Testing Hypotheses of Covariate-Adaptive Randomized Clinical Trials

Wei MA, Feifang Hu, and Lixin Zhang

Covariate-adaptive designs are often implemented to balance important covariates in clinical trials. However, the theoretical properties of conventional testing hypotheses are usually unknown under covariate-adaptive randomized clinical trials. In the literature, most studies are based on simulations. In this article, we provide theoretical foundation of hypothesis testing under covariate-adaptive designs based on linear models. We derive the asymptotic distributions of the test statistics of testing both treatment effects and the significance of covariates under null and alternative hypotheses. Under a large class of covariate-adaptive designs, (i) the hypothesis testing to compare treatment effects is usually conservative in terms of small Type I error; (ii) the hypothesis testing to compare treatment effects is usually more powerful than complete randomization; and (iii) the hypothesis testing for significance of covariates is still valid. The class includes most of the covariate-adaptive designs in the literature; for example, Pocock and Simon’s marginal procedure, stratified permuted block design, etc. Numerical studies are also performed to assess their corresponding finite sample properties. Supplementary material for this article is available online.

KEY WORDS: Conservative tests; Linear models; Pocock and Simon’s marginal procedure; Power; Stratified permuted block design; Type I error

1. INTRODUCTION

In clinical trials, it is usually important to balance treatment arms with respect to key covariates. To do this, many covariate-adaptive designs have been proposed in the literature. One natural idea is doing stratification first and then employing separate randomization within each strata; for example, stratified permuted block design, etc. To deal with many covariates, Taves (1974) and Pocock and Simon (1975) introduced the minimization method, attempting to minimize the weighted sum of marginal differences between numbers of patients for all covariates. Recently Hu and Hu (2012) discussed some limitations of these classical designs and proposed a generalized family of covariate-adaptive designs, and obtained their theoretical properties. For more discussion of handling covariates in clinical trials, see McEntegart (2003), Zhang et al. (2007), Rosenberger and Sverdlov (2008), Hu, ZHU, and Hu (2014), and reference therein.

Both stratified permuted block design and Pocock and Simon’s marginal procedure are extensively implemented in clinical studies. Stratified permuted block design was used in a number of clinical trials, including Iacono et al. (2006) and Jakob et al. (2012). The usage of Pocock and Simon’s marginal method had also increased in last decades. According to Taves (2010), Pocock and Simon’s marginal procedure was implemented in over 400 clinical trials from 1989 to 2008. Some recent examples include Anderson et al. (2000), Gridelli et al. (2003), Krueger et al. (2007), Molander et al. (2007), Ohtori et al. (2012), and so on.

Even though many covariate-adaptive designs have been proposed and implemented in clinical trials, the discussion of statistical inference associated with those methods is limited. In practice, conventional tests are often employed without consideration of covariate-adaptive randomization scheme. It remains a concern if conventional tests are still valid under covariate-adaptive designs. It is now generally accepted that covariates used in trial design should also be incorporated in inference procedures. Forsythe (1987) suggested all covariates used in minimization method should be included into analysis to achieve a valid test through simulation studies. Shao, Yu, and Zhong (2010, p. 347) pointed out that “one way to obtain a valid test procedure is to use a correct model between outcomes and covariates, including those used in randomization.” However, in practice, it is often that not all covariate information used in randomization can be fully used in inference procedures. In a clinical trial described in Anderson et al. (2000), the Pocock and Simon’s marginal procedure was implemented to balance allocation over three covariates including clinical centers, performance status, and disease extent. A continuous primary endpoint was compared between two treatment groups using a two-sample t-test, without adjusting covariate effects at all. In practice, some randomization covariates are omitted in final analysis because: (i) it is difficult to incorporate some covariates in the analysis model, for example, investigation sites, etc.; and (ii) adjusting too many covariates usually means complicated modeling technique.

There have been doubts about the validity of statistical inference for covariate-adaptive designs, especially covariates that are fully or partially omitted in inference procedures. Birkett (1985) and Forsythe (1987) raised concerns about validity of unadjusted analysis under covariate-adaptive designs. They found that the two-sample t-test is conservative in terms of small Type I error if Taves’ minimization is used to allocate patients to
2. HYPOTHESIS TESTING UNDER COVARIATE-ADAPTIVE DESIGNS

In this section, we study hypothesis testing based on a linear model framework for covariate-adaptive designs. Suppose two treatments, 1 and 2, are studied under a covariate-adaptive randomized clinical trial, and \( \mu_1 \) and \( \mu_2 \) are parameters measuring the main effects of treatment 1 and 2, respectively. Let \( N \) be the total number of patients enrolled in the study. Let \( I_i \) be the assignment of \( i \)-th patient, that is, \( I_i = 1 \) for treatment 1 and \( I_i = 0 \) for treatment 2, \( i = 1, 2, \ldots, N \). Conditional on the treatment assignment \( I_i \), the following linear model is assumed for the response of the \( i \)-th patient \( Y_i \):

\[
Y_i = \mu_1 I_i + \mu_2(1 - I_i) + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \gamma_1 Z_{i,1} + \cdots + \gamma_q Z_{i,q} + \varepsilon_i,
\]

where

- the \( X_{i,k}'s \) and \( Z_{i,j}'s \) are discrete or continuous covariates which are independent and identically distributed as \( X_k \) and \( Z_j, k = 1, \ldots, p \) and \( j = 1, \ldots, q \);
- both the \( X_{i,k}'s \) and \( Z_{i,j}'s \) are used in the randomization procedure, but only the \( X_{i,k}'s \) are used in the final statistical inference, \( k = 1, \ldots, p \) and \( j = 1, \ldots, q \);
- all covariates are independent of each other, and \( EX_k = 0 \) and \( EZ_j = 0 \) for all \( k \) and \( j, k = 1, \ldots, p \) and \( j = 1, \ldots, q \); and
- the \( \varepsilon_i's \) are independent and identically distributed random errors with mean zero and variance \( \sigma^2 \) and independent of \( X_k \) and \( Z_j, k = 1, \ldots, p \) and \( j = 1, \ldots, q \).

Notice both \( X_{i,k} \) and \( Z_{i,j} \) are assumed to be scalars in model (1). If \( X_{i,k} \) (or \( Z_{i,j} \)) is a discrete covariate, \( X_{i,k} \) (or \( Z_{i,j} \)) is a scalar that can take several values corresponding to different categories. In practice, a vector is usually used to represent a discrete covariate with multiple categories. For example, a covariate having three categories can be coded as a two-dimensional vector with values of \((0,0), (1,0), \) and \((0,1)\). In this article, \( X_{i,k} \) (or \( Z_{i,j} \)) is assumed to be a scalar for simplicity, but all the results can be extended to the situation where discrete covariates with multiple categories are represented by vectors.

