Biometrika (2022), **109**, 1, *pp.* 33–47
Printed in Great Britain
doi: 10.1093/biomet/asab015
Advance Access publication 3 June 2021

Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods

BY TING YE

Department of Statistics, University of Pennsylvania,
400 Huntsmen Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, U.S.A.
tingye@wharton.upenn.edu

YANYAO YI

Global Statistical Sciences, Eli Lilly and Company,
893 Delaware Street, Indianapolis, Indiana 46285, U.S.A.
yi_yanyao@lilly.com

AND JUN SHAO

School of Statistics, East China Normal University, Shanghai 200241, China
shao@stat.wisc.edu

SUMMARY

Covariate-adaptive randomization schemes such as minimization and stratified permuted blocks are often applied in clinical trials to balance treatment assignments across prognostic factors. The existing theory for inference after covariate-adaptive randomization is mostly limited to situations where a correct model between the response and covariates can be specified or the randomization method has well-understood properties. Based on stratification with covariate levels utilized in randomization and a further adjustment for covariates not used in randomization, we propose several model-free estimators of the average treatment effect. We establish the asymptotic normality of the proposed estimators under all popular covariate-adaptive randomization schemes, including the minimization method, and we show that the asymptotic distributions are invariant with respect to covariate-adaptive randomization methods. Consistent variance estimators are constructed for asymptotic inference. Asymptotic relative efficiencies and finite-sample properties of estimators are also studied. We recommend using one of our proposed estimators for valid and model-free inference after covariate-adaptive randomization.

Some key words: Balancing of treatment assignments; Covariate adjustment; Efficiency; Generalized regression; Model-free inference; Multiple treatment arms; Stratification; Variance estimation.

1. INTRODUCTION

Consider a clinical trial to compare k treatments with given treatment assignment proportions $\pi_1, \ldots, \pi_k$, where $k \geqslant 2$ is a fixed positive integer, $\sum_{t=1}^{k} \pi_t = 1$, and π_t can be any known number strictly between 0 and 1. In many trials patients are not all available for simultaneous assignment of treatments, but rather arrive sequentially and must be treated immediately. Thus,

simple randomization, which assigns patients to treatments completely at random, may yield sample sizes not following the assignment proportions across prognostic factors or covariates, e.g., institution, disease stage, prior treatment, gender and age, which are thought to have significant influence on the responses of interest. A remedy is to apply covariate-adaptive randomization, i.e., the treatment assignment of the ith patient depends on the observed covariate value of this patient and the assignments and covariate values of all $i-1$ previously assigned patients. In this article we focus on enforcing assignment allocation across levels of a covariate vector Z whose components are discrete or discretized continuous covariates. There are model-based approaches of balancing discrete or continuous covariates for estimation efficiency (Atkinson, 1982, 1999, 2002; Rosenberger & Sverdlov, 2008; Senn et al., 2010; Baldi Antognini & Zagoraiou, 2011), which are not further considered in this article. The oldest method of balancing covariates is the minimization method (Taves, 1974) intended to balance treatment assignments across marginal levels of Z: it assigns the ith patient by minimizing a weighted sum of squared or absolute differences between the numbers of patients, up to the ith, assigned to treatments over marginal levels of Z. Pocock & Simon (1975) extended Taves' procedure to achieve minimization with a given probability, which is still referred to as the minimization method. Other popular covariate-adaptive randomization methods include stratified permuted block randomization (Zelen, 1974), the stratified biased coin (Shao et al., 2010; Kuznetsova & Johnson, 2017) and the stratified urn design (Wei, 1977; Zhao & Ramakrishnan, 2016). For nice summaries, see Schulz & Grimes (2002) and Rosenberger & Sverdlov (2008). As pointed out in Taves (2010), from 1989 to 2008, over 500 clinical trials implemented the minimization method to balance important covariates, despite some criticisms by Smith (1984) and Senn et al. (2010). According to a recent review of nearly 300 clinical trials published in 2009 and 2014 (Ciolino et al., 2019), 237 of them used covariate-adaptive randomization.

Although data are collected under covariate-adaptive randomization, conventional inference procedures constructed based on simple randomization are often applied in practice. This has raised concerns because statistical inference on treatment effects should be made using procedures valid under the particular randomization scheme used in data collection. Applying conventional inference after covariate-adaptive randomization may lead to invalid results, especially when the minimization method is used, because its theoretical property remains largely unclear. The European Medicines Agency (2015) guidelines raised concerns and specifically pointed out that

> possible implications of dynamic allocation methods [minimization] on the analysis, e.g., with regard to bias and Type I error control should be carefully considered, …conventional statistical methods do not always control the Type I error.

Starting with Shao et al. (2010), there has been significant progress in understanding the theoretical properties of statistical tests under covariate-adaptive randomization, e.g., Hu & Hu (2012), Shao & Yu (2013), Ma et al. (2015), Bugni et al. (2018, 2019), Ye (2018) and Ye & Shao (2019). Another important stream of statistical inference methods is based on permutation tests or rerandomization inference, e.g., Simon & Simon (2011), Kaiser (2012) and Bugni et al. (2018). However, except for Bugni et al. (2019), all theoretical results are established under the assumption that either a correct model between the responses of interest and covariates is available or the covariate-adaptive randomization procedure has a well-understood property; these are described as type 1 or type 2 in § 2 of the current paper. Model misspecification often occurs, especially when there are many covariates, and the minimization method is neither type 1 nor type 2. The minimization method is applied very often in practice (Pocock & Simon, 1975), mainly because it aims to minimize the imbalance across marginal levels of Z, not every joint level of Z, which is

sufficient in many applications. Enforcing treatment balance in every joint level of Z may cause sparsity of data when the dimension of Z is not small.

To fill the gap, in this paper we propose asymptotically valid inference on the average treatment effect defined as the difference between population response means of every treatment pair, under covariate-adaptive randomization, including minimization. Our main idea is to apply stratification according to the levels of discrete Z, and to adjust for covariates not used in treatment randomization through generalized regression. Our estimator without adjusting for covariates, which is not the most efficient one, coincides with the estimator derived under a different approach in Bugni et al. (2019). Asymptotic normality of the proposed treatment effect estimators is established with explicit limiting variance formulae that can be used for inference as well as comparing relative efficiencies. Our results are not only model free, i.e., only the existence of second-order moments of the responses and covariates are required, but also invariant with respect to covariate-adaptive randomization schemes, i.e., the same inference procedure can be applied under any covariate-adaptive randomization. We also study and compare inference procedures by simulations, and illustrate our method in a real-data example.

2. Preliminaries

Let I be the treatment indicator vector that equals e_t if treatment t is assigned, where e_t denotes the k-dimensional vector whose tth component is 1 and whose other components are 0, $t = 1, \ldots, k$. Let $Y^{(t)}$ be the potential response under treatment t, W be a vector of all observed covariates, and Z be a discrete function of W utilized in covariate-adaptive randomization. For patient i, let I_i, W_i and $Y_i^{(t)}$ $(t = 1, \ldots, k)$ be realizations of I, W and $Y^{(t)}$ $(t = 1, \ldots, k)$ respectively, where $i = 1, \ldots, n$ and n is the total number of patients in all treatment arms. For every patient i, I_i is generated after Z_i is observed, and only the potential response from the treatment indicated by I_i is observed, i.e., we observe $Y_i = Y_i^{(t)}$ if and only if $I_i = e_t$.

