Contemporary Clinical Trials 85 (2019) 105830

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Full Length Article

A gate-keeping test for selecting adaptive interventions under general designs of sequential multiple assignment randomized trials

Xiaobo Zhong[a,b,c,*], Bin Cheng[c], Min Qian[c], Ying Kuen Cheung[c]

[a] *Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
[b] *Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA*
[c] *Department of Biostatistics, Columbia University, New York, NY, USA*

ARTICLE INFO

Keywords:
Adaptive intervention
Gate-keeping approach
Omnibus test
Sample size calculation
Sequential multiple assignment randomized trial

ABSTRACT

This article proposes a method to overcome limitations in current methods that address multiple comparisons of adaptive interventions embedded in sequential multiple assignment randomized trial (SMART) designs. Because a SMART typically consists of numerous adaptive interventions, inferential procedures based on pairwise comparisons of all may suffer a substantial loss in power after accounting for multiplicity. Meanwhile, traditional methods for multiplicity adjustments in comparing non-adaptive interventions require prior knowledge of correlation structures, which can be difficult to postulate when analyzing SMART data of adaptive interventions. To address the multiplicity issue, we propose a likelihood-based omnibus test that compares all adaptive interventions simultaneously, and apply it as a gate-keeping test for further decision making. Specifically, we consider a selection procedure that selects the adaptive intervention with the best observed outcome only when the proposed omnibus test reaches a pre-specified significance level, so as to control false positive selection. We derive the asymptotic distribution of the test statistic on which a sample size formula is based. Our simulation study confirms that the asymptotic approximation is accurate with a moderate sample size, and shows that the proposed test outperforms existing multiple comparison procedures in terms of statistical power. The simulation results also suggest that our selection procedure achieves a high probability of selecting a superior adaptive intervention. The application of the proposed method is illustrated with a real dataset from a depression management study.

1. Introduction

An adaptive intervention (AI) is a multi-stage treatment strategy consisting of a sequence of treatment selections, one per stage of treatment, by which the selection can be adjusted repeatedly according to a patient's ongoing clinical information, such as the treatment history and responses to the previous treatments. AIs have been widely used for treating chronic diseases (e.g., depression). Sequential multiple assignment randomized trial (SMART) is a clinical trial design that randomly assigns patients to a collection of AIs. In many situations, SMART can be viewed as an early-phase developmental trial design leading trialists to the confirmatory trial [1]. Therefore, a natural clinical question in SMART is whether an AI should be selected for further investigation.

By virtue of randomization upon observing the treatment history and tailoring response, the AI values can be consistently estimated using methods such as *G*-computation estimation [2] and inverse probability weighted estimation (IPWE) [3]. Thus, an optimal AI may be selected by comparing the estimated values of all AIs, which entails multiple pairwise comparisons. In a randomized clinical trial with a primary concern to protect against false-positive findings [4], a versatile approach that can be directly applied to pairwise comparisons of AIs in a SMART is Bonferroni's adjustment, which is known to be conservative as the number of AIs increases. Meanwhile, most traditional statistical methods for adjusting multiplicity in comparing non-adaptive interventions, such as methods proposed by Tukey [5] and Hsu [6], require known correlation structures. Since the correlation between estimates derived based on SMART data is typically unknown a priori, we cannot directly apply those methods in SMART settings. Another challenge in analyzing SMART data is “curse of dimensionality”: the number of AIs in SMART typically increases dramatically as the design structure becomes more complex. Consequently, the inferential procedure may suffer a substantial loss in power after accounting for multiplicity, and the sample size calculation based on pairwise comparison may lead to a conservative design.

We propose a likelihood-based gate-keeping method to account for multiplicity whereby an AI selection will be made only after the null hypothesis of no difference among the AIs is rejected. A similar IPWE-

[*] Corresponding author at: Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY 10029, USA.
E-mail address: xiaobo.zhong@mountsinai.org (X. Zhong).

https://doi.org/10.1016/j.cct.2019.105830
Received 12 April 2019; Received in revised form 10 August 2019; Accepted 13 August 2019
Available online 27 August 2019

Table 1
MLEs and P values of pairwise comparison between each AI with the observed best AI (g = 5) in CODIACS trial caption.

