Statistica Sinica **29** (2019), 265-282
doi:https://doi.org/10.5705/ss.202016.0330

SEQUENTIAL MONITORING OF COVARIATE-ADAPTIVE RANDOMIZED CLINICAL TRIALS

Hongjian Zhu and Feifang Hu

University of Texas Health Science Center at Houston
and George Washington University

Abstract: The sequential monitoring of covariate-adaptive randomized clinical trials is standard in modern clinical studies. However, the validity of this sequential procedure is not well studied in the literature. Clinical trialists therefore implement the procedure and perform data analysis based on the theory of the sequential monitoring of fixed designs, leaving many clinical trials open to question. In this paper, we study the theoretical properties of the sequential procedure and propose some important adjustments to classical statistical inference. Under different scenarios, we derive the asymptotic joint distribution of the sequential test statistics. Further, we estimate the decreased variability of the estimated treatment effect due to covariate-adaptive randomization, so that the sequential test statistics can be adjusted to be an asymptotic Brownian motion and the type I error rate can be controlled in real trials. Numerical results from simulation and the redesign of a clinical trial support our theoretical findings, showing that our procedure can control the type I error rate well, and also demonstrating the advantages of our method in terms of power and early stopping. Theoretical and numerical results provide important guidance for future practical clinical trials using covariate-adaptive randomization procedures.

Key words and phrases: Brownian motion, linear regression, personalized medicine, Pocock–Simon's randomization, stratified permuted block randomization, type I error rate.

1. Introduction

Clinical trials are usually complex, involving multiple covariates of interest in addition to the treatment effects. In particular, with the development of bioinformatics, the association between biomarkers and disease has become widely accepted. In the era of personalized medicine, it is desirable to incorporate covariates into clinical trial designs that investigate the heterogeneity of patients' responses to a treatment (Hu (2012); Hu et al. (2015)). The study results may be invalid if there is treatment imbalance over the covariates. Covariate-adaptive

randomization (CAR) procedures, that sequentially assign the next patient based on previous assignments and covariates, and the current covariate profile, have been developed to mitigate such imbalances and are extensively used in clinical trials. Stratified permuted block (SPB) randomization and Pocock and Simon's design (1975) are the most popular CAR procedures. Other CAR designs have been developed by Taves (1974), Wei (1978), Nordle and Brantmark (1977), Signorini et al. (1993), Heritier, Gebski and Pillai (2005), and Hu and Hu (2012). Clinical trials that use these designs include Iacono et al. (2006), Jakob et al. (2012), Anderson et al. (2000), Gridelli et al. (2003), Krueger et al. (2007), Molander et al. (2007), and Ohtori et al. (2012). A detailed discussion of CAR procedures can be found in Rosenberger and Sverdlov (2008). The theoretical properties of hypothesis testing based on CAR procedures have recently been developed by Shao, Yu and Zhong (2010) and Ma, Hu and Zhang (2015). However, these papers focused on the final test statistic, and not the sequential statistics.

While CAR procedures are popular in clinical trials, interim analysis is also common because of its ethical, administrative, and economic advantages (Jennison and Turnbull (2000)). Sequential monitoring arose from the sequential probability ratio test proposed by Wald (1947) for quality control, and its use in medical research was pioneered by Armitage (1975). Influential papers on sequential monitoring in clinical trial designs include Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983). Further, Jennison and Turnbull (1997) discussed a series of group sequential analysis methods incorporating covariate information through linear models, general parametric regression models, and survival models. They did not take into account the problems caused by covariate adaptive designs and the scenario where not all the design covariates are used in the analysis. Tsiatis, Rosner and Tritchler (1985) and Gu and Ying (1995) derived the joint distribution of sequential parameter estimators from proportional hazards models. More details of sequential monitoring can be found in Jennison and Turnbull (2000). These studies considered the scenarios where non-adaptive designs are implemented in clinical trials.

Despite the widespread popularity of the combination of CAR procedures with sequential monitoring in trials and the advantages mentioned, there have been few theoretical investigations of the sequential procedure. The CAR procedure has two limitations: the complicated correlation structure of the within-stratum imbalances and the discreteness of the allocation function. Furthermore, a special situation often arises in clinical trials: only some of the covariates used in the randomization procedures are included in the data analysis. For example,

Lai et al. (2006) investigated the influences of music on maternal anxiety in kangaroos in a randomized controlled trial. Under similar conditions, female infants are believed to have a significantly greater chance of surviving than male infants, hence permuted block randomization stratified on gender was used to allocate the patients. In the data analysis, a t-test was used to analyze the maternal-anxiety outcomes. The reasons for not using all the covariates include, but are not limited to, (i) it is not easy to explain the practical significance of including certain covariates such as investigation sites in the model; (ii) using too many covariates will lead to theoretical difficulties; (iii) the correct model specification is usually unknown. Consequently, theoretical investigation into the sequential monitoring of CAR procedures has been hindered. More importantly, the clinical trials that employ this procedure lack theoretical support, and many of these trials could be open to question.

In this paper, we study clinical trials with the CAR design for randomization and linear regression models for analysis. We obtain the joint distribution of the sequential statistics for three scenarios: (1) all the covariates used in the CAR are included in the data analysis; (2) some of the covariates are included; and (3) no covariates are included, which is Student's t-test. We find that for scenario (1) the joint distribution of the commonly used sequential statistics discussed in Section 2 is asymptotically Brownian motion, the asymptotic joint distribution for complete randomization and fixed designs. Clinical trial practitioners often perform data analysis following the sequential monitoring of CAR procedures, assuming that the data are from the sequential monitoring of complete randomization. This finding, for the first time to our knowledge, justifies and validates all such clinical trials for this scenario.

