Statistical Inference Following Covariate-Adaptive Randomization: Recent Advances

Wei Ma* Feifang Hu† Lixin Zhang‡

Abstract

Covariate-adaptive randomization (CAR) has been increasingly implemented in clinical trials to balance important covariates. However, the properties of statistical inference following CAR are not fully understood. In the literature, most studies are based on simulations. In this paper, we summarize some recent advances on theoretical properties of hypothesis testing under CAR, proposed by Ma *et al.* (2014). We will first give a general review of basic concepts of CAR and motivations to study inference properties following CAR. We next summary the main results in the paper, including describing the framework and assumptions and giving the theoretical properties. In the linear model framework, asymptotical distributions of test statistics are given for testing treatment effects and significance of covariates under null and alternative hypotheses. In particular, it is shown that under a large class of CAR 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. We close with a discussion of related work and possible future directions.

Key Words: Covariate-adaptive design, statistical inference, linear models, Pocock and Simon's marginal procedure, minimization, Type I error, power

1. Introduction

The purpose of this paper is to review properties of statistical inference for covariate-adaptive randomization (CAR) in the linear model framework, proposed by Ma *et al.* (2014). In clinical trials, two types of hypothesis testing are of particular interest, one is to detect significant treatment effects between different treatment groups, which is usually primary goals of clinal trial. The other one is to test whether a covariate is influential on patients' outmans, which becomes more and more important, especially in some biomarker studies. In this paper, the properties of these two types of hypothesis testing will be given for a large family of CAR, which provides theoretical foundation and guidance to apply CAR in practice.

We begin in Section 2 with a brief review of covariate-adaptive randomization used in clinical trials. In Section 3, we summarize the conclusions in Ma *et al.* (2014), which evaluate Type I error and power of statistical inference in the linear model framework for CAR. Section 4 provides discussion about the conclusions proposed, including assumptions in deriving asymptotic properties and how to extend the methodologies to study inference properties for generalized linear model under CAR. In Section 5 some conclusions remarks are given.

*Biogen Idec, Cambridge, MA 02142

†Department of Statistics, George Washington University, Washington. DC 20052

‡Department of Mathematics, Zhejiang University, Hangzhou, P.R. China 310006

2. Covariate-Adaptive Randomization

2.1 Covariate-Adaptive Design

In clinical trials, it is usually important to balance treatment arms with respect to key covariates. There are several advantages to apply covariate-adaptive randomization to clinical trials, such as improving treatment comparability and statistical efficiencies. A natural idea to achieve balance over covariates is stratification. Strata are defined as different combinations of covariates' levels. To get balanced trail, we could apply separate restricted randomization within each stratum to obtain good balance within each stratum and further to obtain overall balance. Depending what restricted randomization is used, we have stratified permuted block design using permuted block design within strata and covariate-adaptive biased coin design using Efron's biased coin design within strata. Stratified permuted block design is the most popular method to balance covariates and is used in most of clinical trials. However, it only works well for a trial with a few strata and large number of patients, otherwise a large portion of incomplete strata would cause imbalance on stratum level and further on the overall level.

To deal with many covariates, several marginal methods (also referred as minimization, dynamic allocation in literature) were proposed. Taves (1974) proposed a minimization method to deal with large number of covariates, but his method didn't involve randomness. Pocock and Simon (1975) generalized Taves' method by incorporating randomness, which has been more popular thereafter. Instead of attempting to eliminate imbalance within each stratum, their method achieves balance by reducing weighted sum of marginal imbalances. A simper version of Pocock and Simon's marginal procedure with two treatments can be described as follows. Suppose $N_{ijk}(n)$ is the number of patients on treatment k in level j of covariate Z_i, $i = 1, ..., I$, $j = 1, ..., m_i$, $k = 1, 2$, after n patients are enrolled in the study. Let $(z_1, ..., z_I)$ be the covariate information for the next patient. Then the marginal imbalance with respect to covariate i is defined as $D_i(n) = N_{iz_i1}(n) - N_{iz_i2}(n)$. The next assignment is based on the weighted average of all marginal imbalances $D(n) = \sum_{i=1}^{I} w_i D_i(n)$ by using the biased coin allocation,

$$\Pr(I_{n+1} = 1) = \begin{cases} p, & \text{if } D(n) < 0 \\ 1 - p, & \text{if } D(n) > 0 \\ 1/2, & \text{if } D(n) = 0 \end{cases}$$

where $1/2 < p < 1$ is the biased coin probability.