After all treatment assignments are made and responses are collected, we would like to make inferences based on the observed data $\{W_i, I_i, Y_i, i = 1, \ldots, n\}$. For our inference procedure studied in § 3, we describe some minimal conditions. The first is about the population for potential responses and covariates.

Condition 1. We have that $(Y_i^{(1)}, \ldots, Y_i^{(k)}, W_i)$ $(i = 1, \ldots, n)$, are independent and identically distributed as $(Y^{(1)}, \ldots, Y^{(k)}, W)$ and $Y^{(t)}$ has finite second-order moment, $t = 1, \ldots, k$.

Condition 1 is model free as there is no assumption on the relationship between W and the potential response $Y^{(t)}$ that may be continuous or discrete.

Under simple randomization, the I_i are independent of the $(Y_i^{(1)}, \ldots, Y_i^{(k)}, W_i)$, and are independent and identically distributed with $\mathrm{pr}(I_i = e_t) = \pi_t$. To enforce assignment proportions at each joint level of Z treated as stratum, three popular covariate-adaptive randomization schemes are the stratified permuted block randomization method (Zelen, 1974), the stratified biased coin method (Shao et al., 2010; Kuznetsova & Johnson, 2017) and the stratified urn design (Wei, 1977; Zhao & Ramakrishnan, 2016).

The minimization method (Taves, 1974; Pocock & Simon, 1975; Han et al., 2009) is the same as the stratified biased coin method if Z is one-dimensional, but is very different from the above three stratification methods with a multivariate Z. It aims to enforce the assignment ratio across marginal levels of Z, not every stratum defined by the joint level of Z. Assignments are made by minimizing a weighted sum of squared or absolute differences between the numbers of patients

assigned to treatment arms across marginal levels of Z. Because only marginal levels of Z are considered in minimization, this method is also called the marginal method in Ma et al. (2015) and Ye & Shao (2019).

We assume the following minimal conditions for covariate-adaptive randomization.

Condition 2. We have that $(I_i, i = 1, \ldots, n)$ and $(Y_i^{(1)}, \ldots, Y_i^{(k)}, W_i, i = 1, \ldots, n)$ are conditionally independent given $Z_1, \ldots, Z_n$.

Condition 3. The covariate vector Z is discrete with finitely many levels given in a set $\mathcal{Z}$. For each $t = 1, \ldots, k$, $\text{pr}(I_i = e_t \mid Z_1, \ldots, Z_n) = \pi_t$ for $i = 1, \ldots, n$, and $\{n(z)\}^{-1} D_t(z)$ converges to 0 in probability as $n \to \infty$ for every $z \in \mathcal{Z}$, where $n(z)$ is the number of patients with $Z_i = z$, and $D_t(z) = n_t(z) - \pi_t n(z)$ with $n_t(z)$ being the number of patients with $Z_i = z$ assigned to treatment t.

Condition 2 is reasonable because given the Z_i, the W_i contain covariates not used in randomization, and treatment assignments do not affect the potential responses, although they do affect the observed responses Y_i. Condition 3 holds for most covariate-adaptive randomization schemes (Baldi Antognini & Zagoraiou, 2015), and certainly for all schemes considered in this paper, minimization, and three stratified designs: the permuted block, biased coin and urn designs.

We classify all covariate-adaptive randomization methods into the following three types in terms of $D_t(z)$ defined in Condition 3.

Type 1: For every t and z, $\{n(z)\}^{-1/2} D_t(z) \to 0$ in probability as $n \to \infty$.

Type 2: For every t, $D_t(z)$, $z \in \mathcal{Z}$, are independent and, for every t and z, $\{n(z)\}^{-1/2} D_t(z) \xrightarrow{d} N(0, v_t)$ with a known $v_t > 0$, where $\xrightarrow{d}$ denotes convergence in distribution as $n \to \infty$.

Type 3: Methods not in type 1 or 2.

The three types are defined based on their degree in enforcing the balancedness according to the given assignment proportions within every joint level of Z. Type 1 is the strongest, since $D_t(z)$ measures the imbalance of assignments within stratum z. The property $\{n(z)\}^{-1/2} D_t(z) \to 0$ in probability is stronger than $\{n(z)\}^{-1} D_t(z) \to 0$ in probability in Condition 3. Type 2 is weaker than type 1 in enforcing the balancedness, as it requires $\{n(z)\}^{-1/2} D_t(z)$ converging in distribution, not in probability, to 0, although it is still stronger than $\{n(z)\}^{-1} D_t(z) \to 0$ in probability.

Representatives of type 1 methods are the stratified permuted block randomization and stratified biased coin methods. Specifically, under stratified permuted block randomization, $D_t(z)$ is bounded by the maximum block size. For the stratified biased coin method, it follows from a result in Efron (1971) that $D_t(z)$ is bounded in probability. The stratified urn design is type 2 with $v_t = 1/12$ when $k = 2$ and $\pi_1 = \pi_2 = 1/2$ (Wei, 1978). Simple randomization treated as a special case of covariate-adaptive randomization is also type 2. Finally, the minimization method is type 3, since it is neither type 1 nor type 2 (Ye & Shao, 2019). Specifically, under minimization, $D_t(z)$ and $D_t(z')$ with $z \neq z'$ are not independent, and their relationship is complicated, because assignments are made according to marginal levels of Z.

For type 1 methods, some theoretical results in statistical testing have been established; see, for example, Shao et al. (2010), Shao & Yu (2013), Bugni et al. (2018, 2019), Ye (2018) and Ye & Shao (2019). Bugni et al. (2018, 2019) and Ye & Shao (2019) also considered type 2 methods. In the next section we propose inference procedures on average treatment effects, and establish their asymptotic validity under general covariate-adaptive randomization, including minimization.

3. Inference on the average treatment effect

In this paper we consider inference on the average treatment effect vector

$$\theta = (\theta_2, \ldots, \theta_k)^{\mathrm{T}}, \qquad \theta_t = E(Y^{(t)} - Y^{(1)}), \tag{1}$$

where a^{T} is the transpose of vector a and E denotes the population expectation. The average treatment effect for any two fixed treatment arms t and s can be obtained as $\theta_t - \theta_s$.

To make asymptotically valid inference on θ defined in (1), the key is to construct an estimator of θ and derive its asymptotic distribution. Under simple randomization, the simplest estimator of θ_t is the response mean difference $\bar{Y}_t - \bar{Y}_1$, where $\bar{Y}_t$ is the sample mean of responses under treatment $t = 1, \ldots, k$. Although $\bar{Y}_t - \bar{Y}_1$ is asymptotically normal under type 1 or 2 covariate-adaptive randomization, it is generally not efficient as covariate information is not utilized in estimation. More seriously, the asymptotic distribution of $\bar{Y}_t - \bar{Y}_1$ is not known under type 3 covariate-adaptive randomization such as the minimization method. Bugni et al. (2018, § 4.2) derived a different estimator of θ, called the strata fixed effect estimator, but its asymptotic normality is established only for type 1 or 2 covariate-adaptive randomization and, thus, it cannot be used under type 3 randomization such as the minimization method.