		Stage-2 treatment for			
AI (g)	Stage-1 treatment	Non-response	Response	$\widehat{\theta}_g$ (se)	P-value
1	Medication	Medication	Medication	6.3 (1.1)	0.135
2	Medication	Medication	Problem-solving therapy	3.3 (1.2)	0.049
3	Medication	Problem-solving therapy	Medication	10.7 (0.6)	0.434
4	Medication	Problem-solving therapy	Problem-solving therapy	7.8 (1.1)	0.210
5	Problem-solving therapy	Medication	Medication	15.5 (6.0)	–
6	Problem-solving therapy	Medication	Problem-solving therapy	9.5 (1.0)	0.320
7	Problem-solving therapy	Problem-solving therapy	Medication	14.2 (6.1)	0.201
8	Problem-solving therapy	Problem-solving therapy	Problem-solving therapy	8.2 (1.1)	0.236

se: estimated asymptotic standard error of $\widehat{\theta}_g$.

based omnibus test was used by Ogbagaber, Karp, and Wahed [7] for sample size calculation in designing SMARTs under three specific design structures. In this article, we derive the asymptotic properties of the proposed test under very general design structure. In addition, our theoretical results leverage the fact that the variance-covariance matrix of the estimator is less than full rank, which leads to an increase in power when compared to the existing test. We will illustrate the proposed method using the CODIACS depression management trial data [8]. Briefly, in this trial, each patient was given medication or problem-solving therapy at baseline and was potentially re-assigned another treatment based on the response intermediately. The objective was to maximize the depression reduction measured by Beck Depression Inventory over a 6-month period. Table 1 lists all AIs embedded in this example, along with some analytical results. We will revisit this example with additional details in *Application*.

2. Methods

2.1. Setting, notation, and model

For brevity in exposition, we consider general SMART designs with two-stage AIs as depicted in Fig. 1(A), although the results can be readily extended to SMART with more than 2 stages. Suppose that there are I treatment options $T_1, \ldots, T_I$ at Stage 1, and under treatment T_i, there are J_i possible intermediate response categories, denoted by $R_{i1}, \ldots, R_{iJ_i}$. Next, suppose that for a patient who receives treatment T_i at Stage 1 and has an intermediate response of R_{ij}, there are K_{ij} treatment options, namely $S_{ij1}, \ldots, S_{ijK_{ij}}$ at Stage 2. Let U_z, X_z, V_z and Y_z denote the Stage-1 treatment, the intermediate outcome, the Stage-2 treatment and the primary outcome for the zth patient in a SMART, where $z = 1, \ldots, n$. Here z is the patient indicator and n is the total sample size. Let $\pi_i = \Pr(U_z = T_i)$ be the randomization probability of assigning T_i to patient z at Stage 1, and $\pi_{ijk} = \Pr(V_z = S_{ijk} | U_z = T_i, X_z = R_{ij})$ be the randomization probability of assigning S_{ijk} to patient z given the history of $(U_z = T_i, X_z = R_{ij})$. The randomization scheme of a two-stage SMART is thus completely specified by

$$\{\pi_i, \pi_{ijk}\text{: } i = 1,\ldots, I; j = 1,\ldots, J_i; k = 1,\ldots, K_{ij}\},$$

where π_i and π_{ijk} are two vectors of randomization probabilities for Stages 1 and 2. The data obtained from the zth patient who has completed a SMART can be summarized as (U_z, X_z, V_z, Y_z) and are assumed to be independent and identical with the following distributions:

$$\Pr(U_z = T_i) = \pi_i;\ i = 1,\ \ldots, I;$$

$$\Pr(X_z = R_{ij} \mid U_z = T_i) = p_{ij};\ j = 1,\ \ldots J_i;\ i = 1,\ \ldots, I;$$

$$\Pr(V_z = S_{ijk} \mid U_z = T_i, X_z = R_{ij}) = \pi_{ijk};\ k = 1,\ \ldots, K_{ij};\ j = 1,\ \ldots, J_i;\ i = 1,\ \ldots, I;$$