The usage of Pocock and Simon's marginal procedure had 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. However, despite its broad applications, the theoretical properties of Pocock and Simon's marginal procedure remain unknown ever since it was proposed. Recently, Hu and Zhang (2013) theoretically proved that, under Pocock and Simon' marginal procedure, the marginal and overall imbalances are bounded in probability, while the within-stratum imbalance increase with the rate of $\sqrt{n}$ as the sample size increases. These conclusions provide foundation for us to further study theoretical properties of statistical inference under these randomization methods.

Recently, Hu and Hu (2012) proposed a new family of covariate-adaptive procedures which simultaneously eliminate imbalances at three different levels, including overall imbalance, marginal imbalance and within-stratum imbalance. When a new patient enters the trial and is read for randomization, the assignment will be based on the weighted average of the above three imbalances, $D(n) = w_1 D_{strtum}(n) + w_2 D_{marginal}(n) + w_3 D_{overall}(n)$, where $D_{marginal}(n)$ is defined as in Pocock and Simon's marginal procedure, $D_{strtum}(n)$

and $D_{overall}(n)$ are within-stratum imbalance and overall imbalance, respectively. Theoretically, they proved that overall imbalance, marginal imbalance and within-stratum imbalance are all bounded in probability under the new covariate-adaptive designs.

2.2 Motivation to Study Inference Properties for CAR

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) theoretically 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, not all covariate information used in randomization can be fully utilized in inference procedures. In a clinical trial described in Anderson *et al.* (2000), Pocock and Simon's marginal procedure is implemented to balance allocation over three covariates including clinical centers, performance status and disease extent. A continuous primary endpoint is compared between two treatment groups using the two sample *t*-test, without adjusting covariate effects at all. In practice, some randomization covariates are omitted in final analysis due to: (i) it is difficult to incorporate some covariates in the analysis model, for example, investigation sites, etc.; (ii) adjusting too many covariates usually means more complicated modeling techniques; and (iii) it requires correct model specification, which is usually unknown in practice.

There have been doubts about validity of statistical inference for covariate-adaptive designs, especially when covariates are fully or partially omitted in inference procedures. Birkett (1985) and Forsythe (1987) had 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 treatments through simulation studies. They also found that the two sample *t*-test is less powerful for minimization than complete randomization for small treatment difference, but more powerful if larger treatment difference exists. In Shao, Yu and Zhong (2010) some theoretical work are done to study conservativeness of the two sample *t*-test. The following linear model with outcomes Y_{ij} for patient i under treatment j, $j = 0, 1$, is assumed for covariate-adaptive biased coin design,

$$Y_{ij} = \mu_j + bZ_i + \varepsilon_{ij}$$

where Z_i is a univariate covariate, Z_is are independent and identically distributed, μ_j and b are unknown parameters and ε_{ij}s are independent and identically distributed random errors and independent of Z_is. They theoretically proved that the two sample *t*-test is conservative by assuming responses follow the above simple homogeneous linear model. Moveover, a bootstrap test is proposed to adjust Type I error under covariate-adaptive biased coin design.

In the literature, the results of statistical inference for covariate-adaptive designs are restricted in several aspects. (1) Conclusions are mainly drawn by simulation, theoretical work is very limited. Shao, Yu and Zhong (2010) proved the property of two sample *t*-test based on covariate-adaptive biased coin design, which is a stratified design and less commonly used in practice. (2) Only two sample *t*-test is discussed, where no covariate information is incorporated in final analysis. In practice, it is often that a subset of randomization covariates are used in final statistical inference. The corresponding theoretical

properties remain unknown. (3) All studies focus on hypothesis testing for comparing treatment effects. There is very little, if any, discussion about inference of covariates under covariate-adaptive designs in the literature. In view of the importance of inference of covariates in clinical and medical studies, for example, in some personalized medicine and biomarker finding studies, we also want to study inference properties for covariates under covariate-adaptive clinical trials.