Let $\bar{Y}_t(z)$ be the sample mean of the Y_i with $Z_i = z$ under treatment $t = 1, \ldots, k$. The following stratified response mean differences with strata being all joint levels of Z is proposed in Bugni et al. (2019, (8)):

$$\hat{\theta} = (\hat{\theta}_2, \ldots, \hat{\theta}_k)^{\mathrm{T}}, \qquad \hat{\theta}_t = \sum_{z \in \mathcal{Z}} \frac{n(z)}{n} \{\bar{Y}_t(z) - \bar{Y}_1(z)\}, \tag{2}$$

although Bugni et al. (2019) provided $\hat{\theta}$ in a different form derived under a fully saturated linear regression. If the weight $n(z)/n$ in (2) is replaced by the population weight $\mathrm{pr}(Z = z)$, then $\hat{\theta}$ is exactly the stratified estimator in survey sampling. We use $n(z)/n$ in (2) as $\mathrm{pr}(Z = z)$ is unknown.

Although $\hat{\theta}_t$ in (2) utilizes information from Z by stratification and is asymptotically more efficient than the simple estimator $\bar{Y}_t - \bar{Y}_1$ or the strata fixed effect estimator in Bugni et al. (2018), it does not make use of covariate information in W, but not in Z. Note that W may contain components that are not in Z, but are related with the potential responses $Y^{(t)}$ $(t = 1, \ldots, k)$ or some components of Z are discretized components of W and the remaining information after discretization is still predictive of $Y^{(t)}$ $(t = 1, \ldots, k)$.

Let X be a function of W that we want to further adjust for. We now consider improving $\hat{\theta}$ in (2) by utilizing X. To maintain model-free estimation, we do not impose any model between $Y^{(t)}$ and X, but adjust for covariate X within each $Z = z$ by applying the generalized regression approach in survey sampling, first discussed in Cassel et al. (1976) and studied extensively in the literature (for example, Särndal et al., 2003; Lin, 2013; Shao & Wang, 2014; Ta et al., 2020). Since this approach is model assisted, but not model based, i.e., a model is used to derive efficient estimators that are still asymptotically valid even if the model is incorrect, it suits our purpose of utilizing covariates without modelling.

Let X_i be the value of covariate X for patient i, $\bar{X}_t(z)$ be the sample mean of the X_i with $Z_i = z$ under treatment t, $n_t(z)$ and $n(z)$ be defined as in Condition 3, and

$$\hat{\beta}_t(z) = \left[\sum_{t=1}^{k} \sum_{i:I_i = e_t, Z_i = z} \{X_i - \bar{X}_t(z)\}\{X_i - \bar{X}_t(z)\}^{\mathrm{T}} \right]^{-1} \frac{n(z)}{n_t(z)} \sum_{i:I_i = e_t, Z_i = z} \{X_i - \bar{X}_t(z)\} Y_i.$$

Within treatment t and $Z = z$, $\hat{\beta}_t(z)$ is a least-squares-type estimator of the coefficient vector in front of X under a linear model between $Y^{(t)}$ and X, but the model is not required to be correct. Then, our first proposed estimator of θ after adjusting for covariates is

$$\hat{\theta}_A = (\hat{\theta}_{2,A}, \ldots, \hat{\theta}_{k,A})^{\mathrm{T}}, \qquad \begin{aligned} \hat{\theta}_{t,A} &= \sum_{z \in \mathcal{Z}} \frac{n(z)}{n} \{\bar{Y}_{t,A}(z) - \bar{Y}_{1,A}(z)\}, \\ \bar{Y}_{t,A}(z) &= \bar{Y}_t(z) - \{\bar{X}_t(z) - \bar{X}(z)\}^{\mathrm{T}} \hat{\beta}_t(z), \end{aligned} \tag{3}$$

where $\bar{X}(z)$ is the sample mean of the X_i of all patients with $Z_i = z$.

Within $Z_i = z$, if we assume that the linear models under all treatments have the same coefficient vector for X, i.e., a homogeneous analysis of covariance model holds between the observed response and (X, I), then we can replace $\hat{\beta}_t(z)$ by

$$\hat{\beta}(z) = \left[\sum_{t=1}^{k} \sum_{i:I_i = e_t, Z_i = z} \{X_i - \bar{X}_t(z)\}\{X_i - \bar{X}_t(z)\}^{\mathrm{T}} \right]^{-1} \sum_{t=1}^{k} \sum_{i:I_i = e_t, Z_i = z} \{X_i - \bar{X}_t(z)\} Y_i,$$

which is in fact a weighted average of the $\hat{\beta}_t(z)$. Again, the model is not required to be correct in order to use $\hat{\beta}(z)$. This leads to an alternative estimator of θ after adjusting for covariates,

$$\hat{\theta}_B = (\hat{\theta}_{2,B}, \ldots, \hat{\theta}_{k,B})^{\mathrm{T}}, \qquad \begin{aligned} \hat{\theta}_{t,B} &= \sum_{z \in \mathcal{Z}} \frac{n(z)}{n} \{\bar{Y}_{t,B}(z) - \bar{Y}_{1,B}(z)\}, \\ \bar{Y}_{t,B}(z) &= \bar{Y}_t(z) - \{\bar{X}_t(z) - \bar{X}(z)\}^{\mathrm{T}} \hat{\beta}(z). \end{aligned} \tag{4}$$

When $k > 2$, $\bar{X}(z)$, $\hat{\beta}_t(z)$ and $\hat{\beta}(z)$ involve data from patients in treatment arms other than treatments t and 1.

The following theorem, proved in the Supplementary Material, derives the asymptotic distributions of $\hat{\theta}_A$ in (3) and $\hat{\theta}_B$ in (4) under covariate-adaptive randomization, including minimization. The asymptotic distribution of $\hat{\theta}$ in (2) is a special case of the result for $\hat{\theta}_A$ by setting $X = 0$.

THEOREM 1. *Assume that Conditions 1–3 hold, and that* $\mathrm{var}(X \mid Z = z)$ *is positive definite for any* $z \in \mathcal{Z}$. *As* $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_A - \theta) \xrightarrow{d} N(0, \Sigma_A + \Sigma_V),$$
$$\sqrt{n}(\hat{\theta}_B - \theta) \xrightarrow{d} N(0, \Sigma_B + \Sigma_V),$$

where

$$\begin{aligned} \Sigma_A &= \mathrm{diag}\langle \pi_t^{-1} E[\mathrm{var}\{Y^{(t)} - X^{\mathrm{T}}\beta_t(Z) \mid Z\}]\rangle + \pi_1^{-1} E[\mathrm{var}\{Y^{(1)} - X^{\mathrm{T}}\beta_1(Z) \mid Z\}]\iota_{k-1}\iota_{k-1}^{\mathrm{T}} \\ &\quad + E\{B(Z)^{\mathrm{T}}\mathrm{var}(X \mid Z)B(Z)\}, \\ \Sigma_B &= \mathrm{diag}\langle \pi_t^{-1} E[\mathrm{var}\{Y^{(t)} - X^{\mathrm{T}}\beta(Z) \mid Z\}]\rangle + \pi_1^{-1} E[\mathrm{var}\{Y^{(1)} - X^{\mathrm{T}}\beta(Z) \mid Z\}]\iota_{k-1}\iota_{k-1}^{\mathrm{T}}, \\ \Sigma_V &= \textit{covariance matrix of the vector } \{E(Y^{(2)} - Y^{(1)} \mid Z), \ldots, E(Y^{(k)} - Y^{(1)} \mid Z)\}^{\mathrm{T}}, \end{aligned}$$

$\beta_t(z) = \{\mathrm{var}(X \mid Z = z)\}^{-1}\mathrm{cov}(X, Y^{(t)} \mid Z = z)$ *for* $t = 1, \ldots, k$, $\beta(z) = \sum_{t=1}^{k} \pi_t \beta_t(z)$, $\mathrm{diag}\langle d_t \rangle$ *denotes the* $(k-1) \times (k-1)$ *diagonal matrix with diagonal elements* $d_2, \ldots, d_k$, ι_{k-1} *is the* $(k-1)$*-dimensional column vector of ones, and* $B(Z) = \{\beta_2(Z) - \beta_1(Z), \ldots, \beta_k(Z) - \beta_1(Z)\}$.