$$Y_z | (U_z = T_i, X_z = R_{ij}, V_z = S_{ijk}) \sim f(y_z \mid \phi_{ijk}, \tau_{ijk}),$$

where ϕ_{ijk} is the parameter of interest, and τ_{ijk}, possibly a vector, is the nuisance parameter. We assume that $f(y_z | \phi_{ijk}, \tau_{ijk})$ satisfies the regularity conditions specified in Theorem 5.39 in van der Vaart (1998) [9], which guarantee the asymptotic efficiency of the maximum likelihood estimator (MLE) of (ϕ_{ijk}, τ_{ijk}). We denote an AI by

$$d_{i;k_{i1},\ldots,k_{iJ_i}} = (T_i; S_{i1k_{i1}}, \ldots, S_{iJ_ik_{iJ_i}})$$

under which a patient receives T_i at the Stage 1, and receives $S_{ijk_{ij}}$ at the Stage 2 if the intermediate response R_{ij} is observed. The value $\theta_{i;\ k_{i1},\ \ldots,\ k_{iJi}}$ of an AI $d_{i;\ k_{i1},\ \ldots,\ k_{iJi}}$ is

$$\theta_{i;k_{i1},\ldots,k_{iJ_i}} = \sum_{j=1}^{J_i} p_{ij} \phi_{ijk_{ij}},$$

where $k_{ij} = 1, \ldots, K_{ij}$, $j = 1, \ldots, J_i$, and $i = 1, \ldots, I$. In the common situations where ϕ_{ijk} is the conditional mean of Y given a patient's clinical history in the trial, called "treatment sequence". The value of an AI can be interpreted as the marginal expected outcome Y across all the possible treatment sequences under this AI. An AI is said to be the best among all the AIs if it has the greatest value among all. We note that there could be more than one truly best AI embedded in a SMART. The goal of our method is to select the unique best AI, or a truly best AI in cases where multiple AIs have the same value.

2.2. Maximum likelihood estimation

We consider the MLE of an AI value, $\widehat{\theta}_{i;k_{i1},\ldots,k_{iJ_i}}$, obtained by plugging in the MLEs of intermediate response rates p_{ij}'s and sequence-specific means of the primary outcome $\phi_{ijk_{ij}}$'s, where $\widehat{p}_{ij}$ and $\widehat{\phi}_{ijk_{ij}}$ can be estimated by maximizing the joint distribution of (U_z, X_z, V_z, Y_z), where $z = 1, \ldots, n$.

Let $\Theta = (\theta_1, \ldots, \theta_G)^T$ be the AI values listed in lexicographical order of $\{i; k_{i1}, \ldots, k_{iJ_i}\}$, and G be the total number of AIs embedded in a SMART. Here, the term "lexicographical order" means the alphabetical order of elements with multiple indices. For example, suppose in a two-stage SMART with Stage-1 treatment options $\{T_1, T_2\}$, intermediate outcomes $\{R_{11}, R_{12}, R_{21}, R_{22}\}$, and Stage-2 treatment options $\{S_{111}, S_{112}, S_{121}, S_{211}, S_{212}, S_{221}\}$, we denote an AI value by $\theta_{i;\ k_{i1},\ k_{i2}}$ and thus can list these AIs as $(\theta_{1;\ 1,\ 1}, \theta_{1;\ 2,\ 1}, \theta_{2;\ 1,\ 1}, \theta_{2;\ 2,\ 1})$ in lexicographical order as done in a dictionary. Considering that an AI value is defined as the weighted sum of the sequence-specific means with the intermediate response rates as the weights in Section 2.1, we can obtain the joint distribution of the MLEs of the intermediate response rates and the MLEs of the sequence-specific means, and then derive the asymptotic distribution of the MLEs of AI values by using the Delta method. In fact, it can be proved that under the regularity conditions given in Theorem 5.39 in van der Vaart [9], as $n \to \infty$,

$$\sqrt{n}(\widehat{\Theta} - \Theta) \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where Σ is a block diagonal matrix whose ith block, Σ_i, is the covariance