3. Statistical Inference for CAR

3.1 Framework

We consider the same setting and follow the notations in Ma *et al.* (2014). Suppose two treatments 1 and 2 are studied under CAR, μ_1 and μ_2 are 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 the ith patient, i.e., $I_i = 1$ for treatment 1 and $I_i = 0$ for treatment 2, $i = 1, 2, ..., N$. The following model is assumed for the response of the ith patient Y_i,

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

where

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

To study statistical inference for CAR, it is assumed only partial randomization covariates, X_k, $k = 1, ..., p$, are implemented into the analysis step. The following working model is used to do statistical inference,

$$\mathsf{E}[Y_i] = \mu_1 I_i + \mu_2(1 - I_i) + \beta_1 X_{i,1} + ... + \beta_p X_{i,p}. \qquad (2)$$

Under the working model (2), the ordinary least squares (OLS) method is used to obtain the estimator of $\boldsymbol{\beta}$, which has the explicit form,

$$\hat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

When model (2) is constructed to study clinical data collected from a CAR, the primary interest is usually to compare treatment effects between different groups. The following hypothesis testing is used to compare treatment effects of μ_1 and μ_2.

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

The test statistic for hypothesis testing (3) has the form.

$$T = \frac{\boldsymbol{L}\hat{\boldsymbol{\beta}}}{(\hat{\sigma}^2 \boldsymbol{L}(\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{L}^\top)^{1/2}}, \qquad (4)$$

where $\boldsymbol{L} = (1, -1, 0, ..., 0)$ and $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N-p-2)$. If $|T| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is $(1-\alpha/2)$ quantile of a standard normal distribution, we will reject the null hypothesis, otherwise accept the null hypothesis.

Besides testing treatment effects, we consider general forms of hypothesis testing for significance of covariates. Let $\boldsymbol{C}$ be an $m \times (p+2)$ matrix of rank m with $m < (p+2)$, where entries of the first two columns of $\boldsymbol{C}$ are all zeros so that $\boldsymbol{C\beta}$ doesn't include any treatment effects. Then the hypothesis testing of interest is,

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi}_0 \text{ versus } H_A : \boldsymbol{C\beta} = \boldsymbol{\xi_1} \neq \boldsymbol{\xi}_0. \tag{5}$$

The test statistic for hypothesis testing (5) is,

$$T^* = \frac{m^{-1}(\boldsymbol{C\beta} - \boldsymbol{\xi}_0)^\top [\boldsymbol{C}(\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{C}^\top]^{-1} (\boldsymbol{C\beta} - \boldsymbol{\xi}_0)}{\hat{\sigma}^2}. \tag{6}$$

If $T^* > \boldsymbol{\chi}^2_{m,(1-\alpha)}/m$, where $\boldsymbol{\chi}^2_{m,(1-\alpha)}$ is $(1-\alpha)$ quantile of a $\boldsymbol{\chi}^2$ distribution with degree of freedom m, we will reject the null hypothesis, otherwise accept the null hypothesis.

A special case of testing (5) is evaluating significance of a single covariate (biomarker). Without loss of generality, we consider the hypothesis testing for β_1, the coefficient of X_1. To test the significance of β_1, the hypothesis is

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0. \tag{7}$$

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

$$T_1 = \frac{\boldsymbol{\ell}\hat{\boldsymbol{\beta}}}{(\hat{\sigma}^2 \boldsymbol{\ell}(\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{\ell}^\top)^{1/2}}, \tag{8}$$

where $\boldsymbol{\ell} = (0, 0, 1, 0, ..., 0)$. If $|T_1| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is $(1-\alpha/2)$ quantile of a standard normal distribution, we will reject the null hypothesis, otherwise accept the null hypothesis.

In clinical trials, CAR 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 the continuous covariate into a discrete variable with several subcategories. Let $C = \{j \mid Z_j \text{ is continuous}, j = 1, ..., q\}$ and $C^* = \{k \mid X_k \text{ is continuous}, k = 1, ..., p\}$. 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. Suppose $\tilde{X}_{i,k}$ and $\tilde{Z}_{i,j}$ are ith observations of covariates $\tilde{X}_k$ and $\tilde{Z}_j$, $k = 1, ..., p$ and $j = 1, ..., q$. $\tilde{X}_{i,k}$ and $\tilde{Z}_{i,j}$ are used in the covariate-adaptive randomization process. We further define three 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 $W_i = (\tilde{X}_{i,1}, ..., \tilde{X}_{i,p}, \tilde{Z}_{i,1}, ..., \tilde{Z}_{i,q})$ represents the covariate profile of the ith patient used in randomization, i.e., $W_i = (x_1^{t_1}, x_2^{t_2}, ..., x_p^{t_p}, z_1^{r_1}, z_2^{r_2}, ..., z_q^{r_q})$ if $\tilde{X}_{i,k}$ is at level $x_k^{t_k}$ and $\tilde{Z}_{i,j}$ is at level $z_j^{r_j}$. For convenience, we use $(t_1, t_2, ..., t_p, r_1, r_2, ..., r_q)$ to denote the stratum formed by patients who have the same covariate profile $(x_1^{t_1}, x_2^{t_2}, ..., x_p^{t_p}, z_1^{r_1}, z_2^{r_2}, ..., z_q^{r_q})$, and use $(k; t_k)$ to denote the margin formed by patients with $\tilde{X}_k = x_k^{t_k}$, and similarly (j, r_j) to denote the margin formed by patients with $\tilde{Z}_j = z_j^{r_j}$. Then let

