Summer 5-2019

# NUMERICAL STUDY FOR SEAMLESS CLINICAL TRIALS WITH COVARIATE ADAPTIVE RANDOMIZATION

MENGXI WANG
*UTHealth School of Public Health*

NUMERICAL STUDY FOR SEAMLESS CLINICAL TRIALS WITH COVARIATE ADAPTIVE RANDOMIZATION

by

MENGXI WANG, BS, PHD

APPROVED:

RUOSHA LI, PHD

HONGJIAN ZHU, PHD

JOHN MICHAEL SWINT, PHD

DEDICATION

To Guwei A. Chen

# NUMERICAL STUDY FOR SEAMLESS CLINICAL TRIALS WITH COVARIATE ADAPTIVE RANDOMIZATION

by

MENGXI WANG
PHD, University of Massachusetts Medical School, 2013
BS, Wuhan University, 2007

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
May, 2019

# NUMERICAL STUDY FOR SEAMLESS CLINICAL TRIALS WITH COVARIATE ADAPTIVE RANDOMIZATION

Mengxi Wang, BS, PhD, MS
The University of Texas
School of Public Health, 2019

Thesis Chair: Hongjian Zhu, PhD

One important goal of the pharmaceutical industry is to evaluate new therapies in a time-sensitive and cost-effective manner without undermining the integrity and validity of clinical trials. Adaptive seamless phase II/III designs (ASD) have gained popularity for accelerating the drug development process and reducing cost. Covariate adaptive randomization (CAR) is the most popular design in randomized controlled trials to ensure valid treatment comparisons by balancing the prognostic characteristics of patients among treatment groups. Although adaptive seamless clinical trials with CAR have been implemented in practice[1], the theoretical understanding of such designs is limited. In addition, current approaches to control the Type 1 error rate in seamless trials are based on theories for complete randomization, which may be invalid under CAR and lead to a Type 1 error rate that deviates from the nominal level. Recently, Ma and Zhu (2019, unpublished) established the theoretical foundation for the adaptive seamless phase II/III trial with CAR and proposed a hypothesis testing approach to control the Type 1 error rate in such trials. In the current

research, numerical studies were conducted to investigate the feasibility and advantages of the proposed approach in the seamless design with stratified permutated block (SPB) randomization. The simulation results revealed that the newly developed method well controlled the Type 1 error rate around the nominal level, improved the statistical power compared to the standard two sample *t*-test and increased the number of replications that the best treatment is selected for Stage II of the seamless trial under the SPB design compared to the complete randomization, which could promote its application in practice.

TABLE OF CONTENTS

LIST OF TABLES

# BACKGROUND

## Literature Review

### 1. Randomized Controlled Trials

Randomized controlled trial (RCT) is a study that divides subjects into distinct groups by random process to compare the effect of treatments or other interventions. It is the gold standard in clinical research due to its potential to minimize various biases and inherent strength to unveil casuality[2]. Randomization, in conjugation with the controlled and blinding provides a powerful tool to achieve accurate and valid estimates of the treatment effects for various medical interventions.

### 1.1 Randomization methods

#### *1.1.1 Simple randomization*

One key feature of RCT is randomization, which helps to ensure that the treatment and control groups are well balanced in both measured and unmeasured factors. Many procedures have been implemented in randomization of RCTs. Simple randomization, which is equivalent to repeated fair coin-tossing, is the most basic method for random assignments[3,4]. It prevents any conscious or unconscious selection bias by assigning subjects to treatment groups completely at random. In a large clinical trial ($n > 200$), simple randomization allows adequate balance in both sample size and prognostic factors (ie. covariates) among treatment groups[5]. However, accidental bias may occur due to chance imbalances in group sizes and pre-specified covariates when the trial size is small ($n < 100$)[5,6]. These imbalances can impair the precision and validity of comparisons among treatments and are often blamed for failures in clinical trials[7]. Therefore, randomization techniques that help

ensure balanced group sizes and baseline characteristics have been largely adopted in practice in contrast to simple randomization[6].

### *1.1.2 Permuted block randomization*

Permuted block randomization (PBR) is the most commonly used method to balance the number of subjects allocated to each treatment group[8]. Blocks are small and balanced with predetermined treatment assignments, ensuring equal number of subjects in each treatment group through the whole trial process. Block size is often a multiple of the group number. All permutations for the block size (ie. all possible combinations of treatment assignments within the block) are listed once the block size is determined, from which the treatment allocation for a subject is decided by random selection of the permutations. Although PBR can help ensure balanced number of subjects in each treatment group to maximize statistical power, the baseline covariates among groups may not be comparable, resulting in confounding bias and impaired power of the study[4,5].

### *1.1.3 Stratified permuted block randomization*

To achieve covariate balance in clinical trials[6,8], multiple coactive-adaptive randomization (CAR) approaches have been proposed, among which stratified permuted block randomization is the most common one in both academia and industry sponsored clinical trials[6,9,10]. In a trial using stratified PBR, subjects are divided into different strata based on measurable prognostic factors that are considered strongly associated with the primary outcome, such as trial centers and disease stages. The total number of strata is the product of the number of levels across the covariates. Permuted block randomization is then performed within each stratum to assign a subject to one of the treatment groups[11]. Stratified

PBR helps achieve proper balance in both group sizes and pre-specified covariates, leading to maximal statistical power and minimal confounding bias. However, only a small number of prognostic factors can be balanced using this procedure. When there are too many influential covariates or covariates with many levels in the trial, a large number of strata will be generated. Some strata may have few or even no subjects if the trial is small, resulting in imbalances in overall treatment allocations and jeopardizing the validity of the study. Therneau *et al.* showed that the balance in covariates begins to fail if the total number of strata is greater than approximately half of the sample size[12]. One early study indicated that the number of strata should be less than N/B to maintain the benefits of stratified PBR, where N is the total sample size and B is the block size[13]. Kerman suggested a more stringent number by multiplying B with a safe factor 4 in the denominator (ie. N/4B)[14]. In practice, a mean of 2.52 (SD = 0.90) stratification variables were used in clinical trials with stratified PBR[6].

**1.2 Traditional clinical trial design**

Conventional drug and medical intervention development consists of a sequence of independent RCTs organized in different phases. To develop a novel drug for certain disease in a classical way, pre-clinical investigations are firstly carried out to study drug's safety, pharmacodynamics and pharmacokinetic on animals and to evaluate drug production and purity. If the study results are promising, the drug is further investigated in human subjects in four sequential phases after approval by US Food and Drug Administration (FDA) for each phase. In phase I trials, drug's safety, maximum tolerated dose and human pharmacodynamics and pharmacokinetic are tested in 20 to 100 healthy volunteers or people

with the disease/condition. Phase II trials, also referred to as exploratory or learning phase trials, provide preliminary evidence of drug efficacy besides further investigating the safety and pharmacological issues in a larger diseased population (up to several hundred). In these trials, multiple doses of a new drug may be compared to a control (ie. conventional treatment or placebo) to decide whether to stop or continue with the development. Data collected in this phase can also help to refine research questions, develop research methods and design research protocols for phase III trials. Known as pivotal studies, phase III trials demonstrate or confirm treatment efficacy and identify incidence of side effects in a population ranging from 300 to 3000 subjects. The trials can last up to 4 years and sometimes more than one phase III trials are required to establish drug efficacy and safety by FDA. Because of the large scope and long duration of the trials, long-term and rare side effects are more likely to be detected in the studies. Statistical analyses for phase III trials are typically conducted by ignoring data collected in previous phases, and the outcome measures are usually different from phase II trials. Phase IV trials are conducted once the drug is approved by FDA during the post-market safety monitoring. More rare adverse reactions can be identified and health economic evaluations could be implemented in this phase[15,16].

**1.3 Adaptive seamless phase II/III design**

To accelerate the process of drug development and reduce its cost, adaptive seamless phase II/III designs (ASD) have been developed, whereby the two phases are combined into a single, uninterrupted trial with two or more stages. Typically, the stages are separated by one or more interim analyses, at which several experimental treatments or drug doses are evaluated. At these interim looks, experimental treatments are either dropped for futility or

continued to be investigated in the later stage(s) due to high treatment efficacy. Adaptions such as sample size reassessment are allowed after the interim analyses[17,18].