We let \( Y = (Y_1, Y_2, \ldots, Y_N)^T \), \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T \), \( \beta = (\mu_1, \mu_2, \beta_1, \ldots, \beta_p)^T \), and \( \gamma = (\gamma_1, \ldots, \gamma_q)^T \), and

\[
X = \begin{bmatrix} I_1 & 1 - I_1 & X_{1,1} & \cdots & X_{1,p} \\
I_2 & 1 - I_2 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\
I_N & 1 - I_N & X_{N,1} & \cdots & X_{N,p} \end{bmatrix}
\] and

\[
Z = \begin{bmatrix} Z_{1,1} & \cdots & Z_{1,q} \\
\vdots & \ddots & \vdots \\
Z_{N,1} & \cdots & Z_{N,q} \end{bmatrix}.
\]

Then model (1) can be written as

\[
Y = X\beta + Z\gamma + \varepsilon.
\]

The working model of inference would be,

\[
E[Y_{i}] = \mu_1 I_i + \mu_2(1 - I_i) + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p},
\]

or in the matrix form,

\[
E[Y] = X\hat{\beta}.
\]

The ordinary least squares method is used to obtain the estimate of \( \beta \), which has the explicit form

\[
\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + Z\gamma + \varepsilon).
\]
When model (2) is constructed to study data from a covariate-adaptive randomized clinical trial, the primary interest is usually to compare treatment effects between different groups. To compare treatment effects of $\mu_1$ and $\mu_2$, the following hypothesis testing is used

$$H_0 : \mu_1 - \mu_2 = 0 \text{ versus } H_A : \mu_1 - \mu_2 \neq 0. \quad (3)$$

The test statistic for (3) is

$$T = \frac{L \hat{\beta}}{(\hat{\alpha}^2 L (X^T X)^{-1} L^T)^{1/2}}, \quad (4)$$

where $L = (1, -1, 0, \ldots, 0)$ and $\hat{\alpha}^2 = (Y - X \hat{\beta})^T (Y - X \hat{\beta}) / (N - p - 2)$. If $|T| > Z_{1-\alpha/2}$, then we reject the null hypothesis; otherwise accept the null hypothesis.

A special case of testing (5) is evaluating the significance of a single covariate (biomarker). This is usually important in personalized medicine (Zhu, Hu, and Zhao 2013). Without loss of generality, we consider the hypothesis testing for $\beta_1$, the coefficient of $X_1$. To test the significance of $\beta_1$, the hypothesis would be

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0. \quad (7)$$

The test statistic for hypothesis testing (7) can be reduced

$$T_1 = \frac{\ell \hat{\beta}}{(\hat{\alpha}^2 \ell (X^T X)^{-1} \ell^T)^{1/2}}, \quad (8)$$

where $\ell = (0, 0, 1, 0, \ldots, 0)$. If $|T_1| > Z_{1-\alpha/2}$, then we reject the null hypothesis; otherwise accept the null hypothesis.

In clinical trials, covariate-adaptive designs are usually based on discrete covariates (Taves 2010). If a continuous covariate is to be used in randomization, a continuous-discrete conversion need be performed to breakdown continuous covariate into a discrete variable with several subcategories. Let $C = \{ j \mid Z_j \text{ is continuous, } j = 1, \ldots, q \}$ and $C^* = \{ k \mid X_k \text{ is continuous, } k = 1, \ldots, p \}$. If $k \in C^*$ or $j \in C$, the covariate-adaptive design is applied with respect to discrete variables, $d_k^*(X_k)$ or $d_j(Z_j)$, where $d_k^*$, $d_j$ are discrete functions. In this case, define $\delta_{i,k}^* = X_{i,k} - E(X_{i,k} | d_k^*(X_{i,k}))$ and $\delta_{i,j} = Z_{i,j} - E(Z_{i,j} | d_j(Z_{i,j}))$. Let

$$\tilde{X}_k = \begin{cases} X_k & \text{if } k \notin C^* \\ d_k^*(X_k) & \text{if } k \in C^* \end{cases}$$

and

$$\tilde{Z}_j = \begin{cases} Z_j & \text{if } j \notin C \\ d_j(Z_j) & \text{if } j \in C \end{cases}$$

where $\tilde{X}_k$ and $\tilde{Z}_j$ are ith observations of covariates $X_k$ and $Z_j$, $k = 1, \ldots, p$ and $j = 1, \ldots, q$. $\tilde{X}_k$ and $\tilde{Z}_j$ are used in covariate-adaptive randomization process. We further define all levels of imbalance between patients in two treatments. Consider $\tilde{X}_k$ have $s_k^*$ levels and $\tilde{Z}_j$ have $s_j$ levels, resulting in $\prod_{k=1}^p s_k^* \prod_{j=1}^q s_j$ strata in total. Let $W_i = (\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,p}, \tilde{Z}_{i,1}, \ldots, \tilde{Z}_{i,q})$ represent the covariate profile of the ith patient used in randomization, that is, $W_i = (x_{i,1}^0, x_{i,2}^0, \ldots, x_{i,p}^0, z_{i,1}^0, z_{i,2}^0, \ldots, z_{i,q}^0)$. If $\tilde{X}_{i,k}$ is at level $x_{i,k}^0$ and $\tilde{Z}_{i,j}$ is at level $z_{i,j}^0$. For convenience, we use $(t_1, t_2, \ldots, t_p, r_1, r_2, \ldots, r_q)$ to denote the stratum formed by patients who have the same covariate profile $(x_{i,1}^0, x_{i,2}^0, \ldots, x_{i,p}^0, z_{i,1}^0, z_{i,2}^0, \ldots, z_{i,q}^0)$, and use $(k; t_k)$ to denote the margin formed by patients with $X_k = x_k^0$, and similarly $(j, r_j)$ to denote the margin formed by patients with $Z_j = z_j^0$. Then let

- $D_N$ be the difference between the numbers of patients in treatment group 1 and 2 as total, that is, the number in group 1 minus the number in group 2;
- $D_N(k; t_k)$ and $D_N(j; r_j)$ be the differences between the numbers of patients in the two treatment groups on the margin $(k; t_k)$ and $(j, r_j)$, respectively;
- $D_N(t_1, t_2, \ldots, t_p, r_1, r_2, \ldots, r_q)$ be the difference between the numbers of patients in the two treatment groups within the stratum $(t_1, t_2, \ldots, t_p, r_1, r_2, \ldots, r_q)$.

These differences play important roles in properties of statistical inference for covariate-adaptive designs (see Section 3 for details).

### 3. THEORETICAL PROPERTIES OF HYPOTHESIS TESTING

Two types of hypothesis testing will be considered in this section, one is comparing main treatment effects between two groups, and the other is testing significance of covariates. The testing hypotheses are performed based on the working model (2) when data are generated from the true model (1). Properties of this hypothesis testing are studied under both null hypothesis and alternative hypothesis. A test is said to be (asymptotically) conservative, if the true Type I error is smaller than the significance level under the null hypothesis.

First we consider the hypothesis tests of comparing treatment effects. We have the following main theorem.

**Theorem 3.1.** Suppose that a covariate-adaptive design satisfies the following two conditions:

A. the overall imbalance is bounded in probability, that is, $D_N = O_p(1)$;
B. the marginal imbalances for all covariates are bounded in probability, that is, $D_N(k; t_k) = O_p(1)$ and $D_N(j; r_j) = O_p(1)$, $k = 1, 2, \ldots, p$, $j = 1, 2, \ldots, q$.

Then:
(i) Under $H_0 : \mu_1 - \mu_2 = 0$,

$$T \xrightarrow{D} N(0, \tau^2), \quad \tau^2 = \frac{\sigma^2 + \sum_{j \in C} \gamma_j^2 \sigma^2_{\delta,j}}{\sigma^2 + \sum_{j=1}^q \gamma_j^2 \text{var}(Z_j)}.$$

where $\sigma^2 = \frac{\sigma^2}{\sigma^2_z}$, $\sigma^2_{\delta,j} = \frac{\gamma_j^2 \sigma^2_{\delta,j}}{\sigma^2_z}$, and $\sigma^2_z = E[\text{var}(\delta_{ij}|d_j(Z_{ij}))].$

Hence,

(a) If $\gamma_j = 0, j = 1, 2, \ldots, q$, then $\tau = 1$. Thus, when all the covariates $Z$s are not related to $Y$, the hypothesis testing (3) can achieve correct Type I error.