- D_N be the difference between the numbers of patients in treatment group 1 and 2 as total, i.e., 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;
- $D_N(t_1, t_2, ..., t_p, r_1, r_2, ..., r_q)$ be the difference between the numbers of patients in the two treatment groups within the stratum $(t_1, t_2, ..., t_p, r_1, r_2, ..., r_q)$.

3.2 Properties

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

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

(A) the overall imbalance is bounded, 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* $D_N(j; r_j) = O_p(1)$*,* $k = 1, 2, ..., p$*,* $j = 1, 2, ..., q$*.*

Then

(i) under $H_0 : \mu_1 - \mu_2 = 0$*,*

$$T \xrightarrow{D} \mathbf{N}(0, \tau^2),\ \tau^2 = \frac{\sigma_\varepsilon^2 + \sum\limits_{j \in C} \gamma_j^2 \sigma_{\delta,j}^2}{\sigma_\varepsilon^2 + \sum\limits_{j=1}^{q} \gamma_j^2 \mathsf{Var}(Z_j)} = \frac{\sigma_\delta^2}{\sigma_z^2}. \tag{9}$$

where $\sigma_z^2 = \sigma_\varepsilon^2 + \sum\limits_{j=1}^{q} \gamma_j^2 \mathsf{Var}(Z_j)$*,* $\sigma_\delta^2 = \sigma_\varepsilon^2 + \sum\limits_{j \in C} \gamma_j^2 \sigma_{\delta,j}^2$ *and* $\sigma_{\delta,j}^2 = \mathsf{E}[\mathsf{Var}(\delta_{i,j} | d_j(Z_{i,j}))]$*.*
Hence,

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

(2) If at least one $\gamma_j \neq 0$*,* $j = 1, 2, ..., q$*, 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, i.e.,* $\mu_1 - \mu_2 = \delta/\sqrt{N}$ *for a fixed* $\delta \neq 0$*, then*

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

Hence, the power increases as more covariates are incorporated into model.

In Theorem 3.1, two mild conditions (A) and (B) are assumed to derive the asymptotic distribution of the test statistic for comparing treatment effects. These conditions are satisfied by a variety of covariate-adaptive designs (see Corollary 3.2 for examples). Under these conditions, the numerator $\boldsymbol{L}\hat{\boldsymbol{\beta}} (= \hat{\mu}_1 - \hat{\mu}_2)$, has a smaller variance than the model-based variance estimator in the denominator if covariates are omitted from the working model. Based on the asymptotic distributions of the test statistic under both the null hypothesis and the alternative hypothesis, Type I error is smaller than the nominal level if at least one $\gamma_j \neq 0$, and power performance can be discussed as well.

Now consider the power of the hypothesis testing (2), under the alternative hypothesis, the power is

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

The power of test (2) under complete randomization (based on the same setting as described in Section 2.2) is

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

From the power expressions for both covariate-adaptive designs and complete randomization above,it can be concluded that the limiting power under covariate-adaptive designs is smaller than that under complete randomization when δ is relatively small, and it is larger than complete randomization when δ is large. This conclusion agrees with simulation studies about two sample *t*-test in literature (Forsythe, 1987; Shao, Yu and Zhong, 2010) for certain covariate-adaptive designs, and our conclusion is more general, which can be applied a general family of CAR.

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 affected if not all covariates are incorporated in the analysis model.