We also derive the joint distribution of the sequential statistics for scenarios (2) and (3), and one can see its difference from standard Brownian motion. As a result, trials that ignore the difference between CAR procedures and complete randomization could give misleading conclusions. These results provide guidance for practical clinical trials. In addition, the asymptotic variances of the sequential statistics for scenarios (2) and (3) indicate that the CAR design shrinks the variability of the estimated treatment effect. We propose an approach to estimate the decreased variance and adjust the sequential statistics, so that the critical values for Brownian motion can still be used, offering clinical trialists practical steps to deal with these complex situations.

We perform extensive numerical studies for these scenarios in terms of the type I error, power, and early stopping. We also redesign a double-blind random-

ized two-arm clinical trial conducted by Tilley et al. (1995) to study the properties of the proposed methods. The numerical results support our theoretical findings and demonstrate the advantages of our methods.

In Section 2, we introduce the notation, describe the framework, and formulate the main theorems. In Section 3, we use generated data to numerically study the sequential monitoring of CAR procedures. Numerical results from the redesign of a clinical trial are discussed in Section 4. Concluding remarks are in Section 5, and the proofs are given in the online supplementary material.

2. Sequential Monitoring of Covariate Adaptive Randomized Clinical Trials

2.1. Framework

We consider a two-arm randomized sequential experiment, in which n subjects are randomly assigned to one of the treatments by CAR procedures. Let T_i $(i = 1, \ldots, n)$ index the treatment (1 if treatment 1; 0 if treatment 2). To incorporate the scenario where some randomization covariates are omitted from the data analysis, we introduce two sets of covariates, $(X_1, \ldots, X_p)$ and $(Z_1, \ldots, Z_q)$. We use one dimensional covariates to describe our framework and theorems. It is easy to generalize our results to multiple dimensional covariates. Let $\boldsymbol{W}_i = (\boldsymbol{W}_i^X, \boldsymbol{W}_i^Z)$ be the covariate vector of the ith subject, where $\boldsymbol{W}_i^X = (X_{i1}, \ldots, X_{ip})$ and $\boldsymbol{W}_i^Z = (Z_{i1}, \ldots, Z_{iq})$. Here, $(X_1, \ldots, X_p)$ represent the covariates used for both randomization and analysis, and $(Z_1, \ldots, Z_q)$ represent those covariates that are used for randomization, but are not included for analysis. Assume the ith subject's response is

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + Z_{i1}\gamma_1 + \cdots + Z_{iq}\gamma_q + \epsilon_i, \quad (2.1)$$

where μ_1 and μ_2 are treatment effects for treatments 1 and 2, $(\beta_1, \ldots, \beta_p)$ and $(\gamma_1, \ldots, \gamma_q)$ are unknown parameters, and the ϵ_i are independent errors with mean 0 and variance σ^2. We assume that all the covariates are independent and, without loss of generality, $E(X_{ik}) = 0, E(Z_{ij}) = 0, i = 1, \ldots, n, k = 1, \ldots, p, j = 1, \ldots, q$. We also assume that the errors are independent of the covariates. We write $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, $\boldsymbol{\eta} = (\mu_1, \mu_2, \beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$, $\boldsymbol{T}(n) = (T_1, \ldots, T_n)^T$, $\boldsymbol{Y}(n) = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\epsilon}(n) = (\epsilon_1, \ldots, \epsilon_n)^T$, and

$$\boldsymbol{X}(n) = \begin{bmatrix} T_1 & 1 - T_1 & X_{11} & \ldots & X_{1p} \\ T_2 & 1 - T_2 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_n & 1 - T_n & X_{n1} & \ldots & X_{np} \end{bmatrix}.$$

When studying CAR, we discretize all the continuous covariates, and apply CAR designs with respect to these discrete covariate variables. Specifically, let

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

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

where $C = \{l : \text{index of continuous covariates among } X_l, l = 1, \ldots, p\}$, $C^* = \{l : \text{index of continuous covariates among } Z_l, l = 1, \ldots, q\}$, and $d_j(\cdot)$ and $d_j^*(\cdot)$ are discrete functions. Write $\tilde{\boldsymbol{W}}_i^X = (\tilde{X}_{i1}, \ldots, \tilde{X}_{ip})$ and $\tilde{\boldsymbol{W}}_i^Z = (\tilde{Z}_{i1}, \ldots, \tilde{Z}_{iq})$.

We need some notation to formulate the main theorem. Suppose $\tilde{X}_k$ has s_k levels and $\tilde{Z}_j$ has s_j^* levels, and let $\boldsymbol{W}_i = (x_{i1}^{c_1}, \ldots, x_{ip}^{c_p}, z_{i1}^{c_1^*}, \ldots, z_{iq}^{c_q^*})$ represent the ith subject's covariate profile if $\tilde{X}_{ik}$ is at level $x_{ik}^{c_k}$ and $\tilde{Z}_{ij}$ is at level $z_{ij}^{c_j^*}$. Let DIF_n be the overall difference in patient numbers between two treatments at the end of the trial. Let $DIF_n^X(k; c_k)$ be the marginal difference with respect to the level $x_k^{c_k}$ of covariate $\tilde{X}_k$, and $DIF_n^Z(j; c_j^*)$ be the marginal difference with respect to the level $z_j^{c_j^*}$ of covariate $\tilde{Z}_j$. Let $DIF_n(c_1, \ldots, c_p, c_1^*, \ldots, c_q^*)$ be the difference in patient numbers in the stratum containing the subjects with covariates $(x_{i1}^{c_1}, \ldots, x_{ip}^{c_p}, z_{i1}^{c_1^*}, \ldots, x_{iq}^{c_q^*})$.