(b) If at least one $\gamma_j \neq 0, j = 1, 2, \ldots, q$, then $\tau < 1$. In this case, the hypothesis testing (3) is conservative.

(ii) Under $H_A : \mu_1 - \mu_2 \neq 0$, consider a sequence of local alternatives, that is, $\mu_1 - \mu_2 = \delta/\sqrt{N}$ for a fixed $\delta \neq 0$, then

$$T \xrightarrow{D} N(\Delta, \tau^2), \quad \Delta = \frac{\delta}{2 \sigma_z}. \quad (10)$$

**Proof.** To prove Theorem 3.1, the asymptotic properties are studied for both the numerator and the denominator of test statistic $T$. We first consider the numerator of test statistic (4),

$$L\hat{\beta} = \hat{\mu}_1 - \hat{\mu}_2 = \mu_1 - \mu_2 + L \left( \frac{X^T X}{N} \right)^{-1} \frac{X^T Z \gamma}{N} + L \left( \frac{X^T X}{N} \right)^{-1} \frac{X^T \varepsilon}{N}.$$

In the online supplementary materials it is proved that, if the conditions (A) and (B) are satisfied for a covariate-adaptive design, then

$$\sqrt{N} \left[ L \left( \frac{X^T X}{N} \right)^{-1} \frac{X^T Z \gamma}{N} + L \left( \frac{X^T X}{N} \right)^{-1} \frac{X^T \varepsilon}{N} \right] \xrightarrow{D} N \left( 0, 4 \left( \sum_{j \in C} \gamma_j^2 \sigma^2_{\delta,j} + \sigma^2_z \right) \right),$$

where $\sigma^2_{\delta,j} = E[\text{var}(\delta_{ij}|d_j(Z_{ij})), j \in C$. Therefore,

$$\sqrt{N}[\hat{\mu}_1 - \hat{\mu}_2] - (\mu_1 - \mu_2)] \xrightarrow{D} N \left( 0, 4 \left( \sum_{j \in C} \gamma_j^2 \sigma^2_{\delta,j} + \sigma^2_z \right) \right).$$

Further, for the denominator of test statistic $T$, we could also show that,

$$\delta^2 L(X^T X)^{-1} L^T = \frac{4}{N} \left[ \sigma^2_z + \sum_{j=1}^q \gamma_j^2 \text{var}(Z_j) \right] + o_p \left( \frac{1}{N} \right).$$

Then, it follows from Slutsky’s Theorem that, under $H_0 : \mu_1 - \mu_2 = 0$,

$$T \xrightarrow{D} N(0, \tau^2), \quad \tau^2 = \frac{\sigma^2_z}{\sigma^2_z},$$

where $\sigma^2_z = \sigma^2 + \sum_{j=1}^q \gamma_j^2 \text{var}(Z_j)$ and $\sigma^2_z = \sigma^2 + \sum_{j \in C} \gamma_j^2 \sigma^2_{\delta,j}.$

Similarly, under $H_A : \mu_1 - \mu_2 \neq 0$ with a sequence of local alternatives, that is, $\mu_1 - \mu_2 = \delta/\sqrt{N}$ for a fixed $\delta \neq 0$,

$$T \xrightarrow{D} N(\Delta, \tau^2), \quad \Delta = \frac{\delta}{2 \sigma_z}.$$

The detailed proofs of Theorem 3.1 and two related lemmas are in the online supplementary materials. □

In Theorem 3.1, theoretical properties of hypothesis testing for treatment effects are obtained under covariate-adaptive designs. In a covariate-adaptive randomized clinical trial, covariate information is incorporated in the design process to reduce imbalance of different levels (within-stratum, within-covariate-margin, and overall). Two mild conditions of covariate-adaptive designs are assumed to derive the asymptotic distribution of the test statistic for comparing treatment effects. Condition (A) states that the overall imbalance is bounded in probability and condition (B) requires that marginal imbalances are all bounded in probability. These conditions are satisfied by various covariate-adaptive designs (see, e.g., Corollary 3.1 and Theorem 3.3). Under these conditions, it can be shown in the proof that the numerator ($\hat{\mu}_1 - \hat{\mu}_2$), that is, the estimate of difference between treatment effects, has a variance $4\sigma^2_z/N$ which is smaller than the model-based variance estimate $4\sigma_z^2/N$ in the denominator. The reduction of variance of ($\hat{\mu}_1 - \hat{\mu}_2$) can be attributed to balanced distributions of covariate profiles across treatment groups, which enhance the comparability of different treatment groups and eliminate the variability between estimates of treatment effects. When a linear model with omitting covariates is implemented to do inference, the covariate-adaptive randomization scheme is ignored and an inflated model-based variance estimate is used, which results in a variance smaller than 1 in the asymptotic distribution under the null hypothesis and thus conservative Type I error smaller than the nominal level. Furthermore, power performance can also be assessed based on the asymptotic distribution of the test statistic under the alternative hypothesis.

Now consider the power of the hypothesis test (3). Under the alternative hypothesis, the power is

$$pr(|T| > Z_{1-\alpha/2}) = \Phi \left( \frac{\delta}{2\sigma_\delta} - \frac{\sigma_z Z_{1-\alpha/2}}{\sigma_\delta} \right) + \Phi \left( -\frac{\delta}{2\sigma_\delta} - \frac{\sigma_z Z_{1-\alpha/2}}{\sigma_\delta} \right) + o(1).$$

The power of test (3) under complete randomization (under the same setting as described in Section 2) is

$$pr(|T| > Z_{1-\alpha/2}) = \Phi \left( \frac{\delta}{2\sigma_z} - Z_{1-\alpha/2} \right) + \Phi \left( -\frac{\delta}{2\sigma_z} - Z_{1-\alpha/2} \right) + o(1).$$

From the power expressions for both covariate-adaptive designs and complete randomization above, we can conclude that the limiting power under covariate-adaptive designs could be smaller than that under complete randomization when $\delta$ is relatively small, and it is usually larger than complete randomization
when $\delta$ is large. This conclusion agrees with some simulation studies about two-sample $t$-test in the literature (Forsythe 1987; Shao, Yu, and Zhong 2010) for some covariate-adaptive designs. Our conclusion is more general, which can be applied to linear models under a large family of covariate-adaptive designs.

The following theorem shows that hypothesis tests regarding significance of covariates can still achieve correct Type I error in covariate-adaptive designs, even though the power would be influenced if not all covariates are incorporated in the analysis model.

**Theorem 3.2.** Under the same conditions as in Theorem 3.1:

(i) Under $H_0 : C \beta = \xi_0$, 

$$T^* \xrightarrow{D} \chi^2_{m}/m. \quad (11)$$

Hence, the hypothesis testing (5) can achieve correct Type I error.

(ii) Under $H_A : C \beta = \xi_1$, consider a sequence of local alternatives, that is, $(\xi_1 - \xi_0) = \eta/\sqrt{N}$ for a fixed $\eta \neq 0$, then

$$T^* \xrightarrow{D} \chi^2_{m}(\lambda)/m,$$

$$\lambda = \eta^2 [CM^{-1}C^T]^{-1}\eta/\sigma_2^2, \quad (12)$$

where $M = \text{diag}(1/2, 1/2, \text{var}(X_1), \ldots, \text{var}(X_p))$ and $\lambda$ is the noncentral parameter. Therefore, the power will increase as more covariates are incorporated into model.

