = \frac{2}{3}$), the random forest-adjusted estimator yields the greatest efficiency gain. In both settings, however, the model averaging estimator attains the smallest variance, illustrating the potential efficiency gain from model averaging.

7. Conclusion

This paper proposes a novel model averaging framework for covariate adjustment under covariate-adaptive randomization, aiming to reconcile reg-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

ulatory requirements for pre-specifying analysis models with the practical difficulty of selecting a single optimal adjustment strategy. The approach employs K-fold cross-validation to compute data-driven weights across a pre-defined set of candidate models, yielding robust and efficient treatment effect estimates. We further establish the asymptotic normality of the proposed estimator and derive a consistent variance estimator, ensuring valid statistical inference under a broad class of covariate-adaptive randomization procedures.

Simulation studies and an application to synthetic clinical trial data demonstrate the favorable performance of our model averaging estimator. Across all simulated scenarios and the clinical trial application, it generally attains the lowest or near-lowest variance, particularly under complex model misspecification where no single candidate is correct, while finite-sample bias diminishes as the sample size increases. In practice, the choice of fold partition is not expected to substantially affect the efficiency of the treatment effect estimator, provided that each training sample remains sufficiently large for nuisance estimation. Our simulations suggest that even when the sample size within each stratum-treatment cell is only approximately 30–50, the proposed estimator can still perform well. Moreover, fold splitting is also required by many machine learning methods to ensure valid

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

inference, so the use of model averaging does not introduce an additional validity burden; rather, it mainly provides an opportunity for efficiency improvement by combining candidate estimators. The main practical limitation is therefore the convergence behavior of the candidate nuisance models. When the sample size is extremely small, machine learning methods may fail to converge sufficiently fast. In such cases, we recommend averaging across faster-converging parametric models, such as linear regression models with different covariate specifications. Importantly, the theoretical validity of the proposed estimator follows under the stated consistency condition on the candidate learners entering the averaging procedure, while the model averaging criterion selects the asymptotically optimal convex combination for the variance-risk objective.

The proposed framework also yields interpretable information through the estimated weights. These weights summarize the relative contribution of candidate models to outcome prediction in the observed data. This post hoc information may help inform future trial designs or refine covariate adjustment strategies in subsequent studies, while remaining compatible with pre-specification requirements. In summary, this work provides a principled and regulatory-compliant approach to covariate adjustment that accounts for model uncertainty. The framework can also be extended to other out-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

come types, including binary and time-to-event outcomes, thereby broadening its applicability in clinical research.

Supplementary Materials

The online supplementary material contains proofs of the main theorems, additional technical lemmas, detailed simulation settings and results, and further details of the clinical trial example.

Acknowledgements

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), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (Grant No. 202630333). Fuyi Tu was supported by the Doctoral Research Start-up Funding of CQUPT (Grant No. E012A2024010). Wei Ma is the corresponding author.

References

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

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.

Cerreia-Vioglio, S., L. P. Hansen, F. Maccheroni, and M. Marinacci (2025). Making decisions under model misspecification. *Review of Economic Studies*, rdaf046.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chi, C.-M., P. Vossler, Y. Fan, and J. Lv (2022). Asymptotic properties of high-dimensional random forests. *The Annals of Statistics 50*(6), 3415–3438.

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

EMA (2015). Guideline on adjustment for baseline covariates in clinical trials. *European Medicines Agency*.

Fang, F., J. Li, and X. Xia (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics 229*(2), 219–245.

FDA (2023). Adjusting for covariates in randomized clinical trials for drugs and biological products. *U.S. Food and Drug Administration*.

Gao, Y., X. Zhang, S. Wang, T. T. L. Chong, and G. Zou (2019). Frequentist model averaging

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

for threshold models. *Annals of the Institute of Statistical Mathematics 71*(2), 275–306.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Holmberg, M. J. and L. W. Andersen (2022). Adjustment for baseline characteristics in randomized clinical trials. *JAMA: The Journal of the American Medical Association 328*(21), 2155–2156.