Let $\lfloor nt \rfloor$ denote the largest integer not greater than nt for $t \in [0, 1]$. We introduce t, the "information time", to formulate this problem using the Skorokhod topology. Let $\mathcal{T}(\lfloor nt \rfloor) = \sigma(T_1, \ldots, T_{\lfloor nt \rfloor})$ be the sigma-algebra generated by the first $\lfloor nt \rfloor$ treatment assignments, and $\mathcal{X}(\lfloor nt \rfloor) = \sigma(\tilde{\boldsymbol{W}}_1^X, \ldots, \tilde{\boldsymbol{W}}_{\lfloor nt \rfloor}^X)$ and $\mathcal{Z}(\lfloor nt \rfloor) = \sigma(\tilde{\boldsymbol{W}}_1^Z, \ldots, \tilde{\boldsymbol{W}}_{\lfloor nt \rfloor}^Z)$ be the sigma-algebras generated by the first $\lfloor nt \rfloor$ covariate vectors $\tilde{X}$ and $\tilde{Z}$. Then, after $N = \lfloor nt \rfloor$ patients have been assigned, the adaptive randomization selects the next treatment assignment based on $\mathcal{F}(N) = \mathcal{T}(N) \bigotimes \mathcal{X}(N+1) \bigotimes \mathcal{Z}(N+1)$.

To compare the two treatment effects, we consider the hypothesis test:

$$H_0 : \mu_1 = \mu_2 \text{ versus } \mu_1 \neq \mu_2. \tag{2.2}$$

A natural statistic including only X to test the above hypothesis at time point $t \in (0, 1]$ is

$$Z_t = \frac{L\hat{\boldsymbol{\eta}}(t)}{\sqrt{\hat{\sigma}(t)^2 L\{\boldsymbol{X}(\lfloor nt \rfloor)^T \boldsymbol{X}(\lfloor nt \rfloor)\}^{-1} L^T}}, \tag{2.3}$$

where $L = (1, -1, 0, \ldots, 0)$, $\hat{\boldsymbol{\eta}}(t) = \{\boldsymbol{X}(\lfloor nt \rfloor)^T \boldsymbol{X}(\lfloor nt \rfloor)\}^{-1} \boldsymbol{X}(\lfloor nt \rfloor)^T \boldsymbol{Y}(\lfloor nt \rfloor)$, and $\hat{\sigma}(t)^2 = \{\boldsymbol{Y}(\lfloor nt \rfloor) - \boldsymbol{X}(\lfloor nt \rfloor)\hat{\boldsymbol{\eta}}(t)\}^T \{\boldsymbol{Y}(\lfloor nt \rfloor) - \boldsymbol{X}(\lfloor nt \rfloor)\hat{\boldsymbol{\eta}}(t)\}/(\lfloor nt \rfloor - p - 2)$.

The sequential statistics (2.3) are the commonly used statistics.

2.2. Asymptotic results

Controlling the type I error rate is the primary challenge when sequentially monitoring a clinical trial. The key is the asymptotic joint distribution of the sequential statistics and the subsequent choices of critical values. Numerous techniques have been proposed for sequentially monitoring the Brownian motion that follows a complete randomization. However, CAR procedures lead to considerable difficulties in deriving the joint distributions of the sequential test statistics: the sequential treatment assignments are not independent of the covariate profiles; the observed responses are not independent of previous treatment assignments and covariates; the observed responses are not independent of each other.

Let

$$Z_t^{adj} = \frac{L\hat{\boldsymbol{\eta}}(t)}{\hat{\epsilon}(t)\sqrt{\hat{\sigma}(t)^2 L\{\boldsymbol{X}(\lfloor nt \rfloor)^T \boldsymbol{X}(\lfloor nt \rfloor)\}^{-1} L^T}}, \tag{2.4}$$

where $\hat{\epsilon}(t)^2$ is any consistent estimator of

$$\frac{\sum_{j\in C^*} \gamma_j^2 \sigma_{\delta j}^2 + \sigma^2}{\sigma^2 + \sum_{j=1}^p Var(Z_j \gamma_j^T)} \tag{2.5}$$

in (S1.9) in the online supplementary material, $\sigma_{\delta j}^2 = E[Var(\delta_j | d_j^*(Z_j))]$, and $\delta_j = Z_j - E\{Z_j | d_j^*(Z_j)\}$. We will discuss $\hat{\epsilon}$ later.

Theorem 1. *Let $B_t^{adj} = \sqrt{t} Z_t^{adj}$ in the space $D[0,1]$ with the Skorohod topology. If a covariate adaptive design satisfies $DIF_n = O_p(1)$, $DIF_n^X(k; c_k) = O_p(1), k = 1, \ldots, p$, and $DIF_n^Z(j; c_j^*) = O_p(1), j = 1, \ldots, q$, then under H_0, B_t^{adj} is asymptotically a standard Brownian motion in distribution. The sequence of test statistics $\{Z_{t_1}^{adj}, \ldots, Z_{t_K}^{adj}, 0 \le t_1 \le t_2 \le \cdots \le t_K \le 1\}$ has the asymptotic canonical joint distribution of* Jennison and Turnbull (2000)*:*
(i) $\{Z_{t_1}^{adj}, \ldots, Z_{t_K}^{adj}\}$ is multivariate normal;
(ii) $EZ_{t_i}^{adj} = 0$;
(iii) $Cov(Z_{t_i}^{adj}, Z_{t_j}^{adj}) = \sqrt{t_i/t_j},\ 0 \le t_i \le t_j \le 1$.