**Proof.** The test statistic can be written as

$$T^* = m^{-1}\sqrt{N}(C\hat{\beta} - \xi_0)^T[C(X^TX/N)C^T]^{-1}\sqrt{N}(C\hat{\beta} - \xi_0).$$

We first have, under $H_0 : C \beta = \xi_0$, 

$$\sqrt{N}(C\hat{\beta} - \xi_0) = \sqrt{N}C(\hat{\beta} - \beta),$$

and under $H_A : C \beta = \xi_1$, 

$$\sqrt{N}(C\hat{\beta} - \xi_0) = \sqrt{N}C(\hat{\beta} - \beta + \xi_1 - \xi_0) = \sqrt{N}C(\hat{\beta} - \beta) + \sqrt{N}(\xi_1 - \xi_0).$$

Further, by the weak law of large numbers and independence of covariates,

$$\frac{1}{N}X^TX \rightarrow_p \text{diag}\left(\frac{1}{2}, \frac{1}{2}, \text{var}(X_1), \ldots, \text{var}(X_p)\right) = M.$$

Then it follows that,

$$\sqrt{N}C(\hat{\beta} - \beta) = \sqrt{N}C\left[\left(\frac{X^TX}{N}\right)^{-1}\frac{X^TZ\gamma}{N}\right]$$

$$\quad + \left(\frac{X^TX}{N}\right)^{-1}\frac{X^Te}{N}$$

$$= \sqrt{N}CM^{-1}\left[\frac{X^TZ\gamma}{N} + \frac{X^Te}{N}\right] + o_p(1).$$

By defining $M = \text{diag}(1/2, 1/2, \hat{\beta})$ and $C = [0_{m \times 2}, \hat{\beta}]$, where $0_{m \times 2}$ is an $(m \times 2)$ matrix with all entries equal to zero, we have

$$\sqrt{N} C(\hat{\beta} - \beta)$$

$$= \tilde{C} \tilde{M}^{-1} \frac{1}{\sqrt{N}} \left[ \sum_j \sum_i X_{i,j}Z_{i,j}\gamma_j + \sum_i X_{i,1}\varepsilon_i \right] + o_p(1).$$

By the central limit theorem and the fact $CM^{-1}C^T = \tilde{C} \tilde{M}^{-1}\hat{C}^T$, we can prove the results. The detailed proofs of Theorem 3.2 are in the online supplementary materials.

From Theorems 3.1 and 3.2, it can be seen that the overall difference and marginal difference play important roles in statistical inference for covariate-adaptive designs. There is no constraint on within-stratum imbalance. For stratified permuted block design, the difference within any stratum is at maximum the half of block size. Since the number of strata is finite, the overall and marginal difference are less than a constant, thus the conditions (A) and (B) are satisfied. Hu and Hu (2012) proposed a large class of covariate-adaptive designs, which satisfy the conditions (A) and (B) under certain conditions. Here we summarize these results in the corollary.

**Corollary 3.1.** Both Theorems 3.1 and 3.2 hold under the following covariate-adaptive designs:

(i) stratified permuted block designs; and

(ii) the class of covariate-adaptive designs proposed by Hu and Hu (2012).

The theoretical properties for Pocock and Simon’s marginal procedure remain unknown for decades. The next theorem demonstrates that marginal difference and overall difference are bounded in probability for Pocock and Simon’s marginal procedure.

**Theorem 3.3.** Under the Pocock and Simon’s marginal procedure (Pocock and Simon 1975), the conditions (A) and (B) are satisfied. Therefore, both Theorems 3.1 and 3.2 hold under Pocock and Simon’s marginal procedures.

**Proof.** Because here we consider only the problem of imbalances for covariates, we can assume that $q = 0$ and all covariates are discrete, for otherwise we can combine $(X_1, \ldots, X_p)$ and $(Z_1, \ldots, Z_q)$ together as a new covariate vector and consider $W_n$ instead. We denote $D_n = (D_n(k; t_k); t_k = 1, \ldots, s^*_k; k = 1, \ldots, p)$ to be the collection of all marginal imbalances after $n$ assignments. Further, at the stage $n$, when the $n$th patient falls within stratum $t^* = (t^*_1, \ldots, t^*_p)$, that is, $W_n = (x^*_1, \ldots, x^*_p)$, we define

$$\text{Im}^1_n = \sum_{k=1}^{p} w_k [D_{n-1}(k; t^*_k) + 1]^2,$$

$$\text{Im}^2_n = \sum_{k=1}^{p} w_k [D_{n-1}(k; t^*_k) - 1]^2,$$

which are the weighted “potential” imbalances that would be caused if the $n$th patient was assigned to treatment 1, 2, respectively. Here $w_k > 0$, $k = 1, \ldots, p$, are weights placed within a covariate margin. Then, under the Pocock and Simon’s randomization procedure, the probability of assigning the $n$th patient
to treatment 1 is \( q \) if \( \text{Imb}^{(1)} > \text{Imb}^{(2)} \), \( 1 - q \) if \( \text{Imb}^{(1)} < \text{Imb}^{(2)} \), and \( 1/2 \) if \( \text{Imb}^{(1)} = \text{Imb}^{(2)} \), where \( 0 < q < 1/2 \).

By defining \( \Lambda_{n-1}(t) = \sum_{k=1}^{p} w_k D_{n-1}(k; t_k) \) and a norm-like function of \( D_n \) by \( V(D_n) = \sum_{k=1}^{p} \sum_{s=1}^{q} w_k D_{n}^{2}(k; t_k) \), we can verify that \( (D_n)_{n \geq 1} \) is a Markov chain with

\[
E \left[ V(D_n) | D_{n-1} \right] = V(D_{n-1}) - 2(1 - 2q) \sum_{t} |\Lambda_{n-1}(t)| p(t),
\]

where \( p(t) = \text{pr}(W_n = (x^t_1, \ldots, x^t_p)) \). Further,

\[
\sum_{t} |\Lambda_{n-1}(t)| p(t) \text{ is bounded } \iff \{D_{n-1}\} \text{ is bounded}.
\]

So, there exist a bounded set \( \mathcal{C} \) in \( \mathbb{R}^p \) and a constant \( b \) such that

\[
E \left[ V(D_n) | D_{n-1} \right] - V(D_{n-1}) \leq \begin{cases} 
-1 & \text{if } D_{n-1} \notin \mathcal{C}, \\
b & \text{if } D_{n-1} \in \mathcal{C}.
\end{cases}
\]

The above drift condition implies that the sequence \( (D_n)_{n \geq 1} \) is a positive (Harris) recurrent Markov chain (see Theorem 11.3.4 of Meyn and Tweedie 1993), and so it is bounded in probability. It follows that \( D_N(k; t_k) \sim \text{Op}(1) \) for all \( t_k \) and \( k \). Finally, note \( \sum_{i=1}^{q} D_N(k; t_k) = D_N \). So, \( D_N \sim \text{Op}(1) \). Therefore, Conditions (A) and (B) are satisfied. The detailed proofs of Theorem 3.3 are in the online supplementary materials.

**Remark.** If we consider complete randomization as a special case of covariate-adaptive design, it does not satisfy the condition on the overall imbalance in Theorem 3.1, because the overall imbalance \( D_N = \sum_{i=1}^{N} (2I_i - 1) = \text{Op}(N^{1/2}) \) by the independence of \( I_i \) and the central limit theorem. The numerical study in next section shows that the test under complete randomization is not conservative.

Based on Theorem 3.2, we can see that hypothesis testing of covariates is still valid in the sense of Type I error under all covariate-adaptive designs. The regression method can be directly used to test significance of prognostic factors with a working model only containing partial covariate information. On the other hand, power will decrease by omitting influential covariates in the working model. Consider the noncentral parameter in (12): it would increase with \( \sigma_i^2 \) reduced, so power would increase with more influential covariates in the model. Therefore, it is helpful to incorporate more influential covariates, if possible, to obtain a more powerful test.