### *1.3.1 Principle of ASD*

A simple scenario of ASD is comparing multiple experimental treatments with one control in a two-stage design with one interim analysis. Based on the data from the learning stage I (analogous to the traditional phase II), the study is either stopped due to futility or continued to the confirmatory stage II (analogous to the traditional phase III) along with the empirically best experimental treatment and the control. The final analysis of the selected treatment includes information from both stages and the overall type one error of the statistical analysis is controlled at a pre-specified level independent of the selection rule at the interim. Bauer and Köhne proposed a test procedure for this scenario in 1994[19]. Generally, a one-sided null hypothesis corresponding to comparing different effects between two treatments is tested. The test procedure for Stage I, the stopping rule for the interim analysis and the combination test for the final analysis are pre-determined. Hypothesis testing for Stage I results in a *p*-value $p_1$, to which the interim decision is made accordingly. If the study is continued, the second stage can be re-designed, e.g, sample size is re-assessed, and a *p*-value $p_2$ is obtained from hypothesis testing for Stage II. The two *p*-values are combined in the end using the pre-specified combination test to decide whether the null hypothesis is rejected or not[17].

### *1.3.2 Combination tests*

In the above adaptive test procedure proposed by Bauer and Köhne, fisher's combination test, i.e. the inverse $\chi^2$ method was recommended[19]. Fisher's criterion results in rejection of the null hypothesis at the end of the two-stage trial if

$$-\log(p_1 p_2) > \chi^2_{4,1-\alpha}/2,$$

where $\chi^2_{4,1-\alpha}$ is $(1-\alpha)th$ quantile of a $\chi^2$ distribution with 4 degrees of freedom.

Another common approach for combination test is the weighted inverse normal method[20]. For an adaptive seamless phase II/III trial with two stages, given two one-sided *p*-values, $p_1$ and $p_2$ from each stage, the method rejects the null hypothesis in the final analysis if

$$w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2) > \Phi^{-1}(1-\alpha),$$

where $w_1$ and $w_2$ are pre-specified weights satisfying $0 < w_i < 1, i = 1,2$ and $w_1^2 + w_2^2 = 1$. $\Phi$ denotes the standard normal CDF. A widely adopted option is $w_i = \sqrt{n_i/n}, i = 1,2$, where $n_i$ are the pre-planned sample sizes for the two stages and $n = n_1 + n_2$ [18].

## 2. Multiplicity in Clinical Trials

Multiplicity is defined as simultaneous assessments of multiple aspects of the efficacy profile in a clinical trial[21]. It may inflate type I error rate and lead to increased risk of false positive conclusions in trials which evaluate multiple end points, compare across several treatment arms, analyze multiple sub-populations and measure the same outcome repeatedly in time[22,23]. For example, in an exploratory dose-comparison study with two dose candidates,

or in a confirmatory trial with two primary end points, two true null hypotheses (i.e. no treatment effect) are tested simultaneously at significance level $\alpha$, which refers to the probability of a type I error. The probability of rejecting at least one true null hypothesis, known as overall type I error rate or familywise error rate (FWER), can be calculated as $1-(1-\alpha)^2$. When $\alpha = 0.05$, FWER is 0.0975, indicating that there is 9.75% of chance to declare at least one significant treatment effect when indeed none is significant by the trial sponsors, whereas others believe that the type I error rate is maintained at the level of 5%. Besides inflating type I error rate, multiplicity also has important implications for sample size determination[24]. Therefore, multiplicity adjustment is mandated by regulatory agencies to ensure accurate efficacy or safety claims[25,26].

**2.1 Multiplicity adjustments in clinical trials**

Numerous multiplicity adjustment methods, i.e. multiple testing procedures (MTPs) have been developed to solve multiplicity problems in clinical research. Identifying the most appropriate MTP for a particular clinical trial is essential to maximize the statistical power. In practice, clinical trialists customize solutions for addressing multiplicity by utilizing all available clinical and statistical information. When definitive clinical information is available, hypotheses for each individual objective, e.g. each end point in a phase III trial, can be ordered in a clinical meaningful way prior to data analysis. Data-driven hypothesis ordering will be used if a meaningful priori ordering cannot be pre-specified. Based on distributional information, MTPs can be classified into nonparametric tests, semiparametric tests and parametric tests, in which parametric tests are the most efficient but rely on specific

statistical assumptions. Methodology and applications of commonly used MTPs in clinical trials are thoroughly reviewed by Bretz *et al*[27]. and Dmitrienko *et al*[21,22]. Two methods, Simes test and Dunnett test are introduced here because they will be used in the proposed numerical study.

### *2.1.1 Simes test*

Simes test is a single-step MTP that assumes non-negative correlations between *p*-values of individual hypotheses and is more powerful than a global test based on Bonferroni test. Assume a multiplicity problem arising in a trial which compares k experimental treatments with a control. K null hypotheses denoted by $H_1,...,H_k$ correspond to evaluations of treatment effects parameterized as $\theta_1,...,\theta_k$, i.e., $H_i:\theta_i=0, i=1,...,k$. The Simes method tests the global null hypothesis

$$H=\bigcap H_i:\theta_1=...=\theta_k=0, i=1,...,k.$$

Let $p_{(i)}, i=1,...,k$ be ordered *p*-values for individual hypotheses. With $p_{(1)}<...<p_{(k)}$, the test rejects $H$ if $p_{(i)}\le i\alpha/k$ for at least one i. Simes' adjusted *p*-value is $\min_i kp_{(i)}/i$.

### *2.1.2 Dunnett test*

Dunnett test is a single-step parametric test assuming that the correlations between the test statistics are known. It provides less conservative multiplicity adjustment and is more powerful than the nonparametric test such as Bonferroni test. For the above multi-arm trial example, we want to test the k null hypotheses

$$H_i:\theta_i=0, i=1,...,k.$$

The test statistics are defined as

$$t_i = \frac{\bar{y}_i - \bar{y}_0}{s\sqrt{1/n_i + 1/n_0}}, i = 1,...,k,$$

where $\bar{y}_i$ and $\bar{y}_0$ are sample mean, and $n_i$ and $n_0$ are sample size for treatment k and control, respectively. s is the pooled sample standard deviation and $s^2 = \sum_{i=0}^{k}(n_i - 1)S_i^2/\nu$, where $S_i^2$ is the sample variance for treatment k and $\nu = \sum_{i=0}^{k} n_i - k - 1$. Under the null hypotheses, $(t_1,...,t_k)$ follows k-variate t-distribution with $\nu$ degrees of freedom and correlations

$$\rho_{i,j} = \sqrt{\frac{n_i}{n_i + n_0}}\sqrt{\frac{n_j}{n_j + n_0}}, i, j = 1,...,k.$$

For the k null hypotheses, dunnett test rejects $H_i$ if $t_i \geq c_{k,1-\alpha}$, where $c_{k,1-\alpha}$ is determined by $\Pr[(t_1,...,t_k) \leq (c_{k,1-\alpha},...,c_{k,1-\alpha})] = \Pr(\max_i t_i \leq c_{k,1-\alpha}) = 1-\alpha$.

**2.2. Closure principle**

Closure principle proposed by Marcus *et al.*[28] is a general construction method which allows one to draw conclusions for individual null hypothesis in multiple testing following a closed test procedure (CTP). It strongly controls FWER and serves as the foundation for all commonly used MTPs in clinical trials[21]. For a multi-arm trial where k experimental treatments are compared with a control, k null hypotheses $H_i : \theta_i = 0, i = 1,...,k$ are to be tested, where $\theta_i$ is the treatment effect. These initial hypotheses are called elementary hypotheses in the CTP. All possible intersection hypotheses, i.e., $H_I = \bigcap_{i \in I} H_i, I \subseteq \{1,...,k\}$ are constructed, and a local level-α test is performed for each of the intersection hypotheses.

An elementary hypothesis $H_i$ is rejected at FWER α if all $H_I$ with $i \in I$ are rejected, each at local level α, i.e., the maximum *p*-value from this set needs to be less than or equal to α. The adjusted *p*-value for $H_i$ is $q_i = \max_{I:i \in I} p_I, i = 1,...,k$, where $p_I$ denotes the p-value for a given intersection hypothesis $H_I, I \subseteq \{1,...,k\}$. Multiple MTPs can be used for different intersection hypotheses[17].