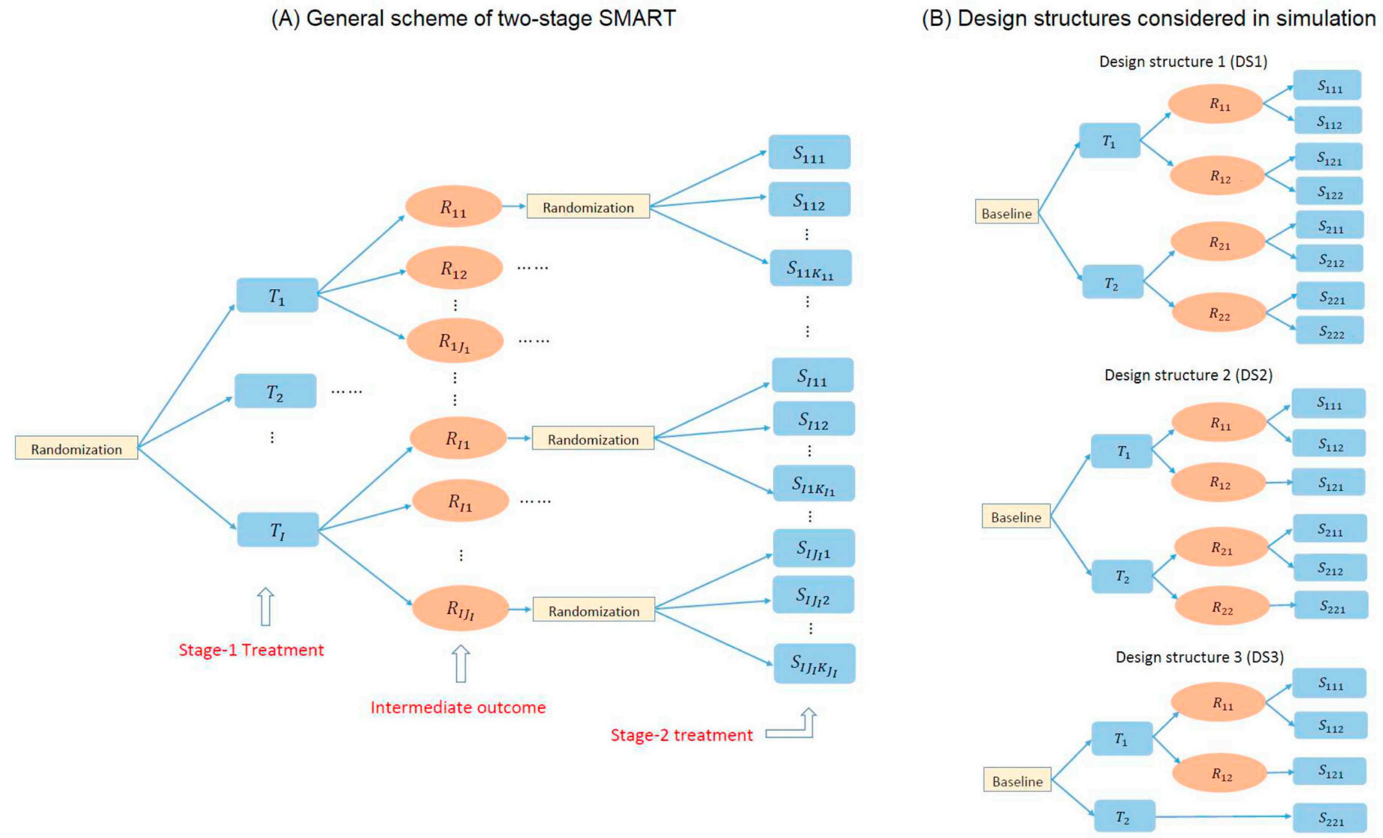

Fig. 1. (A) General scheme of a two-stage SMART design and (B) three design structures considered in simulation.

matrix of $\widehat{\Theta}_i$, the MLEs of the values of AIs sharing T_i, $i = 1, \ldots, I$. Importantly,

$$\mathrm{rank}(\Sigma_i) = \sum_{j=1}^{J_i} K_{ij} - J_i + 1.$$

The details of derivations and proofs are shown in *Appendix*.