**Corollary 3.2.** Under the same conditions as in Theorem 3.1:

(i) Under \( H_0 : \beta_1 = 0 \),

\[
T_1 \xrightarrow{D} N(0, 1).
\]

Hence, the hypothesis testing (7) can achieve correct Type I error.

(ii) Under \( H_A : \beta_1 \neq 0 \), consider a sequence of local alternatives, that is, \( \beta_1 = \delta \beta_1 / \sqrt{N} \) for a fixed \( \delta \), then

\[
T_1 \xrightarrow{D} N(\Delta \beta_1, 1), \quad \Delta \beta_1 = \delta \beta_1 \sigma_1 / \sigma_e,
\]

where \( \sigma_1^2 = \text{var}(X_1) \). Hence, the power will increase as more covariates are incorporated into the model.

Corollary 3.2 gives an important special case of testing covariates, where only a single coefficient is considered.

According to Theorems 3.1 and 3.2, we should select a model with only influential covariates. It is known that too many unnecessary variables in the model increases the variations of estimates and impacts the statistical results. Hence, if only influential variables are incorporated in the model, it will not only reduce unnecessary variations, but will also give valid inference. Some numerical studies are performed in Section 4 about model selections.

### 4. NUMERICAL STUDY

**Case 1: Testing Treatment Effects.** First we consider simulations to study Type I error of hypothesis testing for comparing treatment effects under three designs: Pocock and Simon’s marginal procedure, stratified permuted block design, and complete randomization. For each type of design, both continuous and discrete case are considered. The following linear model (including two covariates \( Z_1 \) and \( Z_2 \)) is assumed for responses \( Y_i \),

\[
Y_i = \mu_1 I_i + \mu_2 (1 - I_i) + \beta_1 Z_{1,i} + \beta_2 Z_{2,i} + \epsilon_i,
\]

where \( \epsilon_i \) is distributed as \( N(0, 1) \), \( \beta_1 = \beta_2 = 1 \). No difference in treatment effects is assumed to study Type I error, that is, \( \mu_1 = \mu_2 \). For the discrete case, \( Z_1 \) follows Bernoulli(\( p_1 \)) and \( Z_2 \) follows Bernoulli(\( p_2 \)); for the continuous case, both \( Z_1 \) and \( Z_2 \) follow normal distributions \( N(0, 1) \). If covariates \( Z_1 \) and \( Z_2 \) are continuous, they are discretized into Bernoulli variables \( Z_1' \) and \( Z_2' \) with the probabilities \( p_1 \) and \( p_2 \) to be used in randomization. More specifically, if \( Z_1 < Z_1(p_1) \), where \( Z_1(p_1) \) is \( p_1 \) quantile of the standard normal distribution, then \( Z_1' = 0 \), otherwise \( Z_1' = 1 \). Original variables (without discretization) are used in statistical inference procedures.

To carry out simulations, the biased coin probability \( 0.75 \) and equal weights are used for Pocock and Simon’s marginal procedure, and the block size \( 4 \) is used for stratified permuted block design. The significance level is \( \alpha = 0.05 \) and sample size \( N = 100, 200, \) or \( 500 \). The hypothesis tests include the two-sample t-test (t-test), the linear model with a single covariate \( Z_1 \) (\( \text{lm}(Z_1) \)), the linear model with a single covariate \( Z_2 \) (\( \text{lm}(Z_2) \)), and the linear model with both covariates \( Z_1 \) and \( Z_2 \) (\( \text{lm}(Z_1, Z_2) \)). By choosing \( (p_1, p_2) = (0.5, 0.5) \), the simulation results for Pocock and Simon’s marginal procedure, stratified permuted block design and complete randomization are demonstrated in Table 1.

In each simulation, Type I error of covariate-adaptive randomization methods is also examined with the bootstrap t-test described in Shao, Yu, and Zhong (2010). To do the test, \( B \) bootstrap samples \( (Y_i^{ab}, Z_{1,i}^{ab}, Z_{2,i}^{ab}) \), \( i = 1, 2, \ldots, B \), are generated independently as simple random samples with replacement from \( (Y_i, Z_{1,i}, Z_{2,i}) \), \( i = 1, 2, \ldots, B \), \( Y_i, Z_{1,i}, Z_{2,i} \) are the covariate-adaptive procedure on the original data is applied on the covariates of each bootstrap sample \( (Z_{1,i}^{ab}, Z_{2,i}^{ab}) \), \( i = 1, 2, \ldots, B \), from which the bootstrap analogues of treatment assignments, \( I_1^{ab}, \ldots, I_N^{ab} \) can be obtained.
Table 1. Simulated Type I error for Pocock and Simon’s marginal procedure (PS), stratified permuted block design (SPB) and complete randomization (CR) in %; simulations based on 10,000 runs

| Z     | Method | N   | t-test lm(Z₁) | lm(Z₂) | lm(Z₁, Z₂) | BS-t |
|-------|--------|-----|---------------|--------|------------|------|
| Discrete | PS     | 100 | 1.75          | 3.05   | 3.09       | 5.21 | 5.18 |
|        |        | 200 | 1.62          | 2.78   | 2.86       | 4.99 | 4.88 |
|        |        | 500 | 1.66          | 2.81   | 2.77       | 4.87 | 4.90 |
|        | SPB    | 100 | 1.85          | 2.86   | 3.05       | 5.29 | 5.67 |
|        |        | 200 | 1.54          | 2.69   | 2.73       | 4.84 | 4.95 |
|        |        | 500 | 1.55          | 2.77   | 2.65       | 4.84 | 5.60 |
|        | CR     | 100 | 5.04          | 5.27   | 5.11       | 5.31 | –    |
|        |        | 200 | 5.00          | 4.95   | 5.12       | 5.21 | –    |
|        |        | 500 | 4.73          | 4.83   | 4.68       | 4.77 | –    |
| Continuous | PS     | 100 | 1.43          | 2.15   | 2.02       | 4.98 | 5.16 |
|        |        | 200 | 1.07          | 1.74   | 1.80       | 4.53 | 5.62 |
|        |        | 500 | 0.91          | 1.72   | 1.73       | 4.72 | 4.79 |
|        | SPB    | 100 | 1.22          | 1.83   | 2.05       | 5.01 | 5.68 |
|        |        | 200 | 0.98          | 1.86   | 1.77       | 5.08 | 5.19 |
|        |        | 500 | 1.15          | 1.98   | 1.84       | 5.48 | 5.61 |
|        | CR     | 100 | 5.20          | 5.31   | 4.82       | 4.92 | –    |
|        |        | 200 | 5.06          | 5.14   | 4.85       | 5.46 | –    |
|        |        | 500 | 4.87          | 5.05   | 4.71       | 4.77 | –    |

Define

\[
\bar{Y}_1 - \bar{Y}_2 = \frac{1}{n_1} \sum_{i=1}^{N} I_i Y_i - \frac{1}{n_2} \sum_{i=1}^{N} (1 - I_i) Y_i, \quad n_1 = \sum_{i=1}^{N} I_i, \\
\quad n_2 = \sum_{i=1}^{N} (1 - I_i),
\]

and

\[
\hat{\theta}^*(b) = \frac{1}{n_1^b} \sum_{i=1}^{N} I_i^b Y_i^b - \frac{1}{n_2^b} \sum_{i=1}^{N} (1 - I_i^b) Y_i^b,
\]

\[
n_1^b = \sum_{i=1}^{N} I_i^b, \quad n_2^b = \sum_{i=1}^{N} (1 - I_i^b).
\]

The bootstrap estimator of the variance of \( \bar{Y}_1 - \bar{Y}_2 \) is then the sample variance of \( \hat{\theta}^*(b) \), \( b = 1, 2, \ldots, B \), represented by \( \hat{\theta}_B \). Then the bootstrap t-test has the form of \( T_B = (\bar{Y}_1 - \bar{Y}_2)/\hat{\theta}_B^{1/2} \). Shao, Yu, and Zhong (2010) showed that the bootstrap t-test can maintain nominal Type I error under covariate-adaptive biased coin design. \( B = 500 \) is used in all following simulations.

