Statistica Sinica Preprint No: SS-2016-0330.R2	
Title	Sequential Monitoring of Covariate-Adaptive Randomized Clinical Trials
Manuscript ID	SS-2016-0330.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0330
Complete List of Authors	Feifang Hu and Hongjian Hu
Corresponding Author	Feifang Hu
E-mail	feifang@gwu.edu
Notice: Accepted version subject to English editing.	

Sequential Monitoring of Covariate-Adaptive Randomized Clinical Trials

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, and many clinical trials are 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. Both 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, stratified permuted block randomization, Pocock–Simon's 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, which 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 et al. (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 et al. (2010) and Ma et al. (2015). However, both papers focused on the final test statistic instead of the sequential statistics (a stochastic process).

While CAR procedures are very 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. However, they did not take into account the problems caused by covariate adaptive designs and the scenario where not all the design covariates were used in the analysis. Tsiatis et al. (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). Note that 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 real trials and the advantages mentioned above, 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 real 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 for decades. More importantly, the clinical trials that employ this procedure lack complete theoretical support, and many of these trials may 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 the following 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, which is the asymptotic joint distribution for complete randomization and fixed designs. As mentioned before, 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, theoretically 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 we can see its difference from standard Brownian motion. As a result, trials that ignore the difference between CAR proce-

dures and complete randomization could give misleading conclusions. The above theoretical results provide guidance for practical clinical trials, and they are one of the major contributions of this paper. 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, which offers clinical trialists practical steps to deal with these complex situations.

Finally, we perform extensive numerical studies for the above three scenarios in terms of the type I error, power, and early stopping. We also redesign a double-blind randomized 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. Conclusion 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 $(Z_1, \ldots, Z_q)$. For simplicity, we use one dimensional covariates to describe our framework and theorems. It is easy to generalize the results in this paper 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 $\boldsymbol{W}_i^Z = (Z_{i1}, \ldots, Z_{iq})$. In the paper, $(X_1, \ldots, X_p)$ 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 Y_i follows the following linear model:

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

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, their expectations are all 0, i.e., $E(X_{ik}) = 0, E(Z_{ij}) = 0, i = 1, \ldots, n, k = 1, \ldots, p, j = 1, \ldots, q$. We also assume that the errors are independent with 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

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

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

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

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 $C = \{l : \text{index of continuous covariates among } X_l, l = 1, \ldots, p\}$, $C^* = \{l : \text{index of continuous covariates among } Z_l, l = 1, \ldots, q\}$, 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 also need the following notation to formulate the main theorem. Suppose $\tilde{X}_k$ has s_k levels and $\tilde{Z}_j$ has s_j^* levels, and let $\boldsymbol{W}_i = (x_{i1}^{c_1}, \ldots, x_{ip}^{c_p}, z_{i1}^{c_1^*}, \ldots, z_{iq}^{c_q^*})$ represents 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 following hy-

pothesis 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 $L = (1, -1, 0, \ldots, 0)$, $\hat{\boldsymbol{\eta}}(t) = (\boldsymbol{X}(\lfloor nt \rfloor)^T \boldsymbol{X}(\lfloor nt \rfloor))^{-1} \boldsymbol{X}(\lfloor nt \rfloor)^T \boldsymbol{Y}(\lfloor nt \rfloor)$, $\hat{\sigma}(t)^2 = [\boldsymbol{Y}(\lfloor nt \rfloor) - \boldsymbol{X}(\lfloor nt \rfloor)\hat{\boldsymbol{\eta}}(t)]^T [\boldsymbol{Y}(\lfloor nt \rfloor) - \boldsymbol{X}(\lfloor nt \rfloor)\hat{\boldsymbol{\eta}}(t)]/(\lfloor nt \rfloor - p - 2)$. The sequential statistics (2.3) are just 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 to this question is the asymptotic joint distribution of the sequential statistics and the subsequent choices of critical values. In the literature, numerous techniques have been proposed for sequentially monitoring a Brownian motion that follows complete randomization. However, CAR procedures lead to considerable difficulties in deriving the joint distributions of the sequential test statistics. Following CAR procedures, the sequential treatment assignments are not independent of the covariate profiles, the observed responses are not independent of previous treatment assignments and covariates, and the observed responses are

not independent of each other. This could be the main reason for the lack of literature on this topic.

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

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

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

The following theorem offers vital theoretical support for the sequential monitoring of CAR procedures, and its implications for the practical procedure will be discussed in Section 2.4.

Theorem 1. *Let $B_t^{adj} = \sqrt{t} Z_t^{adj}$ in the space $D[0,1]$ with the Skorohod topology. Suppose 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. Therefore, the sequence of test statistics $\{Z_{t_1}^{adj}, ..., Z_{t_K}^{adj}, 0 \leq t_1 \leq t_2 \leq ... \leq t_K \leq 1\}$ has the asymptotic canonical joint distribution defined by Jennison and Turnbull (2000), i.e.,

(i) $\{Z_{t_1}^{adj}, ..., Z_{t_K}^{adj}\}$ 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.