Under H_1,

$$B_t^{adj} - \frac{\sqrt{n}(\mu_1 - \mu_2)t}{2\sqrt{\sum_{j\in C^*} \gamma_j^2 \sigma_{\delta j}^2 + \sigma^2}}$$

converges to a standard Brownian motion.

Thus, the effect of CAR procedures on the joint distribution of the sequential

statistics is asymptotically the same as that of complete randomization after adjustment. Clinical trialists implement this procedure assuming that it is the same as complete randomization. Here, we can see the gap and even calculate the difference given the parameter values. To the best of our knowledge, we provide the first theoretical foundation for this procedure.

Remark 1. (1) The conditions on the overall and marginal differences in patient numbers between two treatments in the theorem hold for a variety of CAR procedures such as the stratified permuted block randomization.
(2) The asymptotic variance (2.5) of Z_t is always less than 1, so the variability of the estimated treatment effect has been reduced by the CAR designs.
(3) Because of the reduced variability of the estimated treatment effect, using the traditional estimator of this variance in the statistics leads to a conservative type I error rate. Without adjustment, the power is adversely affected, which effectively increases the necessary sample size and is not consistent with the original aim of sequential monitoring.

2.3. Data analysis with a full dataset and student's t-test statistic

Here, we discuss data analysis with all the covariates used in the randomization, and Student's t-test without any covariates. Let the ith subject's response Y_i be

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \epsilon_i, \tag{2.6}$$

with notation as in model (2.1). We implement the CAR and perform data analysis with all the covariates in model (2.6). To compare the two treatment effects and to perform hypothesis test (2.2), we use the test statistic (2.3) at time point t.

Theorem 2. *Let $B_t = \sqrt{t}Z_t$ in the space $D[0,1]$ with the Skorohod topology. If a covariate adaptive design satisfies $DIF_n = O_p(1)$ and $DIF_n^X(k; c_k) = O_p(1), k = 1, \ldots, p$, then under H_0, B_t is asymptotically a standard Brownian motion in distribution. The sequence of test statistics $\{Z_{t_1}, \ldots, Z_{t_K}, 0 \le t_1 \le t_2 \le \cdots \le t_K \le 1\}$ has the asymptotic canonical joint distribution of* Jennison and Turnbull (2000). *Under H_1, $B_t^{adj} - \{\sqrt{n}(\mu_1 - \mu_2)t\} / (2\sigma)$ converges to a standard Brownian motion.*

In this scenario, we do not have to adjust the sequential statistic (2.3).

Another scenario is that the CAR is used to sequentially allocate patients and the data is analyzed with the following model,

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + \epsilon_i, i = 1, \ldots, n, \tag{2.7}$$

which is equivalent to the Student's t-test. We assume that the responses are

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + Z_{i1}\gamma_1 + \cdots + Z_{iq}\gamma_q + \epsilon_i, i = 1, \ldots, n. \tag{2.8}$$

Let $E = (1, -1)$ and

$$\boldsymbol{Tr}(n) = \begin{bmatrix} T_1 & 1 - T_1 \\ T_2 & 1 - T_2 \\ \vdots & \vdots \\ T_n & 1 - T_n \end{bmatrix}.$$

Via a similar argument to that in Section 2.2, the statistic for testing the hypothesis (2.2) at time point $t \in (0, 1]$ is

$$Z_t^{adj2} = \frac{E\hat{\boldsymbol{\mu}}(t)}{\hat{\epsilon}(t)\sqrt{\hat{\sigma}(t)^2 E(\boldsymbol{Tr}(\lfloor nt \rfloor)^T \boldsymbol{Tr}(\lfloor nt \rfloor))^{-1} E^T}}, \tag{2.9}$$

where $\hat{\boldsymbol{\mu}}(t) = \{\boldsymbol{Tr}(\lfloor nt \rfloor)^T \boldsymbol{Tr}(\lfloor nt \rfloor)\}^{-1} \boldsymbol{Tr}(\lfloor nt \rfloor)^T \boldsymbol{Y}(\lfloor nt \rfloor)$, $\hat{\sigma}(t)^2 = \{\boldsymbol{Y}(\lfloor nt \rfloor) - \boldsymbol{Tr}(\lfloor nt \rfloor)\hat{\boldsymbol{\mu}}(t)\}^T \{\boldsymbol{Y}(\lfloor nt \rfloor) - \boldsymbol{Tr}(\lfloor nt \rfloor)\hat{\boldsymbol{\mu}}(t)\}/(\lfloor nt \rfloor - 2)$, and $\hat{\epsilon}(t)^2$ is a consistent estimator of

$$\frac{\sum_{j \in C^*} \gamma_j^2 \sigma_{\delta j}^2 + \sigma^2}{\sigma^2 + \sum_{j=1}^p Var(Z_j \gamma_j^T)}.$$

Theorem 3. *Let $B_t^{adj2} = \sqrt{t} Z_t^{adj2}$ in the space $D[0, 1]$ with the Skorohod topology. If a covariate adaptive design satisfies $DIF_n = O_p(1)$ and $DIF_n^Z(j; c_j^*) = O_p(1), j = 1, \ldots, q$, B_t^{adj2} and Z_t^{adj2} have the same properties as B_t^{adj} and Z_t^{adj} in Theorem* 1*, respectively.*

Stratified permuted block randomization and Student's t-test are the most popular combination in clinical trials, and this result offers a way to control the type I error rate when sequentially monitoring this procedure.