Theorem 1 is model free and is applicable to any covariate-adaptive randomization method satisfying Conditions 2 and 3, most noticeably the minimization method, for which very little is known about its theoretical property, as the minimization method is neither type 1 nor type 2 as described in § 2. This provides a solid foundation for valid and model-free inference after minimization.

The asymptotic result in Theorem 1 is invariant with respect to randomization methods, i.e., Σ_A, Σ_B and Σ_V do not depend on the randomization scheme. In other words, each estimator of θ in (2)–(4) has the same asymptotic distribution and efficiency regardless of which randomization scheme is used for treatment assignments, including simple randomization. This is intrinsically different from many existing results that are dependent on randomization methods (Shao & Yu, 2013; Ma et al., 2015; Bugni et al., 2018). The only result invariant with respect to randomization methods that can be found in the literature is Bugni et al. (2019, Theorem 3.1) for $\hat{\theta}$ in (2), although Bugni et al. (2019) do not explicitly state this invariance property.

Due to the use of covariate-adaptive randomization, the sample mean $\bar{Y}_t$ is not an average of independent random variables and, thus, the asymptotic distributions of estimators in (2)–(4) cannot be obtained by directly applying the central limit theorem for the sum of independent random variables. We overcome this difficulty by decomposing the tth component of $\hat{\theta} - \theta$ as the sum of the following two uncorrelated terms:

$$U_t = \sum_{z \in \mathcal{Z}} \frac{n(z)}{n} \Big[\{\bar{Y}_t(z) - \bar{Y}_1(z)\} - \{E(Y^{(t)} \mid Z = z) - E(Y^{(1)} \mid Z = z)\} \Big],$$

$$V_t = \sum_{z \in \mathcal{Z}} \frac{n(z)}{n} \{E(Y^{(t)} \mid Z = z) - E(Y^{(1)} \mid Z = z)\} - \theta_t.$$

Conditioned on $(I_1, \ldots, I_n, Z_1, \ldots, Z_n)$, U_t is an average of independent terms, so its limiting distribution can be derived by applying the central limit theorem, which consequently provides the unconditional asymptotic distribution of U_t. For V_t, the only random part is $n(z)$ whose limiting distribution can be easily derived. For $\hat{\theta}_A$ or $\hat{\theta}_B$, a similar decomposition can be obtained with the same V_t and a different U_t incorporating the covariate adjustment term. Details can be found in the Supplementary Material.

This decomposition is not only the key to establishing the asymptotic result, but also identifies two sources of variation. The variation of potential responses after stratifying by Z and adjusting for X is represented by Σ_A. The variation from treatment effect heterogeneity is measured by Σ_V. We allow arbitrary treatment effect heterogeneity, i.e., different subgroups according to levels of Z may benefit differently from the treatment. If there is no treatment effect heterogeneity, then $\Sigma_V = 0$.

In applications, it is often of interest to make inference on the average treatment effect

$$E(Y^{(t)} - Y^{(s)}) = \theta_t - \theta_s \tag{5}$$

between two fixed arms t and s. The estimators of $\theta_t - \theta_s$ under the three methods in (2)–(4) are $\hat{\theta}_t - \hat{\theta}_s$, $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$. Their asymptotic distributions and the asymptotic relative efficiencies among them are summarized in the following theorem.

THEOREM 2. *(i) Under the assumptions in Theorem 1,*

$$\sqrt{n}\{\hat{\theta}_t - \hat{\theta}_s - (\theta_t - \theta_s)\} \xrightarrow{d} N(0, \sigma^2_{t,s,U} + \sigma^2_{t,s,V}),$$
$$\sqrt{n}\{\hat{\theta}_{t,A} - \hat{\theta}_{s,A} - (\theta_t - \theta_s)\} \xrightarrow{d} N(0, \sigma^2_{t,s,A} + \sigma^2_{t,s,V}),$$
$$\sqrt{n}\{\hat{\theta}_{t,B} - \hat{\theta}_{s,B} - (\theta_t - \theta_s)\} \xrightarrow{d} N(0, \sigma^2_{t,s,B} + \sigma^2_{t,s,V}),$$

where

$$\begin{aligned}
\sigma^2_{t,s,U} &= \pi_t^{-1} E\{\mathrm{var}(Y^{(t)} \mid Z)\} + \pi_s^{-1} E\{\mathrm{var}(Y^{(s)} \mid Z)\},\\
\sigma^2_{t,s,A} &= \pi_t^{-1} E[\mathrm{var}\{Y^{(t)} - X^{\mathrm{T}}\beta_t(Z) \mid Z\}] + \pi_s^{-1} E[\mathrm{var}\{Y^{(s)} - X^{\mathrm{T}}\beta_s(Z) \mid Z\}]\\
&\quad + E[\{\beta_t(Z) - \beta_s(Z)\}^{\mathrm{T}} \mathrm{var}(X \mid Z)\{\beta_t(Z) - \beta_s(Z)\}],\\
\sigma^2_{t,s,B} &= \pi_t^{-1} E[\mathrm{var}\{Y^{(t)} - X^{\mathrm{T}}\beta(Z) \mid Z\}] + \pi_s^{-1} E[\mathrm{var}\{Y^{(s)} - X^{\mathrm{T}}\beta(Z) \mid Z\}],\\
\sigma^2_{t,s,V} &= \mathrm{var}\{E(Y^{(t)} - Y^{(s)} \mid Z)\}.
\end{aligned}$$

(ii) The difference between the asymptotic variances of $\hat{\theta}_t - \hat{\theta}_s$ and $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ is

$$\begin{aligned}
\sigma^2_{t,s,U} - \sigma^2_{t,s,A} &= \{\pi_t\pi_s(\pi_t + \pi_s)\}^{-1} E\left[\{\pi_s\beta_t(Z) + \pi_t\beta_s(Z)\}^{\mathrm{T}} \mathrm{var}(X \mid Z)\{\pi_s\beta_t(Z) + \pi_t\beta_s(Z)\}\right]\\
&\quad + \{(\pi_t + \pi_s)^{-1} - 1\} E\left[\{\beta_t(Z) - \beta_s(Z)\}^{\mathrm{T}} \mathrm{var}(X \mid Z)\{\beta_t(Z) - \beta_s(Z)\}\right]\\
&\geqslant 0,
\end{aligned}$$

where the equality holds if and only if, for every $z \in \mathcal{Z}$,

$$\pi_s\beta_t(z) + \pi_t\beta_s(z) = 0 \quad \textit{and} \quad \{\beta_t(z) - \beta_s(z)\}(1 - \pi_t - \pi_s) = 0. \tag{6}$$