**2.3 Multiple testing in ASD**

The general idea for multiple testing in adaptive design is to construct all intersection hypotheses for each stage and to test each of the hypotheses with a suitable combination test[17,29]. Consider a simple two-stage seamless phase II/III design whereby two experimental treatments 1 and 2 are compared with a control in Stage I. Assume treatment 1 is selected in the interim analysis to go forward to Stage II, and we are interested in testing the null hypothesis $H_1 : \theta_1 = 0$ in the final analysis, where $\theta_1$ is the treatment effect for treatment 1. Following the CTP, intersection hypothesis $H_{12}$ is constructed and hypotheses $H_1, H_2$ and $H_{12}$ are to be tested for both stages. Let $p_{i,j}$ denote the one-sided *p*-value for hypothesis $H_j, j \subseteq \{1,2,12\}$ at stage i = 1, 2. Denote combination function C derived from the inverse $\chi^2$ method or the weighted inverse normal method. Since treatment 2 is dropped at the interim and no data is available for it in Stage II, $H_{12}$ for the second stage is equal to $H_1$ and the test is performed on data for treatment 1 in this stage. According to closure principle, $H_1$ is rejected at FWER α in the final analysis if $H_1$ and $H_{12}$ for both stages are simultaneously rejected at the significance level α, i.e., if $C(p_{1,12}, p_{2,1}) \le c$ and $C(p_{1,1}, p_{2,1}) \le c$.

**Public Health Significance**

Randomized controlled trials have been the gold standard for discovering efficient treatments and understanding counter-balancing risks. Many breakthroughs in disease prevention and treatment in the past century are attributable to RCTs, such as the landmark Salk Polio Vaccine Trial[30] and trials for tuberculosis control[31,32]. Ford *et al.* reported that approximately half of the decrease in the age-adjusted death rate for coronary heart disease from 1980 to 2000 can be attributed to medical therapies validated in clinical trials[33]. More than 200,000 clinical trials worldwide have been registered in the US National Institute of Health (NIH) today[34], leading medical innovations to improve health and well-being of human race.

One main barrier to conducting clinical trials in the US is high financial costs[35]. The estimated cost to develop and win marketing approval for a new drug had increased from US$802 million in 2003 to US$2.6 billion in a decade[36]. Lengthy timelines contribute directly to the costs of clinical trials. The average length of time from the start of clinical testing to marketing is 7.5 years[37], and typically 10 to 15 years are spent from discovery to registration with FDA for a drug therapy[38]. Long development process also reduces the time a drug has under patent protection, allowing early entry of generic competitors and reducing revenue for the patent holder. These obstacles may discourage pharmaceutical companies from investing in drug development and consequently limit patients' access to novel treatments.

The motivation behind adaptive seamless phase II/III designs is to save time for drug development. By combining the conventional phase II and III trials into one study, ASD

reduce the lead time between the two phases. It also increases data collection and interpretation efficiency while maintaining the same sample size by combing information in the final analysis. Alternatively, smaller sample sizes are required in the ASD to draw conclusions with the same strength as in the traditional designs. Furthermore, long-term safety data can be obtained earlier because patients in the learning stage I are continued to be monitored in the confirmatory stage II in the ASD.

In contrast to high financial costs, the clinical trial success rate is low. A recent study revealed that only 14% of drugs in clinical trials win FDA approval eventually, and the success rates range from 3.4% for cancer treatments to 33% in vaccines for infectious diseases[39]. Strict control of type I error rate at a two-sided 5% level is mandated for FDA approval[25]. For clinical trials with CAR, conventional tests are usually used without consideration of the randomization scheme[40]. Failing to incorporate all stratification covariates used for randomization into inference procedures results in conservative tests in terms of small type I error rate and reduced statistical power[41], therefore beneficial treatments may be denied to patients.

In this proposal, simulation studies are proposed to examine a new hypothesis testing procedure which is able to well calibrate the type I error rate at the significance level under CAR in adaptive seamless phase II/III clinical trials. The success of the research will facilitate the implementation of the procedure in the adaptive seamless trials with CAR to increase the trial success rate, leading to more efficient and cost effective clinical trials that benefit the general population related to the corresponding medical innovations.

**Hypothesis, Research Question, Specific Aims or Objectives**

**1. Research questions and hypotheses**

For seamless trials, type I error rate tends to inflate due to multiplicity and selection[42]. Classic MTPs such as Dunnett test can well control the type I error rate in the ASD with simple randomization, assuming independence of each patient[43]. Under CAR, however, the assumption is no longer true due to the complicated randomization mechanism aiming to balance covariates over different arms. For example, balancing covariates via stratification leads to correlation between the treatment groups[44]. For a scenario where one experimental treatment is compared to one control, it is reported that the classic tests are too conservative in terms of small type I errors when stratified PBR is used to allocate patients[44]. It is unclear if MTPs such as Simes test and Dunnett test are valid under CAR with multiple treatments in the setting of seamless designs, and if not, how to perform adjustment to achieve valid tests.

Zhu and Ma have studied the properties of intersection tests under CAR with multiple treatments, and have provided theoretical results for a hypothesis testing approach where Simes and Dunnett tests are still valid under CAR in the seamless design (unpublished). Suppose there are $(K+1)$ arms in Stage I of a seamless clinical trial with CAR, in which $K$ experimental treatments are compared to a control, and the total sample size is $N$. Let $\mu_k, k = 0,1,...,K$ be the parameter measuring the main effects of treatment $k$. Testing the $K$ null hypotheses $H_{0,k}: \mu_k = \mu_0, k = 1,...,K$ and their interactions are of interest. Let $\boldsymbol{T_i} = (T_{i0}, T_{i1},...,T_{iK})^T, i = 1,...,N$ be the treatment assignment of the $ith$ subject, where treatment 0 represents the control arm. $T_{ik} = 1, k = 0,1,...,K$, if the $ith$ subject is assigned to

treatment $k$, and $T_{ik} = 0$ otherwise. The number of subjects in treatment $k$ is $N_k = \sum_{i=1}^{N} T_{ik}, k = 0,1,...,K$. Let $\boldsymbol{Y_i} = (Y_{i0}, Y_{i1},..., Y_{iK})^T, i = 1,..., N$ be a random vector of response variables, where $Y_{ik}, k = 0,1,...,K$ is the response of the $ith$ subject assigned to treatment $k$. $\bar{Y}_k = \sum_{i=1}^{N} T_{ik} Y_{ik} \big/ N_k$ is an estimator of $\mu_k$. Let $Z_i$ be the covariate information for the $ith$ subject, which is independent and identically distributed. For simplicity, $Z_i s$ are assumed to be either discrete or continuous covariates. To incorporate continuous covariates in randomization, $Z_i$ is discretized with $D(Z_i)$, a discrete function of $Z_i$ taking values in D. $D(Z_i) = Z_i$ is set for discrete covariates. The response of the $ith$ subject under treatment $k$ follows the linear model

$$Y_{ik} = \mu_k + \beta Z_i + \varepsilon_{ik}, k = 0,1,...,K,$$

where $\beta$ represents the covariate effect, and $\varepsilon_{ik} s$ are independent and identically distributed random errors from the normal distribution $N(0, \sigma_\varepsilon^2)$, and are independent of covariates. Two conditions are introduced for the balancing properties under CAR with multiple treatments: (A) $N_k - N_0 = O_p(1), k = 1,..., K$; (B)

$\sum_{i=1}^{N} (T_{ik} - T_{i0}) I\{D(Z_i) = d\} = O_p(1), k = 1,..., K$ for any $d \in$ D.

The following theorems show how to construct test statistics for $H_{0,k} : \mu_k = \mu_0, k = 1,..., K$ that can be used for the Simes test and the Dunnett test.