2.4. Choice of $\hat{\epsilon}$ and critical values to control the type I error rate

We discuss how to obtain the consistent estimator ($\hat{\epsilon}(t)$) based on the data collected by information time t. In some cases it is preferable to perform data analysis with sequential statistics using partial covariates, but it is reasonable to make adjustments to the critical values or, equivalently, to the test statistics, with all the data available. Different approaches such as bootstraps to obtain $\hat{\epsilon}$ might be available depending on the specific models, and have diverse desirable features. We propose a simple approach based on linear models. For each interim look, we fit model (2.1) with full data to obtain consistent estimators of $\boldsymbol{\gamma}$ and σ.

By the Law of Large Numbers, we can also easily obtain consistent estimators of $\sigma_{\delta j}$ and $Var(Z_j)$ based on the observed covariates, and the consistency of $\hat{\epsilon}$ follows fundamental large-sample theory (Lehmann (2004)).

Although CAR procedures sequentially update information and the allocation probability, the joint distribution of the adjusted sequential test statistics is still a Brownian motion or the canonical joint distribution of Jennison and Turnbull (2000). Thus existing techniques could be used when sequentially monitoring a CAR. These techniques include, but are not limited to, Pocock's test, O'Brien and Fleming's test, the tests of Wang and Tsiatis (1987), the tests of Haybittle (1971) and Peto et al. (1976), the equivalence test, spending functions, stochastic curtailment, and repeated confidence intervals.

We focus on choosing appropriate critical values to control the type I error rate, and we exemplify this procedure by using spending functions. In particular, for the numerical studies in the next section, we assume that sequential hypothesis tests are performed at time points $t_1 = 0.2$, $t_2 = 0.5$, and $t_3 = 1$. We also assume that the three sets of boundaries from Proschan, Lan and Wittes (2006) can be used to control the nominal type I error rate of 0.05: O'Brien–Fleming-like boundaries $(4.877, 2.963, 1.969)$, linear boundaries $(2.576, 2.377, 2.141)$, and Pocock-like boundaries $(2.438, 2.333, 2.225)$. More details can be found in Proschan, Lan and Wittes (2006). In the numerical studies, we give results only for the O'Brien–Fleming boundary; it is the most popular one in clinical trials and the other boundaries give similar conclusions.

3. Numerical Studies

In this section, we report on the finite-sample properties of the procedure and demonstrate our theoretical findings via numerical results. In Tables 1–3 we present our theoretical findings. In Tables 4 and 5 we numerically study the robustness of our method under two scenarios of model mis-specification. In Table 6 we study the performance of our method when sparse samples occur at some levels of covariates that are used for the CAR design.

For Tables 1–3, suppose 500 patients sequentially enter a clinical trial, and the responses are

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + Z_{i1}\gamma_1 + Z_{i2}\gamma_2 + \epsilon_i, i = 1, \ldots, 500, \tag{3.1}$$

where $(\mu_1, \mu_2, \gamma_1, \gamma_2)$ are unknown parameters, and the ϵ_i are independent errors from the normal distribution $N(0, 1)$. We study complete randomization, the Pocock–Simon procedure (PS), and the stratified permuted block randomization

(SPB). The covariate adaptive designs are based on Z_1 and Z_2. We give numerical results for a data analysis with the full dataset and model (3.1) ("Full" in the tables) and a partial dataset and the following model including only Z_1 ("Partial" in the tables):

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + Z_{i1}\gamma_1 + \epsilon_i, i = 1, \ldots, 500. \tag{3.2}$$

We also give results for Student's t-test without any covariates ("t-test" in the tables). We do not distinguish X and Z here for space efficiency. For each CAR, we give results for both the adjusted and unadjusted sequential statistics; PS, PS-adj, SPB, and SPB-adj represent the four cases. In Tables 1–3, we report results where Z_1 and Z_2 are binary covariates with a success rate of 0.5 ("discrete" in the tables) and where Z_1 and Z_2 follow the normal distribution $N(0, 1)$ ("continuous" in the tables). We tried other settings for the covariates and similar results were obtained. When the CAR procedures are implemented with continuous covariates, we discretized them as

$$\tilde{z} = \begin{cases} 1 & \text{if } z < z_{0.4} \\ 0 & \text{if } z \geq z_{0.4} \end{cases},$$

where $z_{0.4}$ is the 0.4-quantile of the standard normal distribution. All results are based on 10,000 replications.

In Table 1, we give the type I error rate assuming that the responses follow model (3.1) with $(\mu_1, \mu_2, \gamma_1, \gamma_2) = (0.5, 0.5, 1, 1)$. We found that when all the covariates were used in the data analysis, the sequential monitoring of all three randomization procedures without adjustment controlled the type I error rate well, consistent with Theorem 2. We do not have to adjust the sequential statistics in this case. The sequential monitoring of complete randomization in all the cases in this section has no problem in controlling the type I error rate. We also found that the sequential monitoring of CAR procedures with the proposed adjusted sequential statistics can protect the type I error rate when not all the covariates are included in the data analysis, whereas the rate is conservative without adjustments. Data analysis with Student's t-test is more conservative than that based on partial covariates. Our theorems allow an explicit calculation of the gap between the unadjusted rate and the adjusted rate for different scenarios, and our numerical results are consistent with the theoretically derived discrepancy.

In Table 2, we give the power assuming that the responses follow model (3.1) with $(\mu_1, \mu_2, \gamma_1, \gamma_2) = (0.5, 0.75, 1, 1)$, and the other settings are the same

Table 1. Type I error rate for different scenarios.