(iii) The difference between the asymptotic variances of $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ and $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ is

$$\begin{aligned}
\sigma^2_{t,s,B} - \sigma^2_{t,s,A} &= \pi_t^{-1} E\left[\{\beta_t(Z) - \beta(Z)\}^{\mathrm{T}} \mathrm{var}(X \mid Z)\{\beta_t(Z) - \beta(Z)\}\right]\\
&\quad + \pi_s^{-1} E\left[\{\beta_s(Z) - \beta(Z)\}^{\mathrm{T}} \mathrm{var}(X \mid Z)\{\beta_s(Z) - \beta(Z)\}\right]\\
&\quad - E\left[\{\beta_t(Z) - \beta_s(Z)\}^{\mathrm{T}} \mathrm{var}(X \mid Z)\{\beta_t(Z) - \beta_s(Z)\}\right]\\
&\geqslant 0,
\end{aligned}$$

where the equality holds if and only if, for every $z \in \mathcal{Z}$,

$$\beta(z) = (\pi_s + \pi_t)^{-1}\{\pi_s\beta_t(z) + \pi_t\beta_s(z)\} \quad \textit{and} \quad \{\beta_t(z) - \beta_s(z)\}(1 - \pi_t - \pi_s) = 0. \tag{7}$$

Theorem 2 indicates that $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ is always asymptotically more efficient than $\hat{\theta}_t - \hat{\theta}_s$ unless (6) holds, in which case the two estimators have the same asymptotic efficiency. This theoretically corroborates the perception that covariate adjustment with a full set of treatment–covariate interactions cannot hurt efficiency. When there are more than two treatments, $1 - \pi_t - \pi_s > 0$ and, consequently, (6) holds only when $\beta_t(z) = \beta_s(z) = 0$ for every z, i.e., X is uncorrelated with the potential responses $Y^{(t)}$ and $Y^{(s)}$ after conditioning on Z so that adjusting for X is unnecessary. When there are only two treatments, (6) also holds if $\pi_t = \pi_s = 1/2$ and $\beta_t(z) = -\beta_s(z)$ for every z. An example is given in § 4.

Note that $\hat{\beta}(z)$ used in $\hat{\theta}_B$ ignores the fact that $\mathrm{cov}(X, Y^{(t)} \mid Z = z)$ may depend on treatment t. That is why $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ is asymptotically not as efficient as $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ in general, and $\sigma^2_{t,s,B} = \sigma^2_{t,s,A}$ when these covariances are the same for every t and every z, i.e., $\beta_1(z) = \cdots = \beta_k(z)$. An exceptional case is that $\sigma^2_{t,s,A} = \sigma^2_{t,s,B}$ when there are only two treatments and $\pi_t = \pi_s = 1/2$. In fact, $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ may be asymptotically less efficient than $\hat{\theta}_t - \hat{\theta}_s$, i.e., covariate adjustment with only the main effects may hurt efficiency, a perspective in Freedman (2008) and Lin (2013). For example, there are scenarios in which (6) holds, but (7) does not. Simulation examples are given in § 4, where comparisons of $\hat{\theta}_t - \hat{\theta}_s$, $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ are made.

To make model-free inference on θ defined in (1), we only need to apply Theorem 1 and construct consistent estimators of limiting variances. We focus on inference for $\theta_t - \theta_s$ defined in (5) with fixed t and s; other parameters of θ can be similarly treated. Let $S^2_t(z)$ be the sample variance of the Y_i in the group of patients under treatment t with $Z_i = z$, $S^2_{t,A}(z)$ be $S^2_t(z)$ with Y_i replaced by $Y_i - X_i^{\mathrm{T}}\hat{\beta}_t(z)$, $S^2_{t,B}(z)$ be $S^2_t(z)$ with Y_i replaced by $Y_i - X_i^{\mathrm{T}}\hat{\beta}(z)$, and $\hat{\Sigma}(z)$ be the sample covariance matrix of the X_i within $Z_i = z$. It is shown in the Supplementary Material that, under Conditions 1–3, the following estimators are consistent for $\sigma^2_{t,s,U}$, $\sigma^2_{t,s,V}$, $\sigma^2_{t,s,A}$ and $\sigma^2_{t,s,B}$, respectively:

$$\hat{\sigma}^2_{t,s,U} = \sum_{z\in\mathcal{Z}} \frac{n(z)}{n}\left\{\frac{S^2_t(z)}{\pi_t} + \frac{S^2_s(z)}{\pi_s}\right\},$$

$$\hat{\sigma}^2_{t,s,A} = \sum_{z\in\mathcal{Z}} \frac{n(z)}{n}\left[\frac{S^2_{t,A}(z)}{\pi_t} + \frac{S^2_{s,A}(z)}{\pi_s} + \{\hat{\beta}_t(z) - \hat{\beta}_s(z)\}^{\mathrm{T}}\hat{\Sigma}(z)\{\hat{\beta}_t(z) - \hat{\beta}_s(z)\}\right],$$

$$\hat{\sigma}^2_{t,s,B} = \sum_{z\in\mathcal{Z}} \frac{n(z)}{n}\left\{\frac{S^2_{t,B}(z)}{\pi_t} + \frac{S^2_{s,B}(z)}{\pi_s}\right\},$$

$$\hat{\sigma}^2_{t,s,V} = \sum_{z\in\mathcal{Z}} \frac{n(z)}{n}\left\{\bar{Y}_t(z) - \bar{Y}_s(z)\right\}^2 - (\hat{\theta}_t - \hat{\theta}_s)^2,$$

regardless of which type of covariate-adaptive randomization method is used. Note that $\hat{\sigma}^2_{t,s,U}$ for $\hat{\theta}_t - \hat{\theta}_s$ is different from the estimator obtained from Bugni et al. (2019, (36)).

4. Simulation results

There have been many publications on empirical studies under covariate-adaptive randomization in the last four decades. Some recent results are presented in Senn et al. (2010), Kahan & Morris (2012) and Xu et al. (2016).

To evaluate and compare our proposed estimators $\hat{\theta}_t - \hat{\theta}_s$, $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ in terms of estimation bias and standard deviation, and to examine variance estimators and the related asymptotic confidence intervals based on Theorem 1, we present some simulation results in this section. We consider two covariates, i.e., $W = (X_1, X_2)$, where X_1 is binary with $\mathrm{pr}(X_1 = 1) = 1/2$ and, conditioned on X_1, $X_2 \sim N(X_1 - 0.5,\ 1)$. For the potential responses, we consider two treatments in cases I–III and three treatments in case IV:

Case I: $Y^{(1)} \mid W \sim N(4X_1 + 2X_2,\ 1)$, $Y^{(2)} \mid W \sim N(\varphi + 4X_1 + 2X_2,\ 1)$.
Case II: $Y^{(1)} \mid W \sim N(4X_1 - 2X_2,\ 1)$, $Y^{(2)} \mid W \sim N(\varphi + 4X_1 + 2X_2,\ 1)$.

Table 1. *Smallest expected number of patients among all stratum–treatment combinations*

			Allocation		
n	Z	No. levels	1 : 1	1 : 2	1 : 2 : 2
100	X_1	2	25.0	16.7	10.0
	X_1, d_2	4	7.7	5.1	3.1
	X_1, d_4	8	2.4	1.6	1.0
500	X_1	2	125.0	83.3	50.0
	X_1, d_2	4	38.6	25.7	15.4
	X_1, d_4	8	12.1	8.1	4.8

Case III: $Y^{(1)} \mid W \sim N(0.25 + 3X_1 + 0.2X_2^2, X_1 + 0.5)$, $Y^{(2)} \mid W \sim N(\varphi + 4X_1 + 2X_2, 1)$.
Case IV: $Y^{(3)} \mid W \sim N(\psi + 1 + 2X_1 - X_2, 1)$, and $Y^{(1)}$ and $Y^{(2)}$ are the same as case III.