**Theorem 1:** Let

$$X_k = \frac{\bar{Y}_k - \bar{Y}_0}{\sigma_d \sqrt{1/N_k + 1/N_0}}, k = 1,...,K ,$$

where $\sigma_d^2 = \sigma_\varepsilon^2 + \beta^2 E[Var\{Z_i \mid D(Z_i)\}]$. (1)

Under conditions (A) and (B), when all $H_{0,k} : \mu_k = \mu_0, k = 1,...,K$ are true, we have

$$(X_1,..., X_K)^T \xrightarrow{d} N(0, \boldsymbol{R}) ,$$

where $\boldsymbol{R} = diag\{\frac{1}{2}\mathbf{I}_K\} + \frac{1}{2}\mathbf{I}_K\mathbf{I}_K^T$.

Based on the theorem, the test statistic follows a standard normal distribution for every single test for $H_{0,k} : \mu_k = \mu_0, k = 1,...,K$ , and the critical value can be selected accordingly to one-sided or two-sided tests. In practice, the value of $\sigma_d$ can be estimated by model-based or bootstrap methods.

**Theorem 2 (The Simes test):** Under conditions (A) and (B), the Type I error rate is controlled for the Simes test with the test statistic $X_k, k = 1,...,K$ under CAR.

Original Dunnett test is based on the multivariate *t* distribution. In Theorem 1, it is proved that the vector of test statistics $(X_1,..., X_k)^T$ asymptotically follows $K$ dimensional normal distribution with unit variances and constant correlations equal to 1/2. Without loss of generality, the alternative hypotheses are assumed to be $H_{1,k} : \mu_k > \mu_0, k = 1,...,K$ . The null hypotheses are rejected if $\max_k X_k \geq c'$, where $c'$ is determined by $\Pr[(X_1,..., X_k) \leq (c',..., c')] = \Pr(\max_k X_k \leq c') = 1 - \alpha$ . The above test procedure is referred to as modified Dunnett test.

**Theorem 3 (The Dunnett test):** Under conditions (A) and (B), the type I error rate is asymptotically $\alpha$ for the modified Dunnett test using test statistic $X_k, k = 1,..., K$ under CAR.

Numerical studies, also known as simulation studies are computer experiments which create data by pseudo-random sampling from known probability distributions. They are excellent tools for evaluating novel statistical methods and comparing competing approaches. To assess the appropriateness and accuracy of the hypothesis testing procedure mentioned above in a seamless design with CAR, numerical studies are proposed to use in the current study with following aims.

**2. Specific Aims**

I. Detect potential problems in type I error control for commonly used hypothesis testing procedures in seamless phase II/III clinical designs with covariate adaptive randomization.

II. Implement the newly developed hypothesis testing procedure in seamless designs with covariate adaptive randomization to better control the type I error rates.

III. Demonstrate the statistical advantage of the new hypothesis testing procedure under the alternative hypothesis in seamless designs with covariate adaptive randomization.

## METHODS

### Study design

A seamless phase II/III trial design with two stages under stratified permuted block randomization (SPB) was considered and numerical studies were conducted based on the following settings. There were $(K+1)$ arms under study in Stage I, in which $K$ experimental treatments were compared to a control. Patients were sequentially assigned to all the arms with SPB based on $M$ covariates. One experimental treatment $k^*$ with the largest estimated treatment effect was selected at the end of the stage. In stage II, the planned number of patients was sequentially assigned to the control and the treatment $k^*$ with SPB based on $M$ covariates. The hypothesis $H_{0,k^*}: \mu_{k^*} = \mu_0$ vs. $\mu_{k^*} > \mu_0$ was tested based on the combined data from both stages. Scenarios with different number of treatments and stratification covariates, i.e., (1) $K = 2, M = 2$; (2) $K = 3, M = 3$; (3) $K = 4, M = 2$ were evaluated. For Scenario 1, both discrete and continuous stratification covariates were studied.

### Simulation procedures

In Scenario 1, two experimental treatments (i.e., treatment 1 and treatment 2) were compared with one control (i.e., treatment 0) in Stage I, and both discrete and continuous stratification covariates were considered. The following linear model with two covariates $Z_1$ and $Z_2$ was used to simulate response $Y_i$,

$$Y_i = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 T_{i2} + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \varepsilon_i;$$

where $T_{i1}$ and $T_{i2}$ were indicator variables indicating the treatment assignment of the $ith$ subject. $T_{ik} = 1, k = 1,2$ if the $ith$ subject was assigned to treatment $k$, and $T_{ik} = 0$ otherwise. $\alpha_1$ and $\alpha_2$ were the addictive effects of treatments on outcome, and $\beta_1$ and $\beta_2$ were coefficients corresponding to the stratification covariates. For the discrete case, $Z_1$ and $Z_2$ followed Bernoulli distributions with success rates $p_1$ and $p_2$, respectively. For the continuous case, $Z_2$ was generated from standard normal distribution and discretized into bernoulli variable $D(Z_2)$ with probability $p_2$ to be used in SPB randomization. More specifically, $D(Z_2) = 1$ if $Z_2 < Z_{(p_2)}$, where $Z_{(p_2)}$ was the $p_2$ quantile of the standard normal distribution, and $D(Z_2) = 0$ otherwise. Continuous covariate was used in statistical inference procedures. $\alpha_0$, $\beta_1$ and $\beta_2$ were assigned value of 1, and $\varepsilon_i \sim N(0, \sigma^2)$. 120 patients were assumed to sequentially enter the trial in Stage I, and the block size of 6 was used for SPB randomization with respect to covariates $Z_1$ and $Z_2$. Complete randomization (CR) was also implemented. To select the experimental treatment to go forward to Stage II, Let

$$W_k = \frac{\bar{Y}_k - \bar{Y}_0}{\sqrt{S_k^2 / N_k + S_0^2 / N_0}}, k = 1,2, \quad (2)$$

where $\bar{Y}_k$ was the mean response for the treatment $k, k = 1,2$, $\bar{Y}_0$ was the mean response for the control, $S_k^2$ and $S_0^2$ were the unbiased estimators of the variance for the treatment $k$ and the control, respectively. The experimental treatment with larger $W$, denoted as treatment $k^*$ was considered more effective and selected to continue.

In Stage II, 500 patients were simulated and allocated into the control and the treatment $k^*$ using SPB randomization with respect to the two covariates or complete randomization. The hypothesis tests for comparing treatment effects in both stages included the two-sample *t*-test without adjustment, full linear model with both covariates $Z_1$ and $Z_2$, the bootstrap *t*-test proposed by Shao *et al.*[45], and the newly developed procedure denoted as *t*-test with adjustment. For the bootstrap *t*-test, B bootstrap samples $(Y_1^{*b}, Z_{1,1}^{*b}, Z_{1,2}^{*b}),\ldots,(Y_N^{*b}, Z_{N,1}^{*b}, Z_{N,2}^{*b})$, b = 1, 2,…, B were generated independently by random sampling with replacement from $(Y_1, Z_{1,1}, Z_{1,2}),\ldots,(Y_N, Z_{N,1}, Z_{N,2})$. SPB randomization with categories defined by the covariate values of each bootstrap sample $(Z_{1,1}^{*b}, Z_{1,2}^{*b}),\ldots,(Z_{N,1}^{*b}, Z_{N,2}^{*b})$ was applied and the bootstrap analogues of treatment allocations $T_{1k}^{*b},...,T_{Nk}^{*b}$ could be obtained, where $T_{ik}^{*b} = 1, k = 0,1,2$ if the $ith$ subject was assigned to treatment $k$, and $T_{ik}^{*b} = 0$ otherwise. Define the treatment effects between the experimental treatments and the control as

$$\bar{Y}_j^{*b} - \bar{Y}_0^{*b} = \frac{1}{n_j^{*b}} \sum_{i=1}^{N} T_{ij}^{*b} Y_i^{*b} - \frac{1}{n_0^{*b}} \sum_{i=1}^{N} T_{i0}^{*b} Y_i^{*b}, n_0^{*b} = \sum_{i=1}^{N} T_{i0}^{*b}, n_j^{*b} = \sum_{i=1}^{N} T_{ij}^{*b}, j = 1,2.$$