Several conclusions can be drawn from Table 1. First, the Type I error is close to 5% for all four tests. We also tried different sample variance of \( \hat{\theta} \). Second, under both Pocock and Simon’s marginal procedure and stratified permuted block design, the two-sample t-test, \( \text{lm}(Z_1) \) and \( \text{lm}(Z_2) \) are all conservative. Among these three tests, the two-sample t-test is the most conservative one with the least Type I error. Furthermore, the Type I error of the bootstrap t-test (BS-t) is close to the nominal level 5% under both Pocock and Simon’s marginal procedure and stratified permuted block design. Under complete randomization, the Type I error is close to 5% for all four tests. We also tried different \( (p_1, p_2) \), similar results are obtained and are not shown here.

Case 2: Power Comparison. Now we compare power for different hypothesis testing methods under Pocock and Simon’s marginal procedure and complete randomization. The same model as in Case 1 is used, except that difference exists

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Simulated power for Pocock and Simon’s marginal procedure (PS) and complete randomization (CR) under discrete (D) and continuous (C) cases. Simulation based on 10,000 runs and number of patients \( N = 100 \).}
\end{figure}
Table 2. Power comparison for Pocock and Simon’s marginal procedure and complete randomization, simulation based on 10,000 runs and sample sizes $N = 32$ and 64.

| $\mu_1 - \mu_0$ | $N = 32$ | | $N = 64$ |
|------------------|-----------|-----------|-----------|
|                  | CR        | PS        | CR        | PS        |
| 0.0              | 4.96      | 5.03      | 5.17      | 5.08      |
| 0.1              | 5.86      | 6.35      | 6.50      | 6.48      |
| 0.2              | 7.81      | 8.51      | 12.12     | 12.68     |
| 0.3              | 12.34     | 13.26     | 22.06     | 22.21     |
| 0.4              | 18.15     | 19.44     | 34.46     | 34.76     |
| 0.5              | 25.28     | 26.54     | 47.99     | 49.35     |
| 0.6              | 33.96     | 36.98     | 63.04     | 65.53     |
| 0.7              | 44.16     | 47.83     | 76.72     | 78.13     |
| 0.8              | 53.74     | 57.28     | 86.97     | 87.95     |
| 0.9              | 64.30     | 68.70     | 93.14     | 94.24     |
| 1.0              | 73.63     | 77.10     | 97.02     | 97.51     |

Table 3. Simulated Type I error for $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ for Pocock and Simon’s marginal procedure (PS) and complete randomization (CR) in %; simulations based on 10,000 runs.

| $Z$ | $N$ | PS | CR | PS | CR |
|-----|-----|----|----|----|----|
| Discrete | 100 | 4.96 | 4.98 | 4.93 | 4.88 |
|       | 200 | 5.35 | 5.28 | 4.76 | 4.82 |
|       | 500 | 5.55 | 5.55 | 5.22 | 5.20 |
| Continuous | 100 | 4.98 | 4.90 | 5.10 | 4.98 |
|         | 200 | 5.14 | 5.14 | 5.05 | 5.18 |
|         | 500 | 5.11 | 5.15 | 5.28 | 5.29 |

We further consider the power comparison for small sample sizes 32 and 64. The simulated power is reported in Table 2 under the same model as in Case 1. In Pocock and Simon’s marginal procedure, the probability of biased coin assignment is 0.8 and equal weights are assigned to the two covariates. The hypothesis testing is based on the model $(\text{lm}(Z_1, Z_2))$ with both covariates $Z_1$ and $Z_2$. Table 2 indicates the advantages between treatment effects $\mu_1$ and $\mu_2$, that is, $\mu_1 - \mu_2 \neq 0$. Sample size $N = 100$ and $(p_1, p_2) = (0.5, 0.5)$ are used in simulations. Several hypothesis testing methods for treatment effects are compared under covariate-adaptive randomization. All the results of power are given in Figure 1, from which several conclusions can be made. The two-sample $t$-test is less powerful than $(\text{lm}(Z_1))$ and $(\text{lm}(Z_2))$, and all those three methods are less powerful than $(\text{lm}(Z_1, Z_2))$ under Pocock and Simon’s marginal procedure. The bootstrap $t$-test has similar power performance as the full model with correct model specification $(\text{lm}(Z_1, Z_2))$ when covariates are discrete. However, the bootstrap $t$-test is less powerful than $(\text{lm}(Z_1, Z_2))$ if covariates are continuous. Furthermore, the power of each test can also be compared between covariate-adaptive randomization and complete randomization. For example, the two-sample $t$-test has smaller power under Pocock and Simon’s marginal procedure than that under complete randomization when $|\mu_1 - \mu_2|$ is relatively small due to conservativeness, but has larger power when $|\mu_1 - \mu_2|$ becomes larger. We also tried stratified permuted block randomization (not reported here), all tests considered have similar power performance as Pocock and Simon’s marginal procedure. Based on Table 1 and Figure 1, we can see that the theoretical properties in Theorem 3.1 hold for sample size around 100.
Table 4. Simulated Type I error under model (14) with backward model selection techniques for Pocock and Simon’s marginal procedure (PS) and complete randomization (CR). The number in parenthesis is the percentage of correct selection; simulation based on 5000 runs.

|   | N   | t-test | lm3  | lm5  | Backward | BS-t |
|---|-----|--------|------|------|----------|------|
| PS| 50  | 1.12   | 4.64 | 4.74 | 4.88(84.34%) | 5.44 |
|   | 100 | 0.74   | 4.98 | 5.06 | 5.02(94.30%) | 5.08 |
|   | 200 | 0.40   | 4.96 | 4.96 | 4.94(96.70%) | 4.70 |
|   | 500 | 0.38   | 5.24 | 5.18 | 5.26(98.84%) | 5.06 |
| CR| 50  | 5.32   | 4.94 | 4.94 | 5.66(83.60%) | –    |
|   | 100 | 5.54   | 5.26 | 5.44 | 5.36(94.40%) | –    |
|   | 200 | 5.06   | 4.98 | 4.82 | 5.02(97.12%) | –    |
|   | 500 | 4.80   | 4.92 | 4.94 | 4.92(98.76%) | –    |

of covariate-adaptive designs over complete randomization for small sample size.

Case 3: Significance of Covariates. Type I error of testing (7) for significance of covariates is studied under Pocock and Simon’s marginal procedure and complete randomization. As the same model in Case 1, we set $\beta_1 = 0$, $\beta_2 = 1$ and $\mu_1 = \mu_2$. To run simulations, the biased coin probability 0.75 and equal weights are used for Pocock and Simon’s marginal procedure. The significance level is $\alpha = 0.05$ and sample size $N$ is 100, 200, or 500. Hypothesis testing of the significance of $\beta_1$ is performed based on the two models with or without $Z_2$, that is, lm(Z1, Z2) and lm(Z1). The Type I error results are given in Table 3, from which it can be seen that the tests of $\beta_1$ are valid in terms of Type I error for both Pocock and Simon’s marginal procedure and complete randomization. Figure 2 reports the power of testing $\beta_1$ under two randomization methods, which indicates that the model lm(Z1, Z2) with $Z_2$ in the analysis is more powerful than lm(Z1). This agrees with Theorem 3.2.