	Full		Partial		t-test	
	discrete	continuous	discrete	continuous	discrete	continuous
CR	0.053	0.051	0.052	0.054	0.055	0.052
PS	0.051	0.052	0.031	0.018	0.017	0.011
SPB	0.051	0.049	0.028	0.019	0.019	0.010
PS-adj	NA	NA	0.050	0.051	0.053	0.048
SPB-adj	NA	NA	0.048	0.050	0.051	0.049

Table 2. Power for different scenarios.

	Full		Partial		t-test	
	discrete	continuous	discrete	continuous	discrete	continuous
CR	0.795	0.796	0.715	0.507	0.634	0.366
PS	0.802	0.795	0.727	0.500	0.651	0.320
SPB	0.800	0.799	0.725	0.501	0.652	0.318
PS-adj	NA	NA	0.800	0.665	0.801	0.566
SPB-adj	NA	NA	0.800	0.663	0.801	0.565

Table 3. Early stopping for different scenarios.

	Full		Partial		t-test	
	discrete	continuous	discrete	continuous	discrete	continuous
CR	1,595	1,680	1,262	630	951	382
PS	1,621	1,643	938	293	516	90
SPB	1,599	1,682	892	314	495	104
PS-adj	NA	NA	1,694	1,083	1,710	771
SPB-adj	NA	NA	1,680	1,062	1,689	741

as before. The value of μ_2 was chosen so that the power is around 0.8 for the sequential monitoring of complete randomization when the "full" model is used. We found that the sequential monitoring of CAR procedures produces similar results to those for the sequential monitoring of complete randomization in terms of power and early stopping when both covariates are included in the data analysis. When only one covariate is included in the data analysis, the sequential monitoring of CAR with adjusted sequential statistics can increase the power. In Table 3, we study early stopping under the scenarios in Table 2. We report the total number of stops at the first two looks, which means early stopping, among 10,000 replications. The sequential monitoring of CAR designs with adjusted sequential statistics stops the trials much earlier than the other approaches.

We discuss the performance of the proposed method when the model is mis-

Table 4. Type I error rate (α), power, and early stopping when two covariates are correlated.

	Full			Partial			t-test		
	α	Power	Early stopping	α	Power	Early stopping	α	Power	Early stopping
CR	0.046	0.799	1,648	0.046	0.726	1,299	0.051	0.570	761
PS	0.052	0.802	1,696	0.033	0.742	1,111	0.011	0.599	358
SPB	0.048	0.791	1,619	0.030	0.738	1,057	0.011	0.592	303
PS-adj	NA	NA	NA	0.058	0.810	1,834	0.052	0.801	1,736
SPB-adj	NA	NA	NA	0.051	0.799	1,726	0.048	0.792	1,632

specified. In Table 4, we consider the case where Z_1 is Bernoulli with a success rate of 0.5 and Z_2 is correlated with Z_1 as

$$P(Z_2 = 1|Z_1 = 1) = 0.8 \text{ and } P(Z_2 = 1|Z_1 = 0) = 0.4.$$

The other settings were the same as before. We report the type I error rate when $(\mu_1, \mu_2) = (0.5, 0.5)$, and (in the same table for space efficiency) the power and early stopping results when $(\mu_1, \mu_2) = (0.5, 0.75)$. We see that our adjusted sequential statistics work well when the two covariates are correlated, and adjustment is not needed when both covariates are included in the data analysis. Our method can greatly increase the power and stop the trial significantly earlier. Without adjustment, using fewer covariates leads to a lower power, and adjustment can help us to obtain similar powers for different scenarios.

In Table 5, we consider model mis-specification when there are unobserved covariates that influence the responses. We took responses as

$$Y_i = \mu_1 T_i + \mu_2(1 - T_i) + Z_{i1}\gamma_1 + Z_{i2}\gamma_2 + Z_{i3}\gamma_3 + \epsilon_i, i = 1, \ldots, 500, \qquad (3.3)$$

where $\gamma_3 = 1$ and Z_3 was Bernoulli distribution with a success rate of 0.6. Other settings are the same as Tables 1-4. Since Z_3 is assumed to be unobservable, the SPB randomization design and the Pocock and Simon's design were implemented with respect to only Z_1 and Z_2, "Full" in Table 5 means that both Z_1 and Z_2 were included in the data analysis, and "Partial" means that only Z_1 was included in the data analysis. Our proposed method is robust under this scenario in terms of the type I error rate; it increases the power, and it stops the trial much earlier compared to using the unadjusted statistics.

In Table 6, we investigate the performance of our method when sparse samples occur at some levels of covariates that are used for the CAR design. We consider the case where Z_1 was Bernoulli with success rate of 0.5, but now Z_2 was Bernoulli with success rate of 0.9. Other settings are the same as Table

Table 5. Type I error rate (α), power, and early stopping when there is one unknown covariate.

	Full			Partial			t-test		
	α	Power	Early stopping	α	Power	Early stopping	α	Power	Early stopping
CR	0.050	0.702	1,194	0.052	0.625	914	0.051	0.561	729
PS	0.048	0.706	1,182	0.031	0.638	704	0.021	0.568	404
SPB	0.053	0.712	1,230	0.033	0.642	746	0.021	0.572	452
PS-adj	NA	NA	NA	0.050	0.709	1,174	0.051	0.708	1,209
SPB-adj	NA	NA	NA	0.053	0.714	1,231	0.054	0.712	1,240

Table 6. Type I error rate (α), power, and early stopping when sparse samples occur at certain covariate levels.