The bootstrap estimator of the variance of $\bar{Y}_j - \bar{Y}_0$ was the sample variance of $\bar{Y}_j^{*b} - \bar{Y}_0^{*b}$, b = 1, 2,…,B, denoted by $\hat{v}_{Bj}$. The bootstrap *t*-test had the test statistic $T_B = (\bar{Y}_j - \bar{Y}_0) / \hat{v}_{Bj}^{1/2}$. B = 200 was used in the current simulations. For the *t*-test with adjustment based on Theorem 1, the value of $\sigma_d$ was computed by formula (1), where values of $\sigma_\varepsilon$ and $\beta$ were estimated by fitting a full linear regression model with both covariates. The closed test procedures using

Dunnett test and Simes test were applied for hypothesis testing to control the familywise error rate (FWER). Hypothesis $H_{0,k^*}: \mu_{k^*} = \mu_0$ vs. $H_{1,k^*}: \mu_{k^*} > \mu_0$ for the seamless design was tested using Fisher's combination test at the end of the stage. According to the closure principle, $H_{0,k^*}$ was rejected at FWER $\alpha$ if $H_{0,12}$ and $H_{0,k^*}$ were both rejected at $\alpha$, i.e., $C(p_{1,12}, p_{2,k^*}) \le c_\alpha$ and $C(p_{1,k^*}, p_{2,k^*}) \le c_\alpha$. The significance level $\alpha$ was 0.05 for all the tests. All the results were based on 10, 000 replications.

In Scenario 2, three experimental treatments (i.e., treatment 1, 2 and 3) were compared with one control (i.e., treatment 0) in Stage I, and three discrete stratification covariates ($Z_1, Z_2$ and $Z_3$) were considered. The following linear model was used to simulate response $Y_i$,

$$Y_i = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 T_{i2} + \alpha_3 T_{i3} + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \varepsilon_i;$$

where $T_{i1}$, $T_{i2}$ and $T_{i3}$ were indicator variables indicating the treatment assignment of the $ith$ subject. $T_{ik} = 1, k = 1,2,3$ if the $ith$ subject was assigned to treatment $k$, and $T_{ik} = 0$ otherwise. $\alpha_1$, $\alpha_2$ and $\alpha_3$ were the addictive effects of treatments on outcome, and $\beta_1$, $\beta_2$ and $\beta_3$ were coefficients corresponding to the stratification covariates. $Z_1, Z_2$ and $Z_3$ followed Bernoulli distributions with success rates $p_1$, $p_2$ and $p_3$, respectively. $\varepsilon_i \sim N(0, \sigma^2)$. CR and SPB randomization with respect to all the three covariates were implemented. In Scenario 3, four experimental treatments (i.e., treatment 1, 2, 3 and 4) were compared with one control (i.e., treatment 0) in Stage I, and two discrete stratification

covariates ( $Z_1$ and $Z_2$ ) were considered. The following linear model was used to simulate response $Y_i$ ,

$$Y_i = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 T_{i2} + \alpha_3 T_{i3} + \alpha_4 T_{i4} + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \varepsilon_i ;$$

where $T_{i1}$, $T_{i2}$, $T_{i3}$ and $T_{i4}$ were indicator variables indicating the treatment assignment of the $ith$ subject. $T_{ik} = 1, k = 1,2,3,4$ if the $ith$ subject was assigned to treatment $k$ , and $T_{ik} = 0$ otherwise. $\alpha_1$, $\alpha_2$, $\alpha_3$ and $\alpha_4$ were the addictive effects of treatments on outcome, and $\beta_1$ and $\beta_2$ were coefficients corresponding to the stratification covariates. $Z_1$ and $Z_2$ followed Bernoulli distributions with success rates $p_1$ and $p_2$, respectively. $\varepsilon_i \sim N(0, \sigma^2)$. CR and SPB randomization with respect to the two covariates were implemented.

In both scenarios, sample sizes for Stage I and II were 200 and 400, respectively. The block sizes for the SPB design are 8 and 10, respectively. Other settings were the same as in Scenario 1.

Type I error rates were calculated as the rate of rejection of $\mathrm{H}_{0,k^*} : \mu_{k^*} = \mu_0$ in the two-stage seamless design among all 10,000 simulation replications assuming $\mu_k = \mu_0$ in Stage I and $\mu_{k^*} = \mu_0$ in Stage II. Powers were computed as the rate of rejection of $\mathrm{H}_{0,k^*} : \mu_{k^*} = \mu_0$ in the two-stage design among all 10,000 simulation replications assuming $\mu_k > \mu_0$ in Stage I and $\mu_{k^*} > \mu_0$ in Stage II. Number of replications that the best treatment is selected for Stage II was the counts that treatment 1 was selected at the end of Stage I among all 10, 000 simulation replications assuming treatment 1 had the largest treatment effect.

## RESULTS

To study Type 1 error rates, no difference in treatment effects was assumed by assigning $\alpha_1 = \alpha_2 = 0$. Results from Scenario 1 were reported in Table 1 (discrete case) and Table 2 (continuous case). Under CR in both cases, Type 1 error rates were close to the nominal level 5% for both the two sample *t*-test (*t-test*) and the full linear model (*lm($Z_1$, $Z_2$)*). Under the SPB randomization with either Dunnett test or Simes test for multiplicity adjustment, the two-sample *t*-tests were conservative with Type I error rates far below 5%, while the *t*-tests with adjustment (*Adjusted-t*) successfully controlled Type 1 error rates around 5%. Type 1 error rates were also well controlled when the full linear model and bootstrap *t*-test (*BS-t*) were used in both discrete and continuous cases. Different values of ( $p_1$, $p_2$, $\sigma$ ) were investigated and similar results were obtained as shown in the tables.

Table 1: Simulated Type I error under stratified permuted block design (SPB) and complete randomization (CR) in % in the seamless trial with three treatments and two discrete covariates.

| MTP | ($p_1$, $p_2$, σ) | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* |
|---|---|---|---|---|---|---|
| Simes | (0.5, 0.5, 1.0) | SPB | 1.73 | 5.26 | 5.14 | 5.20 |
| | | CR | 5.00 | 4.73 | - | - |
| | (0.4, 0.6, 1.0) | SPB | 1.78 | 4.84 | 5.35 | 5.41 |
| | | CR | 4.73 | 4.80 | - | - |
| | (0.4, 0.6, 1.5) | SPB | 3.00 | 4.78 | 5.46 | 5.31 |
| | | CR | 4.61 | 4.65 | - | - |
| Dunnett | (0.5, 0.5, 1.0) | SPB | 1.98 | 5.75 | 5.09 | 5.46 |
| | | CR | 5.20 | 5.30 | - | - |
| | (0.4, 0.6, 1.0) | SPB | 1.91 | 5.38 | 5.23 | 5.36 |
| | | CR | 5.05 | 5.23 | - | - |
| | (0.4, 0.6, 1.5) | SPB | 3.38 | 5.27 | 5.17 | 5.40 |
| | | CR | 5.09 | 5.08 | - | - |

Table 2: Simulated Type I error under stratified permuted block design (SPB) and complete randomization (CR) in % in the seamless trial with three treatments, one discrete covariate and one continuous covariate.

| MTP | ($p_1$, $p_2$, $\sigma$) | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* |
|---|---|---|---|---|---|---|
| Simes | (0.5, 0.5, 1.0) | SPB | 1.10 | 4.53 | 5.45 | 5.16 |
| | | CR | 4.47 | 4.56 | - | - |
| | (0.4, 0.6, 1.0) | SPB | 1.08 | 4.63 | 5.14 | 5.20 |
| | | CR | 4.57 | 4.60 | - | - |
| | (0.4, 0.6, 1.5) | SPB | 2.16 | 4.89 | 5.31 | 4.96 |
| | | CR | 4.55 | 4.58 | - | - |
| Dunnett | (0.5, 0.5, 1.0) | SPB | 1.23 | 4.89 | 5.78 | 5.41 |
| | | CR | 5.02 | 4.97 | - | - |
| | (0.4, 0.6, 1.0) | SPB | 1.27 | 4.94 | 5.46 | 5.19 |
| | | CR | 5.13 | 4.87 | - | - |
| | (0.4, 0.6, 1.5) | SPB | 2.31 | 5.09 | 5.66 | 5.31 |
| | | CR | 4.89 | 5.10 | - | - |

Powers for different designs and analytical approaches were compared by assuming differences in treatment effects in Table 3 (discrete case) and Table 4 (continuous case). Multiple values of $(\alpha_1, \alpha_2)$ with fixed $(p_1, p_2, \sigma) = (0.5, 0.5, 1)$ were investigated in the numerical studies. For both cases, the two sample *t*-test had smaller power under the SPB design than under CR when $|\alpha_1 - \alpha_2|$ was small, but larger power was observed when $|\alpha_1 - \alpha_2|$ increased. Under the SPB randomization, the *t*-test with adjustment was more powerful than the two sample *t*-test, regardless of the methods for multiplicity adjustment. Similar power performance was identified among the *t*-test with adjustment, the bootstrap *t*-test and the full linear model under the SPB design in the discrete case, while the *t*-test with

adjustment and the bootstrap *t*-test were slightly less powerful than the full linear model when one of the covariates was continuous. In addition, the SPB design performed better than CR regarding the number of replications that the better treatment was selected in both discrete and continuous cases, especially when $|\alpha_1 - \alpha_2|$ was relatively large.