Case 4: Model Selection. So far, all simulation results are based on linear models with up to two covariates. Now we consider a linear model with more than two covariates. In this situation, variable selection techniques can be applied to select a subset of influential covariates on outcomes to be used in inference procedures. Suppose outcomes $Y_i$ follows the following model with five covariates,

$$Y_i = \mu_1 I_i + \mu_2(1 - I_i) + X_{i1} \beta_1 + X_{i2} \beta_2^T + X_{i3} \beta_3 + Z_{i1} \gamma_1^T + Z_{i2} \gamma_2 + \epsilon_i,$$

(14)

where $X_1$ is a binary covariate with the probability of 0.5 to take 0 or 1, $X_2$ is a discrete variable with four possible values (0, 0, 0), (1, 0, 0), (0, 1, 0), and (0, 0, 1) for equal probabilities 0.25, $Z_1$ is a discrete covariate with three possible values (0, 0), (0, 1), and (1, 0) for equal probabilities 1/3, $X_3$, and $Z_2$ are standard normal distributed variables, which are discretized to Bernoulli variables with the probability 0.5 for randomization process. Also it is assumed $\beta_1 = 3$, $\beta_2 = (2, 3, 4)$, $\beta_3 = 2$, $\beta_4 = (0, 0)$, $\beta_5 = 0$, so that only $X_1$, $X_2$, and $X_3$ have effects on the outcome $Y$. $\epsilon_i$ is normally distributed with a mean of 0 and a standard deviation of 2. Here only Pocock and Simon’s marginal procedure is considered, since there would be too many strata (total 96 strata) if stratified permuted block design is implemented.

To adjust conservative Type I error, one way is to incorporate all randomization covariates into the analysis model (lm5), where there is no loss of information used in hypothesis testing compared to randomization. A more efficient approach according to Theorem 3.1 is to incorporate only the covariates that are influential on outcomes, that is, constructing the model

![Figure 3. Simulated power under model (14) with backward model selection techniques for Pocock and Simon’s marginal procedure (PS) and complete randomization (CR). Simulation based on 5000 runs and number of patients $N = 100$.](image)
Table 5. Simulated Type I error under model (15) for Pocock and Simon’s marginal procedure (PS), stratified permuted block design (SPB) and complete randomization (CR); simulation based on 10,000 runs

| N  | t-test | lm(Z₁, Z₂) | BS-t |
|----|--------|-------------|------|
| PS | 100    | 1.70        | 3.87 | 3.90 |
|    | 200    | 2.28        | 4.01 | 4.70 |
|    | 500    | 2.32        | 4.11 | 4.25 |
| SPB| 100    | 1.59        | 3.66 | 4.96 |
|    | 200    | 1.70        | 3.27 | 4.46 |
|    | 500    | 1.98        | 3.55 | 4.70 |
| CR | 100    | 3.33        | 3.90 | –   |
|    | 200    | 4.01        | 4.03 | –   |
|    | 500    | 4.47        | 4.11 | –   |

with X₁, X₂, and X₃ (lm3). In Table 4, the model selection with BIC by stepwise algorithm is used to select the analysis model to do statistical inference. The algorithm is realized with “stepAIC” function in R by specifying k = log(N). Stepwise selection can be implemented backward, forward, and with both directions. Here we report only the results for the backward selection(backward). Similar results are obtained from the other two stepwise selection methods (not shown). These methods are able to automatically achieve a final model from a bunch of candidate models with different combinations of multiple covariates, based on which treatment effects can be compared. The results of Type I error of lm3, lm5, and backward are given in Table 4. The results of the two-sample t-test (t-test) and the bootstrap t-test (BS-t) are also included. Power comparison results are shown in Figure 3. From the results of Table 4 and Figure 3, hypothesis testing for treatment effects based on conventional stepwise model selection techniques has valid Type I error and has similar power performance compared to the hypothesis testing with all randomization covariates used in the analysis, as long as the model selection techniques are able to identify the subset of all influential covariates. It is worth pointing out that the bootstrap t-test is less powerful than lm3, lm5, and backward based on Figure 3.

Case 5: Model Misspecification. In Section 2, the underlying response model (1) is assumed to have a form of linearly additive covariate effects in addition to treatment effects. Based on the model, theoretical properties of hypothesis testing are studied under covariate-adaptive designs. Theorem 3.1 shows that valid Type I error can be obtained by using a linear model incorporating all randomization covariates (the full model). This is within expectation since the full model has the correct model specification that coincides with the assumed model. However, the underlying response model is often unknown in practice and it is possible that covariate effects are not linearly additive on responses. For example, a covariate may have a nonlinear form or have interaction effect with other covariates. Under these scenarios, the full model with all randomization covariates in a linearly additive pattern no longer has the correct model specification. In this section, a nonlinear response model with covariate effects in an exponential function is assumed to investigate properties of hypothesis testing for comparing treatment effects. Other situations of model misspecification can be studied similarly and will be left for future research projects. The following model is assumed for the responses Y. Two covariates Z₁ and Z₂ have a nonlinear effect on the response Y in addition to treatment effects.

\[ Y_i = \mu_1 I_i + \mu_2 (1 - I_i) + \exp(Z_{1i} + Z_{2i}) + \epsilon_i, \quad (15) \]

where \( \epsilon_i \) is distributed as \( N(0, 4) \). Z₁ and Z₂ are distributed as normal distributions \( N(0, 1) \), which are discretized into Bernoulli variables with the probability 0.5 to be used in randomization procedures. Three randomization methods are investigated, including Pocock and Simon’s marginal procedure, stratified permuted block design, and complete randomization. The two-sample t-test (t-test), the linear model incorporating both covariates (lm(Z₁, Z₂)), and the bootstrap t-test (BS-t) are used to compare treatment effects. The results of Type I error and power are given in Tables 5 and 6, respectively.

From Table 5, the two-sample t-test is conservative under covariate-adaptive designs. Due to model misspecification, the Type I error of the linear model (lm(Z₁, Z₂)) is deviated from the nominal level 5%, which is particularly obvious under the stratified permuted block design. However, the bootstrap t-test tends to be robust under covariate-adaptive designs with the Type I error closer to 5%. Under complete randomization, the