We use $\varphi = \psi = 1$ in the simulation, which does not affect the relative performance of estimators and coverage probability of related confidence intervals.

Case I has homogeneous treatment effects; case II has treatment effect heterogeneity since the effects of X_2 on $Y^{(1)}$ and $Y^{(2)}$ have different signs; case III has the most severe treatment effect heterogeneity as $Y^{(1)} \mid W$ and $Y^{(2)} \mid W$ have very different distributions; case IV considers multiple treatments.

We consider three different Z in covariate-adaptive randomization. The first one is $Z = X_1$ with two levels, in which case the function of W not used in randomization, but still related to the potential responses is $h(W) = X_2$. The second Z is $Z = (X_1, d_2)$ with four levels, where d_2 is the discretized X_2 with two categories, $(-\infty, 0)$ and $[0, \infty)$, and $h(W)$ is the continuous value of X_2 in $(-\infty, 0)$ or $(0, \infty)$. The third Z is $Z = (X_1, d_4)$ with eight levels, where d_4 is the discretized X_2 with four categories, $(-\infty, -0.8)$, $[-0.8, 0)$, $[0, 0.8)$ and $[0.8, \infty)$, and $h(W)$ is the continuous value of X_2 in $(-\infty, -0.8)$, $(-0.8, 0)$, $(0, 0.8)$ or $(0.8, \infty)$. In all cases, $X = X_2$ is equivalent to $h(W)$ and is used in covariate adjustment.

For the randomization method, we consider minimization with treatment allocation 1 : 1 or 1 : 2 for cases I–III, and 1 : 2 : 2 for case IV. Simulation results for two other randomization methods, the stratified permuted block randomization and the stratified urn design, can be found in the Supplementary Material.

We consider the total sample size $n = 100$ or 500. For these sample sizes, the smallest possible expected numbers of patients within a stratum and treatment according to number of Z levels are given in Table 1. It can be seen that when $n = 100$ and $Z = (X_1, d_4)$ has eight levels, with nonnegligible probability, the number of patients in some stratum–treatment combination is fewer than two and thus calculation of estimators in (2)–(4), and their variance estimators is not possible. Therefore, for cases I–III, we omit the scenario with $n = 100$ and $Z = (X_1, d_4)$. For case IV, we focus on $n = 500$ and $Z = (X_1, d_2)$.

Tables 2 and 3 report the bias, standard deviation, average estimated standard deviation and coverage probability of asymptotic 95% confidence intervals, estimate ± 1.96 SE, of $\hat{\theta}_t - \hat{\theta}_s$, $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ for cases I–IV. Every scenario is evaluated with 2000 simulation runs.

From Tables 2 and 3 we see that all the estimators have negligible biases that are smaller than 1% in most cases. The variance estimators or average estimated standard deviations are very accurate, so that the coverage probabilities of confidence intervals are adequate. Even when the smallest expected number of patients is as small as 5.1 or 7.7 in the case of $n = 100$ and a Z with four levels, $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ perform well.

Table 2. *Bias, standard deviation, average estimated standard deviation and coverage probability of 95% asymptotic confidence intervals under minimization for cases I–III*

				Treatment allocation 1 : 1				Treatment allocation 1 : 2			
n	Case	Z	Estimator	Bias	SD	SE	CP	Bias	SD	SE	CP
500	I	X_1	$\hat{\theta}$	−0.0038	0.1980	0.1999	0.9590	0.0070	0.2159	0.2124	0.9465
			$\hat{\theta}_B$	−0.0016	0.0909	0.0893	0.9445	0.0016	0.0954	0.0949	0.9510
			$\hat{\theta}_A$	−0.0016	0.0908	0.0893	0.9445	0.0017	0.0954	0.0948	0.9490
		X_1, d_2	$\hat{\theta}$	−0.0029	0.1492	0.1466	0.9450	−0.0013	0.1537	0.1553	0.9560
			$\hat{\theta}_B$	−0.0017	0.0900	0.0893	0.9455	−0.0011	0.0962	0.0947	0.9440
			$\hat{\theta}_A$	−0.0018	0.0901	0.0894	0.9455	−0.0009	0.0965	0.0945	0.9435
		X_1, d_4	$\hat{\theta}$	0.0005	0.1143	0.1150	0.9560	0.0003	0.1237	0.1219	0.9485
			$\hat{\theta}_B$	0.0010	0.0903	0.0893	0.9505	0.0013	0.0967	0.0946	0.9425
			$\hat{\theta}_A$	0.0007	0.0908	0.0893	0.9500	0.0010	0.0990	0.0944	0.9415
	II	X_1	$\hat{\theta}$	0.0085	0.2212	0.2191	0.9480	0.0067	0.2320	0.2303	0.9505
			$\hat{\theta}_B$	0.0086	0.2222	0.2185	0.9495	0.0063	0.2563	0.2541	0.9430
			$\hat{\theta}_A$	0.0086	0.2214	0.2191	0.9500	0.0078	0.2255	0.2212	0.9470
		X_1, d_2	$\hat{\theta}$	0.0084	0.2201	0.2191	0.9480	0.0076	0.2284	0.2251	0.9465
			$\hat{\theta}_B$	0.0076	0.2214	0.2178	0.9440	0.0078	0.2407	0.2344	0.9350
			$\hat{\theta}_A$	0.0085	0.2204	0.2190	0.9475	0.0077	0.2242	0.2212	0.9450
		X_1, d_4	$\hat{\theta}$	0.0057	0.2222	0.2192	0.9440	0.0104	0.2256	0.2230	0.9405
			$\hat{\theta}_B$	0.0061	0.2233	0.2177	0.9425	0.0108	0.2289	0.2254	0.9420
			$\hat{\theta}_A$	0.0057	0.2221	0.2190	0.9430	0.0101	0.2252	0.2211	0.9420
	III	X_1	$\hat{\theta}$	0.0003	0.1716	0.1731	0.9475	0.0072	0.1691	0.1667	0.9425
			$\hat{\theta}_B$	0.0029	0.1495	0.1477	0.9500	0.0048	0.1675	0.1656	0.9475
			$\hat{\theta}_A$	0.0031	0.1496	0.1479	0.9480	0.0081	0.1546	0.1533	0.9465
		X_1, d_2	$\hat{\theta}$	0.0016	0.1580	0.1593	0.9490	0.0032	0.1603	0.1595	0.9465
			$\hat{\theta}_B$	0.0050	0.1468	0.1470	0.9460	0.0061	0.1621	0.1576	0.9420
			$\hat{\theta}_A$	0.0032	0.1464	0.1474	0.9480	0.0052	0.1559	0.1523	0.9440
		X_1, d_4	$\hat{\theta}$	0.0042	0.1528	0.1525	0.9465	0.0049	0.1601	0.1557	0.9405
			$\hat{\theta}_B$	0.0080	0.1495	0.1468	0.9425	0.0094	0.1577	0.1539	0.9395
			$\hat{\theta}_A$	0.0048	0.1493	0.1474	0.9450	0.0069	0.1573	0.1521	0.9410
100	I	X_1	$\hat{\theta}$	−0.0001	0.4504	0.4487	0.9500	0.0064	0.4696	0.4743	0.9500
			$\hat{\theta}_B$	0.0019	0.2041	0.1983	0.9370	−0.0003	0.2175	0.2098	0.9365
			$\hat{\theta}_A$	0.0017	0.2044	0.2260	0.9645	0.0015	0.2242	0.2432	0.9615
		X_1, D_2	$\hat{\theta}$	−0.0019	0.3214	0.3308	0.9520	−0.0059	0.3481	0.3489	0.9500
			$\hat{\theta}_B$	0.0005	0.2041	0.1983	0.9440	−0.0046	0.2184	0.2094	0.9375
			$\hat{\theta}_A$	−0.0003	0.2046	0.2202	0.9625	−0.0029	0.2269	0.2372	0.9575
	II	X_1	$\hat{\theta}$	−0.0074	0.4863	0.4903	0.9525	−0.0106	0.5114	0.5151	0.9455
			$\hat{\theta}_B$	−0.0051	0.4930	0.4831	0.9420	−0.0152	0.5695	0.5585	0.9390
			$\hat{\theta}_A$	−0.0048	0.4925	0.4891	0.9455	−0.0117	0.5017	0.4967	0.9455
		X_1, D_2	$\hat{\theta}$	−0.0125	0.4799	0.4892	0.9525	−0.0046	0.4856	0.5014	0.9520
			$\hat{\theta}_B$	−0.0114	0.4928	0.4759	0.9345	−0.0051	0.5239	0.5099	0.9350
			$\hat{\theta}_A$	−0.0106	0.4862	0.4754	0.9375	−0.0066	0.4895	0.4807	0.9380
	III	X_1	$\hat{\theta}$	−0.0048	0.3925	0.3871	0.9455	−0.0025	0.3715	0.3719	0.9500
			$\hat{\theta}_B$	0.0048	0.3350	0.3270	0.9410	−0.0005	0.3733	0.3639	0.9405
			$\hat{\theta}_A$	0.0047	0.3350	0.3331	0.9440	0.0089	0.3459	0.3411	0.9445
		X_1, D_2	$\hat{\theta}$	−0.0109	0.3495	0.3560	0.9535	−0.0055	0.3459	0.3549	0.9525
			$\hat{\theta}_B$	0.0055	0.3341	0.3224	0.9405	0.0089	0.3490	0.3432	0.9355
			$\hat{\theta}_A$	0.0039	0.3314	0.3259	0.9480	0.0191	0.3317	0.3334	0.9410