In summary, the distribution of MLE $\widehat{\Theta}_i$ is asymptotically normal and the asymptotic covariance matrix Σ_i is not of full rank when $J_i \geq 2$. This is a surprising but fundamental distributional result on which we build the Wald test in next section. For an intuitive explanation about why the asymptotic covariance matrix is less than full rank, note that any AI value is a linear combination of the sequence-specific endpoints. In a typical SMART, the number of embedded AIs is larger than the number of treatment sequences. For example, in a SMART with two Stage-1 treatment options, binary intermediate response, and three Stage-2 treatment options given any intermediate response, *i. e.*, $\{T_i, R_{ij}, S_{ijk}; i = 1,2; j = 1,2; k = 1,2,3\}$, the total number of treatment sequences is 12, but the number of AIs is 18. Therefore, the values of AIs embedded in a SMART are linear dependent. Thus, their estimates are asymptotically linearly dependent, i.e., the asymptotic covariance is less than full rank.

2.3. Wald test and its asymptotic distributions

For ease of exposition, we use θ_g to denote the gth component of Θ, where $g = 1, \ldots, G$ and G is the total number of AIs embedded in a SMART. We consider a statistical test for the following hypotheses:

$$H_0\colon \theta_1 = \cdots = \theta_G \quad \text{versus} \quad H_1\colon \ \theta_g\text{'s are not all equal.} \tag{1}$$

Let $\mathbf{C} = (\mathbf{1}_{G-1} | -\mathbf{I}_{G-1})$ be a $(G-1) \times G$ contrast matrix such that the 1st column is a $(G-1)$ vector of 1's and the jth column is a $(G-1)$ vector in which the $(j-1)$th entry is -1 and other entries are all 0's for $2 \leq j \leq G$. For example, the contrast matrix for a SMART with 4 AIs embedded in is

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}.$$

Let Σ be the covariance matrix of $\widehat{\Theta}$, and $\widehat{\Sigma}$ be the plug-in estimator of Σ by replacing p_{ij}, ϕ_{ijk}, τ_{ijk} with their MLEs $\widehat{p}_{ij}$, $\widehat{\phi}_{ijk}$, $\widehat{\tau}_{ijk}$, respectively. Then a Wald-type test statistic can be written as

$$Q = n(\mathbf{C}\widehat{\Theta})^T(\mathbf{C}\widehat{\Sigma}\mathbf{C}^T)^-(\mathbf{C}\widehat{\Theta}), \tag{2}$$

where $\mathbf{M}^-$ denotes a generalized inverse of a square matrix $\mathbf{M}$. Under H_0 in (1) and as $n \to \infty$, Q follows a chi-squared distribution with degrees of freedom

$$\nu = \sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij} - \sum_{i=1}^{I} J_i + I - 1. \tag{3}$$

In addition, under a sequence of local alternatives $\{\Theta_n\}$ such that

$$\lim_{n\to\infty} \ n(\mathbf{C}\Theta_n)^T(\mathbf{C}\Sigma\mathbf{C}^T)^-(\mathbf{C}\Theta_n) = \lambda^* > 0,$$

Q follows a noncentral chi-squared distribution of ν degrees of freedom and noncentrality parameter λ^*. Therefore, an asymptotic level α test rejects H_0 if $Q > \chi_{\nu,\,\alpha}{}^2$, the $(1-\alpha)$th percentile of a chi-squared distribution with ν degrees of freedom. Interestingly, in the special case of comparing non-adaptive intervention sequences, $\mathrm{rank}(\Sigma_i) = 1$ and the test reduces to the regular Wald test.