Table 3: Power comparison (in %) and number (M) of replications the better treatment is selected for the confirmation stage under stratified permuted block design (SPB) and complete randomization (CR) in the seamless trial with three treatments and two discrete covariates.

| MTP | ($\alpha_1$, $\alpha_2$) | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* | M |
|---|---|---|---|---|---|---|---|
| Simes | (0.26, 0.16) | SPB | 77.28 | 89.44 | 88.84 | 89.73 | 6667 |
| | | CR | 75.00 | 88.93 | - | - | 6420 |
| | (0.24, 0.16) | SPB | 69.76 | 84.50 | 84.35 | 85.21 | 6374 |
| | | CR | 69.04 | 83.99 | - | - | 6139 |
| | (0.22, 0.16) | SPB | 61.76 | 78.61 | 79.02 | 79.03 | 6042 |
| | | CR | 62.46 | 77.84 | - | - | 5837 |
| | (0.20, 0.16) | SPB | 52.99 | 71.56 | 72.42 | 72.41 | 5697 |
| | | CR | 55.31 | 70.85 | - | - | 5517 |
| | (0.18, 0.16) | SPB | 44.41 | 63.68 | 64.58 | 64.63 | 5370 |
| | | CR | 48.46 | 63.50 | - | - | 5255 |
| Dunnett | (0.26, 0.16) | SPB | 78.00 | 90.08 | 89.52 | 90.13 | 6667 |
| | | CR | 75.81 | 89.57 | - | - | 6420 |
| | (0.24, 0.16) | SPB | 70.87 | 85.35 | 84.91 | 85.72 | 6374 |
| | | CR | 70.20 | 84.85 | - | - | 6139 |
| | (0.22, 0.16) | SPB | 62.71 | 79.28 | 79.56 | 79.73 | 6042 |
| | | CR | 63.59 | 78.90 | - | - | 5837 |
| | (0.20, 0.16) | SPB | 54.08 | 72.42 | 73.30 | 73.11 | 5697 |
| | | CR | 56.56 | 71.88 | - | - | 5517 |
| | (0.18, 0.16) | SPB | 45.75 | 64.60 | 65.32 | 65.13 | 5370 |
| | | CR | 49.87 | 64.71 | - | - | 5255 |

Table 4: Power comparison (in %) and number (M) of replications the better treatment is selected for the confirmation stage under stratified permuted block design (SPB) and complete randomization (CR) in the seamless trial with three treatments, one discrete covariate and one continuous covariate.

| MTP | $(\alpha_1, \alpha_2)$ | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* | M |
|---|---|---|---|---|---|---|---|
| Simes | (0.26, 0.16) | SPB | 57.32 | 88.96 | 79.31 | 79.31 | 6547 |
| | | CR | 58.74 | 88.92 | - | - | 6154 |
| | (0.24, 0.16) | SPB | 49.72 | 84.42 | 73.39 | 73.71 | 6243 |
| | | CR | 52.97 | 83.95 | - | - | 5970 |
| | (0.22, 0.16) | SPB | 42.33 | 78.44 | 67.08 | 67.36 | 5944 |
| | | CR | 47.53 | 78.06 | - | - | 5709 |
| | (0.20, 0.16) | SPB | 35.27 | 71.84 | 59.71 | 60.41 | 5632 |
| | | CR | 41.96 | 71.25 | - | - | 5495 |
| | (0.18, 0.16) | SPB | 28.76 | 64.26 | 52.83 | 53.12 | 5316 |
| | | CR | 36.72 | 63.77 | - | - | 5278 |
| Dunnett | (0.26, 0.16) | SPB | 58.49 | 89.48 | 80.17 | 80.14 | 6547 |
| | | CR | 60.41 | 89.37 | - | - | 6154 |
| | (0.24, 0.16) | SPB | 50.85 | 84.90 | 74.39 | 74.54 | 6243 |
| | | CR | 54.21 | 84.85 | - | - | 5970 |
| | (0.22, 0.16) | SPB | 43.66 | 79.41 | 68.07 | 68.04 | 5944 |
| | | CR | 49.07 | 78.85 | - | - | 5709 |
| | (0.20, 0.16) | SPB | 36.42 | 72.44 | 60.92 | 61.28 | 5632 |
| | | CR | 43.20 | 72.30 | - | - | 5495 |
| | (0.18, 0.16) | SPB | 29.70 | 65.61 | 53.77 | 53.89 | 5316 |
| | | CR | 38.15 | 64.75 | - | - | 5278 |

Similar results on Type 1 error rates, power and the number of replications the best treatment is selected for Stage II were obtained in numerical studies for Scenario 2 (Table 5-6) and Scenario 3 (Table 7-8).

## DISCUSSION

In practice, unadjusted analysis is widely used in clinical trials with CAR for simplicity and to avoid model misspecification[40,44,46]. However, ignoring the stratification covariates in the analysis may lead to a reduction in the Type 1 error rate and a decrease in power[40,44,46]. In the current numerical study, the two sample *t*-test resulted in conservative Type 1 error rates and decreased power under the SPB randomization as shown in previous findings. The newly proposed *t*-test with adjustment well controlled the Type 1 error rate under the SPB randomization with either Simes test or Dunnett test, consistent with the theoretical results in Theorem 2 and Theorem 3. The bootstrap *t*-test has been shown to control the Type 1 error rate at the nominal level under CAR in a single phase design with

Table 5: Simulated Type I error under stratified permuted block design (SPB) and complete randomization (CR) in % in the seamless trial with four treatments and three discrete covariates.

| MTP | ($p_1$, $p_2$, $p_3$, σ) | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* |
|---|---|---|---|---|---|---|
| Simes | (0.5, 0.5, 0.5, 1.0) | SPB | 0.81 | 4.44 | 5.00 | 5.19 |
| | | CR | 4.56 | 4.70 | - | - |
| | (0.4, 0.5, 0.6, 1.0) | SPB | 0.76 | 4.50 | 5.16 | 4.93 |
| | | CR | 4.57 | 4.67 | - | - |
| | (0.4, 0.5, 0.6, 1.5) | SPB | 2.05 | 4.49 | 5.22 | 4.76 |
| | | CR | 4.57 | 4.30 | - | - |
| Dunnett | (0.5, 0.5, 0.5, 1.0) | SPB | 1.03 | 5.16 | 5.58 | 5.75 |
| | | CR | 5.18 | 4.97 | - | - |
| | (0.4, 0.5, 0.6, 1.0) | SPB | 0.90 | 5.00 | 5.66 | 5.42 |
| | | CR | 5.37 | 5.03 | - | - |
| | (0.4, 0.5, 0.6, 1.5) | SPB | 2.37 | 5.15 | 5.78 | 5.24 |
| | | CR | 5.32 | 5.05 | - | - |

two treatments[45,46]. Here the Type 1 error rate was also well controlled by the bootstrap *t*-test under the SPB randomization in a seamless trial design.