Since $k = 2$, $\hat{\theta} = \hat{\theta}_2 - \hat{\theta}_1$, $\hat{\theta}_A = \hat{\theta}_{2,A} - \hat{\theta}_{1,A}$ and $\hat{\theta}_B = \hat{\theta}_{2,B} - \hat{\theta}_{1,B}$. SD, standard deviation; SE, average estimated standard deviation; CP, coverage probability.

Table 3. *Bias, standard deviation, average estimated standard deviation and coverage probability of 95% asymptotic confidence interval under minimization for case IV with* $n = 500$

t	s	$\theta_t - \theta_s$	Estimator	Bias	SD	SE	CP
2	1	1	$\hat{\theta}_t - \hat{\theta}_s$	−0.0007	0.1907	0.1901	0.9515
			$\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$	0.0040	0.1840	0.1821	0.9470
			$\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$	0.0058	0.1777	0.1726	0.9375
3	1	1	$\hat{\theta}_t - \hat{\theta}_s$	−0.0004	0.1541	0.1546	0.9445
			$\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$	0.0037	0.1616	0.1615	0.9445
			$\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$	0.0052	0.1479	0.1460	0.9395
3	2	0	$\hat{\theta}_t - \hat{\theta}_s$	0.0004	0.2094	0.2082	0.9505
			$\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$	−0.0002	0.2077	0.2048	0.9445
			$\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$	−0.0006	0.2019	0.2007	0.9495

With homogeneous treatment effects in case I, a more informative Z leads to a more efficient $\hat{\theta}_t - \hat{\theta}_s$. However, the same phenomenon may not exist when treatment effect heterogeneity exists, though a more informative Z does not lead to a less efficient $\hat{\theta}_t - \hat{\theta}_s$.

Adjusting for covariates, i.e., using $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ or $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$, may lead to substantial improvements over $\hat{\theta}_t - \hat{\theta}_s$ in terms of standard deviation, which again agrees with our theory. The improvement is larger when a less informative Z is utilized in randomization, such as $Z = X_1$. Another interesting observation is that a different Z used in randomization does not affect the standard deviation of $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ or $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ very much.

The comparison of $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ is also consistent with our theory in § 3. Under $1:1$ treatment allocation or homogeneous treatment effects, $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ is as good as $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$. When the treatment allocation is $1:2$ and treatment effect heterogeneity exists, $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ is not as good as $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and could be even worse than $\hat{\theta}_t - \hat{\theta}_s$. The same is observed when the treatment allocation is $1:2:2$.

In case II with $1:1$ treatment allocation, $\mathrm{cov}(X, Y^{(2)} \mid Z = z) = -\mathrm{cov}(X, Y^{(1)} \mid Z = z)$, i.e., (6) holds and, thus, $\hat{\theta}_t - \hat{\theta}_s$ and $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ have very similar standard deviations, as predicted by Theorem 2. In this particular case, $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ is also as good as $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$.

5. REAL-DATA EXAMPLE

For illustration, we apply our methods to a real-data example from Chong et al. (2016), whose goal is to evaluate the contribution of low dietary iron intake to human capital attainment by measuring the causal effect of reducing adolescent anemia on school attainment. The dataset is publicly available at `https://www.openicpsr.org/openicpsr/project/113624/version/V1/view`. In brief, Chong et al. (2016) conducted an experiment on students aged from 11 to 19 in rural Peru, where iron deficiency is high, to study whether the following three promotional videos, considered as three treatments, can encourage students to increase their iron intake and hence improve their school performance. The first video shows a popular soccer player encouraging iron supplements to maximize energy; the second video shows a physician encouraging iron supplements for overall health; and the third placebo video shows a dentist encouraging oral hygiene without mentioning iron at all. A stratified permuted block randomization design was applied to assign 219 students to the three treatments with allocation $1:1:1$, using the student's school grade as covariate Z with $\mathcal{Z} = \{1, 2, 3, 4, 5\}$. Four students were excluded from the analysis for various reasons (Chong et al., 2016). The number of students in each stratum by

Table 4. *Number of students in stratum by treatment combination in real-data example*

	Soccer	Physician	Placebo
$z = 1$	16	17	15
$z = 2$	19	20	19
$z = 3$	15	15	16
$z = 4$	10	11	12
$z = 5$	10	10	10

Table 5. *Estimate, SE, and p-value for the real-data example*

	Soccer versus placebo			Physician versus placebo		
	Estimate	SE	p-value	Estimate	SE	p-value
$\hat{\theta}_t$	−0.051	0.205	0.803	0.409	0.207	0.048
$\hat{\theta}_{t,B}$	−0.089	0.203	0.661	0.444	0.202	0.028
$\hat{\theta}_{t,A}$	−0.045	0.198	0.821	0.484	0.197	0.014

Treatment 1 is placebo.

treatment is given in Table 4, which shows that no stratum by treatment has too few units to apply our methods.