2.4. Sample size determination

A formal sample size calculation formula is derived based on the proposed Wald test. In designing a SMART aiming to select the best AI to move forward to further clinical investigation, the sample size determination may proceed prescriptively as follows:

General approaches to calculate the sample size for a SMART

Step 1. For a given design structure of $\{T_i, R_{ij}, S_{ijk}\}$, where $i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $k = 1, \ldots, K_{ij}$, calculate the degrees of freedom ν according to (3).

Step 2. For a prespecified type I error rate α and a targeted statistical power, determine the noncentrality parameter λ^* required under the alternative hypothesis by solving

$$\chi^2_{\nu,\text{power}}(\lambda^*) = \chi^2_{\nu,\alpha}(0) \tag{4}$$

where $\chi_{\nu,\ \text{power}}{}^2(\lambda^*)$ denotes the $(1 - \text{power})$th percentile of a noncentral chi-squared distribution with ν degrees of freedom and the noncentrality parameter λ^*.

Step 3. For given design parameters $\{\pi_i, \pi_{ijk}\}$, assumed intermediate response probabilities $\{p_{ij}\}$ and primary outcome parameter values $\{\phi_{ijk}, \tau_{ijk}\}$ for the conditional outcome distribution f, calculate the targeted AI values Θ^* and its covariance Σ^*, so that standardized overall effect size, Δ, can be calculated according to

$$\Delta = (\mathbf{C}\Theta^*)^T(\mathbf{C}\Sigma^*\mathbf{C}^T)^-(\mathbf{C}\Theta^*). \tag{5}$$

Step 4. The total number of patients needed for a SMART is

$$n = \frac{\lambda^*}{\Delta}. \tag{6}$$

The values of λ^* under some commonly used type I error rates and statistical powers in clinical trials are given in the Table 2. Generally, smaller type I error rates, larger statistical powers, larger degrees of freedom (which reflect the number of treatment options and the number of intermediate response categories) require a larger λ^*, and hence a larger sample size per Step 4 above.

2.5. Gate-keeping approach for AI selection

We apply the proposed Wald test as a gate-keeping method: if the test fails to reject H_0 in (1), we stop further comparison and conclude that there is no sufficient evidence to support any AI being better than the others. Otherwise, if H_0 is rejected, we proceed to select the AI with the highest estimated value, and recommend it for further clinical evaluation.

The gate-keeping approach is proposed for selecting the best AIs upon rejecting the null hypothesis of no difference in a developmental trial. The idea is to screen a family of candidates under the strict control of false positive finding, so as to quickly select one AI that can potentially be more effective than the others and move it to next phase of investigation, which is a confirmatory trial to compare the selected best AI with a appropriate control.

Table 2
Non-centrality parameters (λ^*) calculated by solving Eq. (4) in Section 2.4 under degrees of freedom (ν), type I error (α), and statistical power commonly used in trials. The calculated value of λ^* is to be used in Eq. (6) in Section 2.4 for the determination of sample size.

	(α, power)					
Degrees of freedom (ν)	(0.01, 0.10)	(0.01, 0.20)	(0.05, 0.10)	(0.05, 0.20)	(0.10, 0.10)	(0.10, 0.20)
2	17.42	13.88	12.65	9.63	10.45	7.71
3	19.24	15.45	14.17	10.90	11.79	8.80
4	20.73	16.75	15.41	11.94	12.88	9.68
5	22.02	17.87	16.47	12.83	13.81	10.44
6	23.18	18.87	17.42	13.62	14.65	11.13
7	24.23	19.78	18.28	14.35	15.41	11.75
8	25.20	20.63	19.08	15.02	16.11	12.32
9	26.12	21.42	19.81	15.65	16.76	12.86
10	26.98	22.17	20.53	16.24	17.38	13.36
11	27.79	22.88	21.20	16.80	17.96	13.84
12	28.57	23.56	21.83	17.34	18.52	14.30
13	29.31	24.21	22.44	17.85	19.05	14.74
14	30.03	24.83	23.02	18.34	19.56	15.16
15	30.71	25.43	23.58	18.81	20.06	15.56
16	31.38	26.01	24.13	19.27	20.53	15.95
17	32.02	26.57	24.65	19.71	20.99	16.33
18	32.65	27.11	25.16	20.14	21.43	16.69
19	33.25	27.64	25.65	20.56	21.87	17.05
20	33.84	28.16	26.13	20.96	22.29	17.39

3. Simulations

Having established the gate-keeping approach for AI selection in the previous section, we evaluate its performances in finite sample size settings using simulation in this section. The properties of the Wald test and the gate-keeping method are examined under a variety of SMART designs and outcome scenarios.