	Full			Partial			t-test		
	α	Power	Early stopping	α	Power	Early stopping	α	Power	Early stopping
CR	0.053	0.796	1,588	0.053	0.755	1,365	0.053	0.670	1,018
PS	0.049	0.795	1,650	0.040	0.767	1,337	0.023	0.693	776
SPB	0.051	0.792	1,591	0.042	0.765	1,324	0.024	0.693	739
PS-adj	NA	NA	NA	0.050	0.793	1,646	0.052	0.793	1,670
SPB-adj	NA	NA	NA	0.052	0.792	1,629	0.051	0.791	1,641

4. The advantages of our methods displayed in previous tables remain under this scenario. The proposed method is robust when there are sparse samples at certain covariate levels.

4. Redesign of Clinical Trial Evaluating Treatment for Rheumatoid Arthritis

Rheumatoid arthritis is a chronic inflammatory disorder typically affecting the small joints and causing painful swelling. It eventually results in bone erosion and joint deformity. Tilley et al. (1995) conducted a clinical trial to assess the safety and efficacy of minocycline in the treatment of rheumatoid arthritis. This was a double-blind randomized trial of oral minocycline or a placebo. A total of 219 patients entered the trial; 109 were assigned to the treatment group and 110 to the placebo group.

Here we redesign the clinical trial and focus on the measurement of the change in hematocrit. Low hematocrit is common in patients with rheumatoid arthritis. After removing some missing data, we obtained summary statistics and parameter estimators in a linear model using information for 205 patients (108

Table 7. Evaluation of power and early stopping for stratified permuted block randomization in real-data analysis.

		SPB		SPB-adj	
	Sample size	Power	Early stopping	Power	Early stopping
Full		0.94	8,178	NA	NA
Partial	205	0.935	8,081	0.942	8,196
t-test		0.928	7,911	0.942	8,202
Full		0.681	4,770	NA	NA
Partial	100	0.668	4,621	0.686	4,808
t-test		0.650	4,399	0.689	4,880

treatment, 107 control). Two binary covariates were used in the model: Z_1 was the indicator of "oral corticosteroids used at entry" with a success rate of 0.32, and Z_2 was education status with a success rate of 0.46 ($Z_2 = 0$ for high school graduation or below, $Z_2 = 1$ for at least some college). The fitted model was

$$\hat{y}_i = -1.66 + 1.67T_i + 1.69Z_1 + 1.21Z_2, \tag{4.1}$$

with residual $N(0, 3.39^2)$.

We generated covariate data based on these summary statistics, sequentially allocated the patients using CAR, and generated responses based on the fitted model (4.1). To provide more information, we used different time points than in the previous sections to perform the sequential monitoring; $t_1 = 0.5$, $t_2 = 0.8$, and $t_3 = 1$. The corresponding boundaries to keep the overall type I error at 0.05 are O'Brien–Fleming-like boundaries $(2.963, 2.266, 2.028)$, linear boundaries $(2.241, 2.252, 2.247)$, and Pocock-like boundaries $(2.157, 2.288, 2.347)$. We report results (see Table 7) only for stratified permuted block randomization and O'Brien–Fleming-like boundaries, since this is the most popular combination and other settings give similar results. The results are consistent with the previous numerical studies. CAR procedures work well if all the covariates used for the randomization are included in the model. Otherwise, our adjustments are needed to improve the power. In addition, our method with adjusted sequential statistics can stop the trial earlier, based on the number of stops at the first two looks. Note that this data has a relatively large variance of error. It is dominant in the asymptotic variance of the sequential statistics, and the effect of covariate adaptive design is not quite significant. Even in this special situation, we can see that our method shows improvement. We also provide results for a sample size of 100 to show the small-sample performance of our method. Our method greatly improves the performance in this case.

5. Discussion

There are important directions for future research. First, we have studied data analysis for continuous responses with linear regression, so binary responses with logistic regression are a natural generalization. Other types of responses and models deserve study and difficulties could be introduced by the nonexistence of a closed form of the parameter estimators. We have made use of the α-spending function to control the type I error rate. Other methods may provide diverse advantages; these include optimal spending functions (Anderson (2007)) and beta spending functions. A generalized structure of covariates could be investigated for other scenarios in clinical trials. Other approaches to adjust the sequential statistics could be developed. Finally, Hu and Rosenberger (2006) classified adaptive randomization procedures into four categories: restricted randomization, response-adaptive randomization (RAR), CAR, and covariate-adjusted response-adaptive (CARA) randomization. Zhu and Hu (2010, 2012) studied sequential monitoring of RAR in clinical trials. The sequential monitoring of CARA deserves investigation.

Supplementary Materials

The proofs are in the online supplementary materials.

Acknowledgment

We appreciate the constructive suggestions from the referees and the editors. Research is supported by grant DMS-1612970 from the National Science Foundation (USA) and by grant No. 11371366 from the National Natural Science Foundation of China.

References

Anderson, K. M. (2007). Optimal spending functions for asymmetric group sequential designs. *Biometrical Journal* **49**, 337–345.

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., McIllmurray, M. B. and Carmichael, J. (2000). Gemcitabine 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.

Armitage, P. (1975). *Sequential Medical Trials*. Blackwell, Oxford.

Gridelli, C., Gallo, C., Shepherd, F. A., Illiano, A., Piantedosi, F., Robbiati, S. F., Manzione, L., Barbera, S., Frontini, L., Veltri, E., Findlay, B., Cigolari, S., Myers, R., Ianniello, G.

P., Gebbia, V., Gasparini, G., Fava, S., Hirsh, V., Bezjak, A., Seymour, L. and Perrone, F. (2003). Gemcitabine plus vinorelbine compared with cisplatin plus vinorelbine or cisplatin plus gemcitabine for advanced non-small-cell lung cancer: A phase III trial of the Italian GEMVIN Investigators and the National Cancer Institute of Canada Clinical Trials Group. *Journal of Clinical Oncology* **21**, 3025–3034.