As an example, we follow Bugni et al. (2019) and consider the outcome of academic achievement, which is a standardized average of a student's academic grades from a given semester in subjects of math, foreign language, social sciences, science and communications. Estimates $\hat{\theta}_t - \hat{\theta}_s$, $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ and their average estimated standard deviations are reported in Table 5 for the average treatment effect between the soccer player and placebo videos, or physician and placebo videos, together with the p-values associated with two-sided tests of no treatment effect. The estimates from $\hat{\theta}_t - \hat{\theta}_s$ are the same as those in Bugni et al. (2019). The covariate X used in $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ and $\hat{\theta}_{t,B} - \hat{\theta}_{s,B}$ is the baseline anemia status thought to have an interactive effect with treatment on the outcome, as mentioned in Chong et al. (2016). It can be seen that the average estimated standard deviation of $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$ is the smallest and, in terms of p-values, the effect between physician and placebo videos is only marginally significant when $\hat{\theta}_t - \hat{\theta}_s$ is used, but very significant based on our proposed $\hat{\theta}_{t,A} - \hat{\theta}_{s,A}$.

6. Concluding remarks

To improve asymptotic efficiency, we recommend $\hat{\theta}_A$ in (3) since it is asymptotically better than $\hat{\theta}$ in (2) or $\hat{\theta}_B$ in (4). In the special case of two treatment arms with equal allocation, we recommend $\hat{\theta}_B$, since it is asymptotically equivalent to $\hat{\theta}_A$ and has better empirical performance.

As a full stratification according to Z is required, one limitation of the estimators in (2)–(4) is that all strata need to have large enough sizes. In view of the empirical results in § 4 and § 5 and in Ye & Shao (2019), we recommend our procedures when there are at least 10 units in every stratum-treatment combination. Our future research is to study how to combine strata with too small sizes. Both covariate-adaptive randomization in treatment assignment and adjustment for covariates in estimation can gain efficiency, and covariate-adaptive randomization has the practically important advantage of balancing assignments across prognostic factors. Thus, another future research area is to study how to choose Z and X from the entire W.

ACKNOWLEDGEMENT

We thank all reviewers for useful comments and suggestions. Our research was supported by the National Natural Science Foundation of China and the U.S. National Science Foundation. Shao is also affiliated with the University of Wisconsin.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online contains all technical proofs and further simulation results. `R` code for the methods proposed in this paper can be found in the `R` package `RobinCar` at `https://github.com/tye27/RobinCar`.

REFERENCES

ATKINSON, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* **69**, 61–7.

ATKINSON, A. C. (1999). Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statist. Med.* **18**, 1741–52.

ATKINSON, A. C. (2002). The comparison of designs for sequential clinical trials with covariate information. *J. R. Statist. Soc.* A **165**, 349–73.

BALDI ANTOGNINI, A. & ZAGORAIOU, M. (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika* **98**, 519–35.

BALDI ANTOGNINI, A. & ZAGORAIOU, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli* **21**, 881–908.

BUGNI, F. A., CANAY, I. A. & SHAIKH, A. M. (2018). Inference under covariate-adaptive randomization. *J. Am. Statist. Assoc.* **113**, 1784–96.

BUGNI, F. A., CANAY, I. A. & SHAIKH, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quant. Econom.* **10**, 1747–85.

CASSEL, C. M., SÄRNDAL, C. E. & WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–20.

CHONG, A., COHEN, I., FIELD, E., NAKASONE, E. & TORERO, M. (2016). Iron deficiency and schooling attainment in Peru. *Am. Econom. J.: Appl. Econom.* **8**, 222–55.

CIOLINO, J. D., PALAC, H. L., YANG, A., VACA, M. & BELLI, H. M. (2019). Ideal vs. real: A systematic review on handling covariates in randomized controlled trials. *BMC Med. Res. Methodol.* **19**, 136.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403–17.

EUROPEAN MEDICINES AGENCY (2015). *Guideline on Adjustment for Baseline Covariates in Clinical Trials*. EMA/CHMP/295050/2013.

FREEDMAN, D. A. (2008). On regression adjustments in experiments with several treatments. *Ann. Appl. Statist.* **2**, 176–96.

HAN, B., ENAS, N. H. & MCENTEGART, D. (2009). Randomization by minimization for unbalanced treatment allocation. *Statist. Med.* **28**, 3329–46.

HU, Y. & HU, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Ann. Statist.* **40**, 1794–815.

KAHAN, B. C. & MORRIS, T. P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. *Statist. Med.* **31**, 328–40.

KAISER, L. D. (2012). Dynamic randomization and a randomization model for clinical trials data. *Statist. Med.* **31**, 3858–73.

KUZNETSOVA, O. M. & JOHNSON, V. P. (2017). Approaches to expanding the two-arm biased coin randomization to unequal allocation while preserving the unconditional allocation ratio. *Statist. Med.* **36**, 2483–98.

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Statist.* **7**, 295–318.

MA, W., HU, F. & ZHANG, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *J. Am. Statist. Assoc.* **110**, 669–80.

POCOCK, S. J. & SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31**, 103–15.

ROSENBERGER, W. F. & SVERDLOV, O. (2008). Handling covariates in the design of clinical trials. *Statist. Sci.* **23**, 404–19.

SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. (2003). *Model Assisted Survey Sampling*. New York: Springer.

Schulz, K. F. & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: Chance, not choice. *Lancet* **359**, 515–19.
Senn, S., Anisimov, V. V. & Fedorov, V. V. (2010). Comparisons of minimization and atkinson's algorithm. *Statist. Med.* **29**, 721–30.
Shao, J. & Wang, S. (2014). Efficiency of model-assisted regression estimators in sample surveys. *Statist. Sinica* **24**, 395–414.
Shao, J. & Yu, X. (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics* **69**, 960–9.
Shao, J., Yu, X. & Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika* **97**, 347–60.
Simon, R. & Simon, N. R. (2011). Using randomization tests to preserve type I error with response-adaptive and covariate-adaptive randomization. *Statist. Prob. Lett.* **81**, 767–72.
Smith, R. L. (1984). Sequential treatment allocation using biased coin designs. *J. R. Statist. Soc.* B **46**, 519–43.
Ta, T., Shao, J., Li, Q. & Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statist. Sinica* **30**, 1135–54.
Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Therap.* **15**, 443–53.
Taves, D. R. (2010). The use of minimization in clinical trials. *Contemp. Clin. Trials* **31**, 180–4.
Wei, L.-J. (1977). A class of designs for sequential clinical trials. *J. Am. Statist. Assoc.* **72**, 382–6.
Wei, L.-J. (1978). The adaptive biased coin design for sequential experiments. *Ann. Statist.* **6**, 92–100.
Xu, Z., Proschan, M. & Lee, S. (2016). Validity and power considerations on hypothesis testing under minimization. *Statist. Med.* **35**, 2315–27.
Ye, T. (2018). Testing hypotheses under covariate-adaptive randomisation and additive models. *Statist. Theory Rel. Fields* **2**, 96–101.
Ye, T. & Shao, J. (2019). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *arXiv:*1811.07232v2.
Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *J. Clin. Epidemiol.* **27**, 365–75.
Zhao, W. & Ramakrishnan, V. (2016). Generalization of Wei's urn design to unequal allocations in sequential clinical trials. *Contemp. Clin. Trials Commun.* **2**, 75–9.

[*Received on* 26 *February* 2020. *Editorial decision on* 22 *February* 2021]