Theorem 3.2 *Under the same conditions as in Theorem 3.1,*

(i) under $H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi}_0$,

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

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

(ii) under $H_A : \boldsymbol{C\beta} = \boldsymbol{\xi}_1$, *consider a sequence of local alternatives, i.e.,* $(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_0) = \boldsymbol{\eta}/\sqrt{N}$ *for a fixed* $\boldsymbol{\eta} \neq \mathbf{0}$, *then*

$$T^* \xrightarrow{D} \boldsymbol{\chi}^2_{(m)}(\lambda)/m,\ \lambda = \boldsymbol{\eta}^\top [\boldsymbol{C}\boldsymbol{M}^{-1}\boldsymbol{C}^\top]^{-1}\boldsymbol{\eta}/\sigma_z^2, \tag{12}$$

where $\boldsymbol{M} = \mathtt{diag}(1/2, 1/2, \mathsf{Var}(X_1), ..., \mathsf{Var}(X_p))$ *and* λ *is the noncentral parameter. Therefore, the power increases as more covariates are incorporated into model.*

Theorem 3.1 and Theorem 3.2 imply that the overall difference and marginal imbalances play important roles in statistical inference for covariate-adaptive designs. For stratified permuted block design, the difference between the number of patients in two treatments within any stratum is the half of block size at maximum. Since the number of strata is finite for any covariate-adaptive design, the overall and marginal imbalance are less than a constant, thus the conditions (A) and (B) are satisfied. The theoretical properties for Pocock and Simon's marginal procedure remains unknown for decades and recently are derived by Hu and Zhang (2013). In their paper, the authors demonstrate the marginal imbalances and overall imbalance are bounded in probability for Pocock and Simon's marginal procedure, thus the conditions (A) and (B) are also satisfied. Furthermore, Hu and Hu (2012) proposed a large class of covariate-adaptive designs, which satisfy the conditions (A) and (B). Here we summarize these results in the following corollary.

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

(i) Pocock and Simon's marginal procedures (Pocock and Simon, 1975);

(ii) stratified permuted block designs; and

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

REMARK 3.1 *If we consider complete randomization as a special case of covariate-adaptive design, it does not satisfy the conditions in Theorem 3.1, because the marginal imbalance* $D_N(j; r_j) = O_p(N^{1/2})$ *for complete randomization. The numerical study in the next section shows that the test of treatment effect 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 covariate-adaptive designs. A linear regression model can be directly used to test significance of prognostic factors with a working model only containing partial covariate information. One the other hand, however, the power may be harmed by omitting important covariates in the working model. Consider the non-central parameter in (12), it increases with σ_z^2 reduced, so the power increases with fewer important covariates omitted from the model. Therefore, it is helpful of incorporating more important covariates, if possible, to obtain a more powerful test.

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

Corollary 3.2 *Under the same conditions as in Theorem 3.1,*

(i) under $H_0 : \beta_1 = 0$*,*

$$T_1 \xrightarrow{D} \mathbf{N}(0, 1). \tag{13}$$

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, i.e.,* $\beta_1 = \delta_{\beta_1}/\sqrt{N}$ *for a fixed* δ_{β_1}*, then*

$$T_1 \xrightarrow{D} \mathbf{N}(\Delta_{\beta_1}, 1),\ \Delta_{\beta_1} = \frac{\delta_{\beta_1}\sigma_1}{\sigma_z},$$

where $\sigma_1^2 = \mathsf{Var}(X_1)$*. Hence, the power increases as more covariates are incorporated into model.*

According to Theorem 3.1 and Theorem 3.2, a model with only influential covariates can achieve valid tests. It is known that too many unnecessary variables in the model will increase variations of estimators and affect statistical results. Hence, if only influential variables are incorporated in the model, it will not only reduce unnecessary variations, but also give valid inference. Detailed proofs of Theorem 3.1 and Theorem 3.2 and extensive simulation studies are in Ma *et al.* (2014).

4. Discussions

4.1 Assumption of Independent Covariates

In Section 3, the properties of statistical inference are studied for linear models under a large family of covariate-adaptive randomization. Among several assumptions to derive the theoretical results, it is assumed that all the covariates used in randomization are independent of each other. Based on this critical assumption of independence and other assumptions, it is proved that the hypothesis testing to compare treatment effects between two groups is conservative and the hypothesis testing for a linear combination of covariates remains valid. However, despite its importance in theory, the independence assumption is very strong and may be not satisfied in practice. Therefore, there are concerns whether the conclusions in the last section still hold if covariates are correlated. A simulation studies show that the conclusion holds under certain cases and more theoretical adjustment is desired.