Hu, Y. and F. Hu (2012). Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics 40*(3), 1794–1815.

ICH E9 (1998). Statistical principles for clinical trials. *International Conference on Harmonisation*.

ICH E9 (R1) (2017). Addendum on estimands and sensitivity analysis in clinical trials. to the guideline on statistical principles for clinical trials. *Clinical Trials*.

Keller, M. B., J. P. McCullough, D. N. Klein, B. Arnow, D. L. Dunner, A. J. Gelenberg, J. C. Markowitz, C. B. Nemeroff, J. M. Russell, M. E. Thase, 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.

Le, T. M. and B. S. Clarke (2022). Model averaging is asymptotically better than model selection

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

for prediction. *Journal of Machine Learning Research 23*(33), 1–53.

Li, J., J. Lv, A. T. Wan, and J. Liao (2022). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association 117*(537), 495–509.

Li, N., Y. Fei, Y. Yang, and X. Zhang (2025). Averaging estimators of heterogeneous treatment effects under additive models. *Econometric Theory*, 1–34.

Liao, J., X. Zong, X. Zhang, and G. Zou (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics 209*(1), 35–60.

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.

Morris, T. P., A. S. Walker, E. J. Williamson, and I. R. White (2022). Planning a method for covariate adjustment in individually randomised trials: a practical guide. *Trials 23*(1), 328.

Pocock, S. J., S. E. Assmann, L. E. Enos, and L. E. Kasten (2002). Subgroup analysis, covari-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

ate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine 21*(19), 2917–2930.

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

Rafi, A. (2023). Efficient semiparametric estimation of average treatment effects under covariate adaptive randomization. *arXiv preprint arXiv:2305.08340*.

Robinson, L. D. and N. P. Jewell (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 227–240.

Rolling, C. A., Y. Yang, and D. Velez (2019). Combining estimates of conditional treatment effects. *Econometric Theory 35*(6), 1089–1110.

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

Rosenblum, M. and M. J. Van Der Laan (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics 6*(1), 13.

Singh, A., S. Mishra, and G. Ruskauff (2010). Model averaging techniques for quantifying conceptual model uncertainty. *Groundwater 48*(5), 701–715.

Tsiatis, A. A., M. Davidian, M. Zhang, and X. Lu (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Statistics in Medicine 27(23), 4658–4677.

Tu, F., W. Ma, and H. Liu (2024). A unified framework for covariate adjustment under stratified randomisation. *Stat 13*(4), e70016.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics 156*(2), 277–283.

Wei, L. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association 73*(363), 559–563.

Yang, L. and A. A. Tsiatis (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician 55*(4), 314–321.

Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis 74*(1), 135–161.

Ye, T., J. Shao, Y. Yi, and Q. Zhao (2023). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association 118*(544), 2370–2382.

Ye, T., Y. Yi, and J. Shao (2022). Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika 109*(1), 33–47.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0502

Zhang, X., D. Yu, G. Zou, and H. Liang (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association 111*(516), 1775–1790.

Zhang, X., G. Zou, and R. J. Carroll (2015). Model averaging based on kullback-leibler distance. *Statistica Sinica 25*, 1583.

Zhao, Z., X. Zhang, G. Zou, A. T. Wan, and G. K. Tso (2024). Model averaging for estimating treatment effects. *Annals of the Institute of Statistical Mathematics 76*(1), 73–92.

Fuyi Tu
School of Mathematics and Statistics, Chongqing University of Posts and Telecommunications
E-mail: tufy@cqupt.edu.cn

Jiahui Xin
Institute of Statistics and Big Data, Renmin University of China
E-mail: xinjh@ruc.edu.cn

Wei Ma
Institute of Statistics and Big Data, Renmin University of China
E-mail: mawei@ruc.edu.cn