Gu, M. and Ying, Z. (1995). Group sequential methods for survival data using partial likelihood score processes with covariate adjustment. *Statistica Sinica.* **5**, 793–804.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* **44**, 793–797.

Heritier, S., Gebski, V. and Pillai, A. (2005). Dynamic balancing randomization in controlled clinical trials. *Statistics in Medicine* **24**, 3729–3741.

Hu, F. (2012). Statistical issues in trial design and personalized medicine. *Clinical Investigation* **2**, 121–124.

Hu, F., Hu, Y., Ma, W., Zhang L. X. and Zhu, H. (2015). Statistical inference of adaptive randomized clinical trials for personalized medicine. *Clinical Investigation* **5**, 415–425.

Hu, F. and Rosenberger, W. F. (2006). The Theory of Response-Adaptive Randomization in Clinical Trials (525). John Wiley & Sons.

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

Iacono, A. T., Johnson, B. A., Grgurich, W. F., Youssef, J. G., Corcoran, T. E., Seiler, D. A., Dauber, J. H., Smaldone, G. C., Zeevi, A., Yousem, S. A., Fung, J. J., Burckart, G. J., McCurry, K. R. and Griffith, B. P. (2006). A randomized trial of inhaled cyclosporine in lung-transplant recipients. *The New England Journal of Medicine* **354**(2), 141–150.

Jakob, S. M., Ruokonen, E., Grounds, R. M., Sarapohja, T., Garratt, C., Pocock, S. J., Bratty, J. R. and Takala, J. (2012). Dexmedetomidine vs midazolam or propofol for sedation during prolonged mechanical ventilation: Two randomized controlled trials. *Journal of American Medical Association* **307**, 1151–1160.

Jennison, C. and Turnbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall/CRC.

Krueger, G. G., Langley, R. G., Leonardi, C., Yeilding, N., Guzzo, C., Wang, Y., Dooley, L. T. and Lebwohl, M. (2007). A human interleukin-12/23 monoclonal antibody for the treatment of psoriasis. *The New England Journal of Medicine* **356**, 580–592.

Lai, H. L., Chen, C. J., Peng, T. C., Chang, F. M., Hsieh, M. L., Huang, H. Y., and Chang, S. C. (2006). Randomized controlled trial of music during kangaroo care on maternal state anxiety and preterm infants' responses. *International Journal of Nursing* **43**, 139–146.

Lan, K. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Lehmann, E. L. (2004). *Elements of Large-Sample Theory.* Springer.

Ma, W., Hu, F. and Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association* **110**, 669–680.

Molander, A., Warfvinge, J., Reit, C. and Kvist, T. (2007). Clinical and radiographic evaluation of one- and two-visit endodontic treatment of asymptomatic necrotic teeth with apical periodontitis: A randomized clinical trial. *Journal of Endodontics* **33**, 1145–1148.

Nordle, O. and Brantmark, B. (1977). A self-adjusting randomization plan for allocation of patients into two treatment groups. *Clinical Pharmacology & Therapeutics* **22**, 825–830.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Ohtori, S., Miyagi, M., Eguchi, Y., Inoue, G., Orita, S., Ochiai, N., Kishida, S., Kuniyoshi, K., Nakamura, J., Aoki, Y., Ishikawa, T., Arai, G., Kamoda, H., Suzuki, M., Takaso, M., Furuya, T., Kubota, G., Sakuma, Y., Oikawa, Y., Toyone, T. and Takahashi, K. (2012). Efficacy of epidural administration of anti-interleukin-6 receptor antibody onto spinal nerve for treatment of sciatica. *European Spine Journal* **21**, 2079–2084.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* **34**, 585–612.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

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

Proschan, M. A., Lan, K. and Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials, A Unified Approach*. Springer Science+Business Media, LLC.

Rosenberger, W. F. and Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science* **23**, 404–419.

Shao, J., Yu, X. and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika* **97**, 347–360.

Signorini, D. F., Leung, O., Simes, R. J., Beller, E. and Gebski, V. J. (1993). Dynamic balanced randomization for clinical trials. *Statistics in Medicine* **12**, 2343–2350.

Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *J. Clin. Pharmacol. Therap.* **15**, 443–453.

Tilley, B. C., Alarcón, G. S., Heyse, S. P., Trentham, D. E., Neuner, R., Kaplan, D. A., Clegg, D. O., Leisen, J. C., Buckley, L., Cooper, S. M., Duncan, H., Pillemer, S. R., Tuttleman, M., and Fowler, S. E. (1995). Minocycline in rheumatoid arthritis. A 48-week, double-blind, placebo-controlled trial. *Annals of Internal Medicine.* **122**, 81–89.

Tsiatis, A. A., Rosner, G. L., and Tritchler, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika* **72**, 365–373.

Wald, A. (1947). *Sequential Analysis*. John Wiley and Sons Inc., New York.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–200.

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

Zhu, H. and Hu, F. (2010). Sequential monitoring of response-adaptive randomized clinical trials. *The Annals of Statistics* **38**, 2218–2241.

Zhu, H. and Hu, F. (2012). Interim analysis of clinical trials based on urn models. *Canadian Journal of Statistics* **40**, 550–568.

Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, 1200 Pressler Street, W922, Houston, TX 77030, USA.

E-mail: hongjian.zhu@uth.tmc.edu

Department of Statistics, George Washington University, Rome Hall, 801 22nd St NW, Washington, DC, 20052, USA.

E-mail: feifang@gwu.edu

(Received July 2016; accepted July 2017)