A simulation case. The purpose is to study Type I error of hypothesis testing for comparing treatment effects when randomization covariates are correlated. Simulations are carried out for three randomization procedures: Pocock and Simon's marginal procedure, stratified permuted block design and complete randomization. The following linear model

Table 1: Simulated Type I error for Pocock and Simon's marginal procedure (PS), stratified permuted block design (SPB) and complete randomization (CR) in % with $\rho = 0.5$. Simulations based on 10000 runs.

Randomization	N	$t\text{-}test$	$lm(Z_1)$	$lm(Z_2)$	$lm(Z_1, Z_2)$
PS	100	1.07	3.34	3.40	4.90
	200	0.82	3.27	3.23	4.98
	500	0.82	3.12	3.27	4.88
SPB	100	1.10	3.42	3.63	5.15
	200	1.07	3.36	3.37	5.11
	500	0.86	3.38	3.40	5.13
CR	100	5.05	4.97	4.99	5.08
	200	5.27	5.10	5.44	5.36
	500	4.90	4.98	4.89	4.67

with two discrete covariates Z_1 and Z_2 is assumed to be the underlying true model with $\mu_1 = \mu_2$,

$$Y_i = \mu_1 I_i + \mu_2(1 - I_i) + Z_{i,1} + Z_{i,2} + \varepsilon_i,$$

where ε_i is distributed as $\mathbf{N}(0, 1)$. Z_1 and Z_2 follow Bernoulli(0.5) with the correlation between Z_1 and Z_2 equal to $\rho = 0.5$. The hypothesis tests are based on the two sample *t*-test ($t\text{-}test$), the linear model with a single covariate Z_1 ($lm(Z_1)$), the linear model with a single covariate Z_2 ($lm(Z_2)$), the linear model with both covariate Z_1 and Z_2 ($lm(Z_1, Z_2)$).

In simulations, the following setting of parameters is used. In Pocock and Simon's marginal procedures, equal weights are assigned to two covariates and the biased coin probability is equal to 0.75. The block size 4 is used for stratified permuted block design. The simulation results are presented in Table 1.

4.2 Inference for GLM

Most work in this paper is in the linear model framework where outcomes of clinical trials are continuous. When responses are binary, Feinstein and Landis (1976) and Green and Byar (1978) studied statistical problems comparing successful rate between two treatment groups on a special case where there are two strata and two treatment considered. Under this restricted assumption, they showed that Type I error is smaller than the nominal level under stratified randomization. Gail, Wieand and Piantadosi (1984) studied estimates of treatment effect in randomized experiments with nonlinear regression and omitted covariates. Gail (1988) studied properties of the score test for perfectly balanced studies across strata (not randomization) on a large family of generalized linear models. The following underlying model between response and treatment is assumed,

$$E(Y|T, X = i) = h(\alpha T + \beta_i),$$

where Y is the response variable, α is the treatment effect, and β_i is the stratum parameter for $X = i$. This covers a large range of generalized linear model. For example, $h(\eta) = \exp(\eta)/[1 + \exp(\eta)]$ in logistic regression. The properties of the score test is studied under perfectly balanced studies with no covariates included in analysis, so the working model is

$$E(Y|T, X = i) = h(\mu + \alpha T).$$

The properties of Type I error for several kinds of generalized linear models are given under the studies that are perfectly balanced. For example, Type I error is shown to be conservative for logistic regression. However, the properties of the score test and other tests, such as the Wald test and the likelihood ratio test, are unknown for general covariate-adaptive designs. Considering wide applications of clinical trials with response variables that are not continues, the inference properties of generalized linear models and even more advanced models under covariate-adaptive randomized clinical trials are desired. Those topics are left for future research.

5. Conclusion

In this paper, we reviewed theoretical properties of statistical inference under CAR based on linear models. In Section 2, the asymptotic distributions of several test statistics under both null and alternative hypotheses are given. Instead of focusing on a specific covariate-adaptive design, the problem is studied from the angle of imbalance measure of different levels (overall, marginal, within-stratum). So the conclusions can be applied to a broad range of covariate-adaptive designs, including stratified permuted block design and Pocock and Simon's marginal 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) to see if they are satisfied. Furthermore, in Section 3 we discussed topics about inference properties based on more general assumptions and statistical models. The results summarized in this paper provide new insights about balance and efficiency of clinical trials, and the framework can be used to study other statistical methods under covariate-adaptive designs.

REFERENCES

Ma, W., Hu, F., and Zhang, L. (2014), "Testing Hypotheses of Covariate-Adaptive Randomized Clinical Trials," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2014.922469.