Table 6: Power comparison (in %) and number (M) of replications the best treatment is selected for the confirmation stage under stratified permuted block design (SPB) and complete randomization (CR) in the seamless trial with four treatments and three discrete covariates.

| MTP | ($\alpha_1$, $\alpha_2$, $\alpha_3$) | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* | M |
|---|---|---|---|---|---|---|---|
| Simes | (0.28, 0.16, 0.14) | SPB | 67.11 | 88.17 | 88.19 | 88.45 | 6006 |
| | | CR | 66.73 | 87.44 | - | - | 5407 |
| | (0.26, 0.16, 0.14) | SPB | 59.75 | 83.52 | 83.87 | 84.36 | 5565 |
| | | CR | 60.88 | 82.76 | - | - | 5091 |
| | (0.24, 0.16, 0.14) | SPB | 52.01 | 78.28 | 78.41 | 79.47 | 5138 |
| | | CR | 55.43 | 77.50 | - | - | 4758 |
| | (0.22, 0.16, 0.14) | SPB | 44.35 | 72.23 | 72.61 | 73.42 | 4741 |
| | | CR | 49.64 | 71.35 | - | - | 4446 |
| | (0.20, 0.16, 0.14) | SPB | 36.70 | 65.57 | 65.96 | 66.77 | 4276 |
| | | CR | 44.21 | 64.43 | - | - | 4129 |
| Dunnett | (0.28, 0.16, 0.14) | SPB | 69.50 | 89.21 | 89.20 | 89.48 | 6006 |
| | | CR | 68.67 | 88.46 | - | - | 5407 |
| | (0.26, 0.16, 0.14) | SPB | 61.92 | 84.99 | 85.14 | 85.45 | 5565 |
| | | CR | 63.30 | 83.97 | - | - | 5091 |
| | (0.24, 0.16, 0.14) | SPB | 54.25 | 79.73 | 79.91 | 80.55 | 5138 |
| | | CR | 57.50 | 78.99 | - | - | 4758 |
| | (0.22, 0.16, 0.14) | SPB | 46.38 | 74.05 | 74.14 | 75.14 | 4741 |
| | | CR | 52.08 | 73.17 | - | - | 4446 |
| | (0.20, 0.16, 0.14) | SPB | 39.11 | 67.45 | 67.71 | 68.78 | 4276 |
| | | CR | 46.54 | 65.93 | - | - | 4129 |

Higher power of the *t*-test with adjustment compared to the two sample *t*-test under SPB randomization demonstrates the statistical advantage of the newly developed hypothesis testing procedure, which had similar power performance as the bootstrap *t*-test and the adjusted analysis using full linear model for the discrete case. When one stratification

covariate was continuous, however, the *t*-test with adjustment and the bootstrap *t*-test were less powerful than the full linear model, probably due to loss of information in discretizing the continuous covariate for treatment allocation while failing to fully adjust for it in the hypothesis testing process using the former two methods. In addition, the two sample *t*-test had lower power under the SPB design compared to CR due to conservativeness of the test when $|\alpha_1 - \alpha_2|$ was small. As $|\alpha_1 - \alpha_2|$ became larger, the power increased more rapidly under SPB design than under CR, resulting in higher power under SPB randomization when $|\alpha_1 - \alpha_2|$ was large. Similar results were observed from previous simulation studies under CAR in the single phase design with two treatments[45-47]. Further theoretical research is required to explain the pattern of power differences in the seamless trial setting.

Table 7: Simulated Type I error under stratified permuted block design (SPB) and complete randomization (CR) in % in the seamless trial with five treatments and two discrete covariates.

| MTP | ($p_1$, $p_2$, σ) | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* |
|---|---|---|---|---|---|---|
| Simes | (0.5, 0.5, 1.0) | SPB | 1.24 | 5.01 | 4.72 | 4.88 |
| | | CR | 4.82 | 4.34 | - | - |
| | (0.4, 0.6, 1.0) | SPB | 1.19 | 4.31 | 4.65 | 4.68 |
| | | CR | 4.59 | 4.32 | - | - |
| | (0.4, 0.6, 1.5) | SPB | 2.37 | 4.48 | 4.71 | 4.57 |
| | | CR | 4.61 | 4.62 | - | - |
| Dunnett | (0.5, 0.5, 1.0) | SPB | 1.63 | 5.10 | 5.34 | 5.32 |
| | | CR | 5.06 | 5.14 | - | - |
| | (0.4, 0.6, 1.0) | SPB | 1.53 | 5.22 | 5.42 | 5.38 |
| | | CR | 5.15 | 5.09 | - | - |
| | (0.4, 0.6, 1.5) | SPB | 2.98 | 5.05 | 5.45 | 5.08 |
| | | CR | 5.26 | 5.12 | - | - |

The probability that the best treatment is selected at the interim look to proceed is of interest for clinical trial practitioners. Here the number of replications that the best treatment is selected for Stage II was higher under the SPB design than under CR, indicating that the best treatment was more likely to be selected under CAR according to the selection rule as defined in formula (2). One intuitive explanation is that the covariates are more balanced

Table 8: Power comparison (in %) and number (M) of replications the best treatment is selected for the confirmation stage under stratified permuted block design (SPB) and complete randomization (CR) in the seamless trial with five treatments and two discrete covariates.

| MTP | $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ | Allocation | *t-test* | *lm($Z_1$, $Z_2$)* | *BS-t* | *Adjusted-t* | M |
|---|---|---|---|---|---|---|---|
| Simes | (0.28, 0.16, 0.14, 0.12) | SPB | 70.74 | 85.67 | 86.16 | 86.29 | 5142 |
| | | CR | 70.09 | 85.27 | - | - | 4666 |
| | (0.26, 0.16, 0.14, 0.12) | SPB | 63.31 | 81.11 | 81.45 | 81.77 | 4692 |
| | | CR | 64.51 | 80.38 | - | - | 4326 |
| | (0.24, 0.16, 0.14, 0.12) | SPB | 55.79 | 75.56 | 75.97 | 76.66 | 4304 |
| | | CR | 58.43 | 74.93 | - | - | 4033 |
| | (0.22, 0.16, 0.14, 0.12) | SPB | 48.50 | 69.24 | 69.82 | 70.36 | 3896 |
| | | CR | 52.21 | 68.08 | - | - | 3679 |
| | (0.20, 0.16, 0.14, 0.12) | SPB | 40.83 | 62.19 | 62.98 | 63.12 | 3547 |
| | | CR | 46.15 | 61.34 | - | - | 3350 |
| Dunnett | (0.28, 0.16, 0.14, 0.12) | SPB | 73.52 | 87.28 | 87.62 | 87.58 | 5142 |
| | | CR | 72.52 | 87.04 | - | - | 4666 |
| | (0.26, 0.16, 0.14, 0.12) | SPB | 66.30 | 82.78 | 83.19 | 83.34 | 4692 |
| | | CR | 67.20 | 82.38 | - | - | 4326 |
| | (0.24, 0.16, 0.14, 0.12) | SPB | 58.68 | 77.79 | 78.13 | 78.51 | 4304 |
| | | CR | 61.54 | 76.83 | - | - | 4033 |
| | (0.22, 0.16, 0.14, 0.12) | SPB | 51.42 | 71.72 | 71.94 | 72.54 | 3896 |
| | | CR | 55.30 | 70.82 | - | - | 3679 |
| | (0.20, 0.16, 0.14, 0.12) | SPB | 44.34 | 64.53 | 65.24 | 65.73 | 3547 |
| | | CR | 49.30 | 63.68 | - | - | 3350 |

between treatment arms under the SPB randomization, which decreases the standard error of the treatment effect estimate and thereby increases precision in the estimation. The improved precision of the estimated treatment difference can also explain the greater statistical power of the trial under CAR than under CR in the seamless design.

## CONCLUSION

In this study, numerical simulations were conducted to investigate the feasibility and advantages of a newly developed hypothesis testing approach in the seamless phase II/III design with CAR. The proposed method has been shown to well control the Type 1 error rate around the nominal level and increase the statistical power compared to the two sample *t*-test, which is not valid under CAR but still commonly used in clinical trials with such a design.