This theorem reveals the effect of CAR procedures on the joint distribution of the sequential statistics, which is asymptotically the same as that of complete randomization after adjustment. In practice, clinical trialists implement this procedure assuming that it is exactly the same as complete randomization. From this theorem, we can easily see the gap and even calculate the difference given the parameter values. To the best of our knowledge, this paper provides the theoretical foundation for this procedure for the first time. Some other remarks are as follows.

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) Note that the asymptotic variance (2.5) of Z_t is always less than 1; it represents that 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 will lead to a conservative type I error rate. In addition, without adjustment, the power will also be adversely affected, which effectively increases the sample size needed and is not consistent with the original aim of sequential monitoring.

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

Here, we discuss two special cases of the above scenario, i.e., data analysis with all the covariates used in the randomization, and Student's t-test without any covariates. First, assume that the ith subject's response Y_i follows the following linear model:

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

where the notation is the same 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. Then we have the following theorem.

Theorem 2. *Let $B_t = \sqrt{t}Z_t$ in the space $D[0,1]$ with the Skorohod topology. Suppose 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. Therefore, the sequence of*

test statistics $\{Z_{t_1}, ..., Z_{t_K}, 0 \leq t_1 \leq t_2 \leq ... \leq t_K \leq 1\}$ *has the asymptotic canonical joint distribution defined by 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.*

A major difference between the first two theorems is that we do not have to adjust the sequential statistic (2.3) in this case, because its asymptotic properties are exactly the same as those of complete randomization.

Another special case occurs when the CAR is used to sequentially allocate patients, and the data is analyzed with Student's t-test, or equivalently using the following model:

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

To make the notation consistent with that of the previous sections, we assume that the responses follow the following model:

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

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

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

We then have the following theorem.

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.*

As explained in the Introduction, stratified permuted block randomization and Student's t-test are the most popular combination in real clinical trials. The above theorem 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.

First, we discuss how to obtain the consistent estimator ($\hat{\epsilon}(t)$) based on the data collected by information time t. In some cases it may be

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 these estimators may have diverse desirable features. To make our methods acceptable to a wide audience, 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 defined by Jennison and Turnbull (2000). As a result, numerous 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.

In this paper, 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 will be performed at three time points: $t_1 = 0.2$, $t_2 = 0.5$, and $t_3 = 1$. We also assume that the following three sets of boundaries from Proschan et al. (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 et al. (2006). In the numerical studies, to save space 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 study 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 misspecification. In Table 6, we specially 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 follow

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

where $(\mu_1, \mu_2, \gamma_1, \gamma_2)$ are unknown parameters, and ϵ_i are independent errors from the normal distribution $N(0, 1)$. In this paper, we study three randomization procedures, i.e., 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). Note that 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 discretize them in the following way:

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

where $z_{0.4}$ is the 0.4-quantile of the standard normal distribution. All the results are based on 10000 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 are used in the data analysis, the sequential monitoring of all three randomization procedures without adjustment can control the type I error rate well, which is consistent with Theorem 2. As a result, we do not have to adjust the sequential statistics in this case. Actually, 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 find 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. Further, data analysis with Student's t-test is more conservative than that based on partial covariates. As mentioned

before, the theorems allow an explicit calculation of the gap between the unadjusted rate and the adjusted rate for different scenarios. The above numerical results are consistent with the theoretically derived discrepancy.

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

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 as before. The value of μ_2 is chosen so that the power is around 0.8 for the sequential monitoring of complete randomization when the "full" model is used. We find 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 10000 replications. The sequential monitoring of CAR designs with adjusted sequential statistics stops the trials much earlier than the other approaches do.

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

Next, we discuss the performance of the proposed method when the model is mis-specified. In Table 4, we consider the case where Z_1 follows a Bernoulli distribution with a success rate of 0.5 and Z_2 is correlated with

Table 3: Early stopping for different scenarios

	Full		Partial		t-test	
	discrete	continuous	discrete	continuous	discrete	continuous
CR	1595	1680	1262	630	951	382
PS	1621	1643	938	293	516	90
SPB	1599	1682	892	314	495	104
PS-adj	NA	NA	1694	1083	1710	771
SPB-adj	NA	NA	1680	1062	1689	741

Z_1 in the following way:

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

The other settings are 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 can 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 will lead to a lower power, and adjustment can help us to obtain similar

powers for different scenarios.