Table 6. Simulated power under model (15) for Pocock and Simon’s marginal procedure (PS), stratified permuted block design (SPB) and complete randomization (CR); simulation based on 10,000 runs and number of patients \( N = 100 \)

| \( \mu_1 - \mu_0 \) | t-test | lm(Z₁, Z₂) | BS-t | t-test | lm(Z₁, Z₂) | BS-t | t-test | lm(Z₁, Z₂) |
|-------------------|--------|-------------|------|--------|-------------|------|--------|-------------|
| 0.0               | 1.87   | 3.67        | 4.14 | 1.40   | 3.13        | 4.92 | 3.60   | 4.37        |
| 0.4               | 4.21   | 8.31        | 7.71 | 3.01   | 6.35        | 7.57 | 5.69   | 7.73        |
| 0.8               | 10.71  | 19.56       | 16.85| 9.26   | 17.63       | 17.14| 12.66  | 18.39       |
| 1.2               | 21.91  | 35.81       | 29.90| 19.79  | 34.39       | 30.61| 21.85  | 33.71       |
| 1.6               | 35.79  | 53.98       | 44.25| 34.73  | 52.37       | 46.14| 35.60  | 51.25       |
| 2.0               | 50.37  | 66.52       | 57.94| 49.97  | 66.99       | 60.75| 48.32  | 64.83       |
| 2.4               | 62.81  | 77.09       | 69.23| 62.34  | 76.80       | 70.67| 60.48  | 75.32       |
| 2.8               | 71.56  | 83.22       | 76.54| 71.30  | 83.59       | 77.96| 69.56  | 82.37       |
| 3.2               | 79.88  | 87.99       | 83.16| 79.09  | 88.32       | 83.93| 76.86  | 87.27       |
| 3.6               | 84.69  | 91.36       | 87.39| 84.22  | 91.25       | 87.85| 82.97  | 91.06       |
| 4.0               | 88.47  | 93.30       | 90.37| 87.87  | 93.25       | 90.43| 86.97  | 92.48       |
Type I error should be close to 5% as sample size increases based on the central limit theorem. Regarding the power comparison in Table 6, the model \( \text{Im}(Z_1, Z_2) \) has better performance than the two-sample \( t \)-test by using more covariate information under all these three design methods. In addition, the power of the model \( \text{Im}(Z_1, Z_2) \) is slightly larger than the bootstrap \( t \)-test under covariate-adaptive designs.

## 5. DISCUSSION

In this article, we studied the theoretical properties of hypothesis testing under general covariate-adaptive designs based on a linear model framework. We derived the corresponding asymptotic distributions of the testing statistics under both null and alternative hypotheses. Instead of focusing on a specific covariate-adaptive design, we studied the problem from the angle of imbalance measure of different levels (overall, marginal, within-stratum). So the conclusion can be applied to a broad range of designs, including stratified permuted block design and Pocock and Simon’s procedure. For example, to apply Theorems 3.1 and 3.2 to a specific covariate-adaptive randomized clinical trial, one just need to check the conditions (A) and (B). The results in this article provide new insights about balance and efficiency of clinical trials, and the framework can be used to study properties of other statistical methods under covariate-adaptive designs.

We study hypothesis testing based on a linear regression framework, where response of patients is a continuous variable. In reality, the endpoints of clinical trials may be other types of variable. For instance, a clinical trial can be designed to compare success rates between a new medication and a standard treatment, where the response would be a binary variable. In the literature, Feinstein and Landis (1976) and Green and Byar (1978) studied the properties of unadjusted tests under stratified randomization for binary response, and concluded that Type I error would decrease when stratified randomization is used rather than unstratified randomization. The results for other types of covariate-adaptive design, including Pocock and Simon’s procedure, remain unknown. As future research projects, we will study the behavior of conventional statistical inference for different types of response based on generalized linear models, survival models, and so on. There are some other kinds of covariate-adaptive randomization procedures in the literature, including Zelen (1974), Wei (1978), Begg and Iglewicz (1980), and Atkinson (1982). Theorems 3.1 and 3.2 may not apply to these designs, because it is unknown whether the conditions (A) and (B) remain true for these designs.

Based on Theorems 3.1 and 3.2, incorporating all randomization covariates in the final analysis model can achieve valid hypothesis testing for treatment effects and covariates. In practice, model selection techniques can also be used as in Case 4 of Section 4 to target the subset of all influential covariates, based on which valid tests can be obtained. However, sometimes not all important randomization covariates are used in the inference step for a covariate-adaptive design. Then the actual Type I error may be not equal to the nominal level when comparing treatment effects and adjustment is necessary to achieve a valid test in this situation. If no covariate is incorporated in the final analysis model, a bootstrap \( t \)-test as described in Shao, Yu, and Zhong (2010), can be implemented to restore Type I error, and this method is shown to be more powerful than the conventional two-sample \( t \)-test under covariate-adaptive designs. Similar bootstrap adjustment methods can be considered to correct the variance estimation and to make \( \tau = 1 \) in Theorem 3.1. We leave this as a future research project.

The proposed properties of hypothesis testing for covariate-adaptive designs can be generalized in several ways. First, all covariate-adaptive designs considered in this article are based on discrete covariates. We may consider covariate-adaptive designs (Lin and Su 2012; Ma and Hu 2013) that directly use continuous covariates without discretization. However, related theoretical work is limited in the literature. Second, one important assumption to derive theoretical results is the independence between covariates. We may apply the similar idea to dependent covariates by incorporating correlation structure. Third, the proposed properties are based on clinical trials with two treatments, which can be generalized to multiple treatments (Tynofeyeyev, Rosenberger, and Hu 2007). Those topics are left for future research.

### SUPPLEMENTARY MATERIALS

The supplementary materials contain detailed proofs of all theorems in the main article and intermediate lemmas.

[Received September 2013. Revised April 2014.]

### REFERENCES

Aickin, M. (2009), "A Simulation Study of the Validity and Efficiency of Design-Adaptive Allocation to Two Groups in the Regression Situation," *The International Journal of Biostatistics*, 5, 14. [670]

Anderson, H., Hopwood, P., Stephens, R. J., Thatcher, N., Cottier, B., Nicholson, M., Milroy, R., Maughan, T. S., Falk, S. J., Bond, M. G., Burt, P. A., Connolly, C. K., McIlmurray, M. B., and Carmichael, J. (2000), "Germicidal Plus Best Supportive Care (BSC) vs BSC in Inoperable Non-Small Cell Lung Cancer—A Randomized Trial With Quality of Life as the Primary Outcome," *British Journal of Cancer*, 83, 447–453. [669]

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Hugdins, L., Gong, L., Hodges, L. M., Berlin, D. S., Thorn, C. F., Sangkuhl, K., Hebert, J. M., Woon, M., Sugriya, H., Whaley, R., Morgan, A. A., Pushkarev, D., Neff, N. F., Knowles, J. W., Chou, M., Thakuria, J., Rosenbaum, A., Zaratek, A. W., Huch, G., Greely, H. T., Quake, S. R., and Altman, R. B. (2010), “Clinical Evaluation Incorporating a Personal Genome,” *The Lancet*, 375, 1525–1535. [670]

Atkinson, A. C. (1982), “Optimum Biased Coin Designs for Sequential Clinical Trials With Prognostic Factors,” *Biometrika*, 69, 61–67. [679]

Begg, C. B., and Iglewicz, B. (1980), “A Treatment Allocation Procedure for Sequential Clinical Trials,” *Biometrics*, 36, 81–90. [679]

Birkett, N. J. (1985), “Adaptive Allocation in Randomized Controlled Trials,” *Controlled Clinical Trials*, 6, 146–155. [669]

Cheung, S. H., Zhang, L. X., Hu, F., and Chan, W. S. (2014), “Covariate-Adjusted Response-Adaptive Designs for Generalized Linear Models,” *Journal of Statistical Planning and Inference*, 149, 152–161. [671]

Efron, B. (1977), “Forcing a Sequential Experiment to be Balanced,” *Biometrika*, 58, 403–417. [670]

Feinstein, A. R., and Landis, J. R. (1976), “The Role of Prognostic Stratification in Preventing the Bias Permitted by Random Allocation of Treatment,” *Journal of Chronic Diseases*, 29, 277–284. [670,679]

Forsythe, A. B. (1987), “Validity and Power of Tests When Groups Have Been Balanced for Prognostic Factors,” *Computational Statistics Data Analysis*, 5, 193–200. [669,673]

Green, S. B., and Byar, D. P. (1978), “The Effect of Stratified Randomization on Size and Power of Statistical Tests in Clinical Trials,” *Journal of Chronic Diseases*, 31, 445–454. [670,679]