One reason that practitioners perform unadjusted analysis, such as the two sample *t*-test instead of the adjusted analysis in clinical trials with CAR is that model misspecification may occur when using linear regression models for covariate adjustment[44,46]. In practice, the underlying response model is usually unknown and the covariate effects may be not linearly additive on responses. For example, a stratification covariate may have a non-linear form or correlate with other stratification covariates. Under these scenarios, fitting a full model incorporating all stratification covariates in a linearly additive pattern in adjusted analysis will lead to biased standard errors[44]. Therefore, future numerical studies to investigate the performance of the newly developed hypothesis testing approach in different situations of model misspecification are of great interest. Another interesting future direction is to

examine the hypothesis testing procedure in a seamless design with CAR by simulations using parameters from real trial data, which could further promote its application in practice.

For simplicity, the current numerical study assumed that multiple experimental treatments were compared with one control, i.e., a placebo, and primary endpoint was evaluated at the interim look. The selection rule used at the end of Stage I was derived from a previous study, which selected treatments to continue based on standardized treatment effects[48]. In practice, multiple experimental treatments are often compared with active controls that are known, effective treatments besides placebo. Treatment selection is based on a threshold value, which is the maximum of Minimal Clinically Important Difference (MCID)[49] and treatment effects of the active controls against placebo. Experimental treatments with effects larger than the threshold value or the closest to that value if no treatment effect exceeds are selected to continue to Stage II[50]. In addition, early endpoint is evaluated at the interim analysis for treatment selection when the primary endpoint of interest is only available after long-term follow-up in practice[51]. Selection methods for incorporating early endpoint data in the seamless trials have been proposed by Stallard[52] and Friede *et al*[53]. It will be interesting to investigate the proposed hypothesis testing procedure by numerical studies that implement the above selection designs.

## REFERENCES


1. Barnes PJ, Pocock SJ, Magnussen H, et al. Integrating indacaterol dose selection in a clinical study in COPD using an adaptive seamless design. *Pulm Pharmacol Ther*. 2010;23(3):165-171.

2. Bondemark L, Ruf S. Randomized controlled trial: The gold standard or an unobtainable fallacy? *Eur J Orthod*. 2015;37(5):457-461.

3. Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: Conclusions and recommendations. *Control Clin Trials*. 1988;9(4):365-374.

4. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: Chance, not choice. *Lancet*. 2002;359(9305):515-519.

5. Kang M, Ragan BG, Park JH. Issues in outcomes research: An overview of randomization techniques for clinical trials. *J Athl Train*. 2008;43(2):215-221.

6. Lin Y, Zhu M, Su Z. The pursuit of balance: An overview of covariate-adaptive randomization techniques in clinical trials. *Contemp Clin Trials*. 2015;45(Pt A):21-25.

7. Leyland-Jones B, BEST Investigators and Study Group. Breast cancer trial with erythropoietin terminated unexpectedly. *Lancet Oncol*. 2003;4(8):459-460.

8. Hu F, Hu Y, Ma Z, Rosenberger WF. Adaptive randomization for balancing over covariates. *WIREs Comp Stat*. 2014;6:288-303.

9. Pond GR, Tang PA, Welch SA, Chen EX. Trends in the application of dynamic allocation methods in multi-arm cancer clinical trials. *Clin Trials*. 2010;7(3):227-234.

10. Taves DR. The use of minimization in clinical trials. *Contemp Clin Trials*. 2010;31(2):180-184.

11. Broglio K. Randomization in clinical trials: Permuted blocks and stratification. *JAMA*. 2018;319(21):2223-2224.

12. Therneau TM. How many stratification factors are "too many" to use in a randomization plan? *Control Clin Trials*. 1993;14(2):98-108.

13. Hallstrom A, Davis K. Imbalance in treatment assignments in stratified blocked randomization. *Control Clin Trials*. 1988;9(4):375-382.

14. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol*. 1999;52(1):19-26.

15. Umscheid CA, Margolis DJ, Grossman CE. Key concepts of clinical trials: A narrative review. *Postgrad Med*. 2011;123(5):194-204.

16. Food and Drug Admistration (FDA). The drug development process. . 2018;Retrieved from https://www.fda.gov/ForPatients/Approvals/Drugs/default.htm.

17. Bretz F, Schmidli H, Konig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biom J*. 2006;48(4):623-634.

18. Stallard N, Todd S. Seamless phase II/III designs. *Stat Methods Med Res*. 2011;20(6):623-634.

19. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50(4):1029-1041.

20. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55(4):1286-1290.

21. Dmitrienko A, D'Agostino RS. Traditional multiplicity adjustment methods in clinical trials. *Stat Med*. 2013;32(29):5172-5218.

22. Dmitrienko A, D'Agostino RB S. Multiplicity considerations in clinical trials. *N Engl J Med*. 2018;378(22):2115-2122.

23. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: The what, why, when and how. *Int J Epidemiol*. 2017;46(2):746-755.

24. Lazzeroni LC, Ray A. The cost of large numbers of hypothesis tests on power, effect size and sample size. *Mol Psychiatry*. 2012;17(1):108-114.

25. Food and Drug Administration. Multiple endpoints in clinical trials: Guidance for industry. . 2017.

26. European Medicines Agency. Guideline on multiplicity issues in clinical trials. . 2016.

27. Bretz FX,X. Tutorial: Introduction to multiplicity in clinical trials. . 2014.

28. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63(3):655-660.

29. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med*. 1999;18(14):1833-1848.

30. Dawson L. The salk polio vaccine trial of 1954: Risks, randomization and public involvement in research. *Clin Trials*. 2004;1(1):122-130.

31. Churchyard GJ, Fielding KL, Lewis JJ, et al. A trial of mass isoniazid preventive therapy for tuberculosis control. *N Engl J Med*. 2014;370(4):301-310.

32. Hanson ML, Comstock GW, Haley CE. Community isoniazid prophylaxis program in an underdeveloped area of alaska. *Public Health Rep*. 1967;82(12):1045-1056.

33. Ford ES, Ajani UA, Croft JB, et al. Explaining the decrease in U.S. deaths from coronary disease, 1980-2000. *N Engl J Med*. 2007;356(23):2388-2398.

34. U.S. National Library of Medicine. Trends, charts and maps. . 2018.

35. ASPE, U.S. Department of Health & Human Services. Examination of clinical trial costs and barriers for drug development. . 2018.

36. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*. 2016;47:20-33.

37. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: New estimates of drug development costs. *J Health Econ*. 2003;22(2):151-185.

38. Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation. . 2010.

39. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics*. 2018;January 31:kxx069.

40. Kahan BC, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: Review and reanalysis. *BMJ*. 2012;345:e5840.

41. Ma W, Hu F, Zhang L. Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*. 2015;110(510):669-680.

42. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias--two hostile brothers. *Stat Med*. 2010;29(1):1-13.

43. Hampson LV, Jennison C. Optimizing the data combination rule for seamless phase II/III clinical trials. *Stat Med*. 2015;34(1):39-58.

44. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med*. 2012;31(4):328-340.

45. SHAO J, YU X, ZHONG B. A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*. 2010;97(2):347-360.

46. Ma W, Hu F, Zhang L. Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*. 2015;110(510):669-680.

47. Forsythe AB. Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics & Data Analysis*. 1987;5(3):193-200.

48. Parsons N, Friede T, Todd S, et al. An R package for implementing simulations for seamless phase II/III clinical trials using early outcomes for treatment selection. *Computational Statistics & Data Analysis*. 2012;56(5):1150-1160.

49. Cook CE. Clinimetrics corner: The minimal clinically important change score (MCID): A necessary pretense. *J Man Manip Ther*. 2008;16(4):E82-3.

50. Bretz F, Lawrence D, Thomas P. Phase II/III adaptive design with treatment selection: A case study. Lecture presented: EMEA/EFPIA Workshop on Adaptive Designs in Confirmatory Clinical Trials; December 14, 2007; London, UK.

51. Kunz CU, Friede T, Parsons N, Todd S, Stallard N. A comparison of methods for treatment selection in seamless phase II/III clinical trials incorporating information on short-term endpoints. *J Biopharm Stat*. 2015;25(1):170-189.

52. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Stat Med*. 2010;29(9):959-971.

53. Friede T, Parsons N, Stallard N, et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Stat Med*. 2011;30(13):1528-1540.