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	1648	0.046	0.726	1299	0.051	0.570	761
PS	0.052	0.802	1696	0.033	0.742	1111	0.011	0.599	358
SPB	0.048	0.791	1619	0.030	0.738	1057	0.011	0.592	303
PS-adj	NA	NA	NA	0.058	0.810	1834	0.052	0.801	1736
SPB-adj	NA	NA	NA	0.051	0.799	1726	0.048	0.792	1632

In Table 5, we consider another scenario of model mis-specification where there are unobserved covariates that influence the responses. We assume that the responses follow

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

where $\gamma_3 = 1$ and Z_3 follows 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 are implemented with respect to only Z_1 and Z_2, "Full" in Table 5 means that both Z_1 and Z_2 are included in the data analysis, and "Partial" means that only Z_1 is included in the data analysis. Our proposed method (i) is robust under this scenario in terms of the type I error rate,

(ii) increases the power, and (iii) stops the trial much earlier compared to using the unadjusted statistics.

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	1194	0.052	0.625	914	0.051	0.561	729
PS	0.048	0.706	1182	0.031	0.638	704	0.021	0.568	404
SPB	0.053	0.712	1230	0.033	0.642	746	0.021	0.572	452
PS-adj	NA	NA	NA	0.050	0.709	1174	0.051	0.708	1209
SPB-adj	NA	NA	NA	0.053	0.714	1231	0.054	0.712	1240

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. Specifically, we consider the case where Z_1 follows a Bernoulli distribution with a success rate of 0.5, but Z_2 follows a Bernoulli distribution with a success rate of 0.9. Other settings are the same as Table 4. The advantages of our methods displayed in previous tables remain under this scenario. Therefore, 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 af-

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	1588	0.053	0.755	1365	0.053	0.670	1018
PS	0.049	0.795	1650	0.040	0.767	1337	0.023	0.693	776
SPB	0.051	0.792	1591	0.042	0.765	1324	0.024	0.693	739
PS-adj	NA	NA	NA	0.050	0.793	1646	0.052	0.793	1670
SPB-adj	NA	NA	NA	0.052	0.792	1629	0.051	0.791	1641

fecting the small joints and causing painful swelling. It will eventually result 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 is 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 treatment, 107 control). Two binary covariates are used in the model: Z_1 is the indicator of "oral corticosteroids used at entry" with a success rate of 0.32, and Z_2 is 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 is

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

with the residual following the normal distribution $N(0, 3.39^2)$.

In this section, we generate covariate data based on the above summary statistics, sequentially allocate the patients using CAR, and generate responses based on the fitted model (4.1). To provide more information, we use different time points from the previous sections to perform the sequential monitoring; these are $t_1 = 0.5$, $t_2 = 0.8$, 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 real data has a relatively large variance of error.

It becomes 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.

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	205	0.94	8178	NA	NA
Partial		0.935	8081	0.942	8196
t-test		0.928	7911	0.942	8202
Full	100	0.681	4770	NA	NA
Partial		0.668	4621	0.686	4808
t-test		0.650	4399	0.689	4880

5. Conclusion

The properties of the sequential monitoring of covariate-adaptive ran-

domized clinical trials is not well studied in the literature. In this paper, we have derived the joint distribution of the sequential statistics for three common scenarios in clinical trials that use CAR procedures for randomization and linear regression models for analysis. Based on these theoretical properties, we have proposed practical approaches to make use of existing critical values to control the type I error rate. We have also numerically studied different procedures and redesigned a clinical trial. The results demonstrated that the type I error rate can be protected as indicated by our theoretical conclusions, and they also showed the advantages of the combination of sequential monitoring and covariate adaptive designs.

There are several important directions for future research. First, we have studied data analysis for continuous responses with linear regression. Binary responses with logistic regression are a natural generalization. Many other types of responses and models deserve study; difficulties could be introduced by the nonexistence of a closed form of the parameter estimators. Second, 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. Third, a generalized structure of covariates could be investigated for other scenarios in real clinical trials. Fourth, other approaches to adjust the se-

quential statistics could be developed. Finally, Hu and Rosenberger (2006) classified adaptive randomization procedures into four categories, i.e., 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 is worth an investigation.

Supplementary Materials

The proofs are in the online supplementary materials.

Acknowledgements

We highly 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. *Biom. J.* **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. *Brit. J. Radiology* **44**, 793-797.

Heritier, S., Gebski, V. and Pillai, A. (2005). Dynamic balancing randomization in controlled clinical trials. *Stat. Med.* **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 (Vol. 525). John Wiley & Sons.

Hu, Y. and Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. *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. *J. Amer. Statist. Assoc.* **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. *Int. J. Nurs. Stud.* **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. *J. Amer. Statist. Assoc.* **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. *Clin. Pharm. Therap.* **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. *Brit. J. 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. *Stat. Med.* **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. *Ann. Intern. Med.* **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. *J. Amer. Statist. Assoc.* **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.