3.1. SMART designs

Fig. 1(B) describes three design structures of two-stage SMARTs considered in the simulation. The first design structure (DS1) mimics CODIACS (cf. Table 1) and many other situations where there are two treatment options at each decision making point, that is, T_i, $S_{ijk} \in \{0, 1\}$, and binary intermediate response, that is, $R_{ij} \in \{0, 1\}$ for $i, j, k = 1, 2$. As a result, there are eight possible AIs embedded in DS1. Under DS2 and DS3, there are also two treatment options at Stage 1. However, randomization at Stage 2 may be restricted for patients with certain intermediate responses; as a result, there are 4 and 3 embedded AIs under DS2 and DS3, respectively.

With a given design structure, a SMART design will be completely specified by the set $\{\pi_i, \pi_{ijk}\}$ of randomization probabilities defined in *Methods*. In the simulation, we considered three sets of randomization probabilities for each design structure as shown in Fig. 1(B). First, we considered balanced randomization (BR), that is, $\Pr(U = 1) = 0.5$ at Stage 1 and $\Pr(V = 1 | U, X) = 0.5$ whenever there is an option of randomization at Stage 2. Second, we considered an unbalanced randomization (UBR) scheme, where $\Pr(U = 1) = 0.7$ and $\Pr(V = 1 | U, X) = 0.7$ whenever there is an option of Stage-2 randomization. Third, we considered $\Pr(U = 1) = 0.5$ at Stage 1, $\Pr(V = U | U, X = 0) = 0.3$ and $\Pr(V = U | U, X = 1) = 0.7$, whenever there is an option of second stage randomization. Under this scheme, Stage 2 implements a randomized play-the-winner (RPTW) rule for the situations where the first and the second stage treatment options are identical.

In summary, three design structures (DS1, DS2, DS3) and three randomization schemes (BR, UBR, RPTW) yield a total of 9 SMART designs under which the proposed method is evaluated.

3.2. Outcome scenarios

In a simulated SMART, the treatment assignment (U_z, V_z) of the zth patient was generated in accordance with the randomization schemes described in Section 3.1. The intermediate response rate was set as $\Pr(X_z = 1 | U_z = T_i) = 1/3$ for $T_i \in \{0, 1\}$. Given the zth patient's treatment history and intermediate response (T_i, R_{ij}, S_{ijk}), the primary outcome Y_z was randomly generated from a normal distribution with mean $\phi_{ijk} = \phi(T_i, R_{ij}, S_{ijk})$ and variance $\sigma^2 = 100$, where the conditional mean ϕ_{ijk} was specified by

$$\phi(T_i, R_{ij}, S_{ijk}) = \beta_0 + \beta_1 T_i + \beta_2 R_{ij} + \beta_3 S_{ijk} + \beta_4 T_i R_{ij} + \beta_5 T_i S_{ijk} + \beta_6 R_{ij} S_{ijk} + \beta_7 T_i R_{ij} S_{ijk} \tag{7}$$

for T_i, R_{ij}, $S_{ijk} \in \{0, 1\}$. The parameters

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^T$$

was chosen so that the true values $\theta_{i;\, k_{i1}, \ldots, k_{iJi}}$'s would follow the patterns displayed in Fig. 2. Under Value Pattern 1 (VP1), AIs with the same Stage-1 treatment had the same values; under VP2, the values of the AIs were uniformly higher if their Stage-1 treatment was $U = 1$; under VP3, the best AI had Stage-1 treatment $U = 1$ while the second best AI had Stage-1 treatment $U = 0$, and so on and so forth, following an alternating pattern. The value of $\boldsymbol{\beta}$ was chosen so that the effect size was $\Delta = 0.05$ or

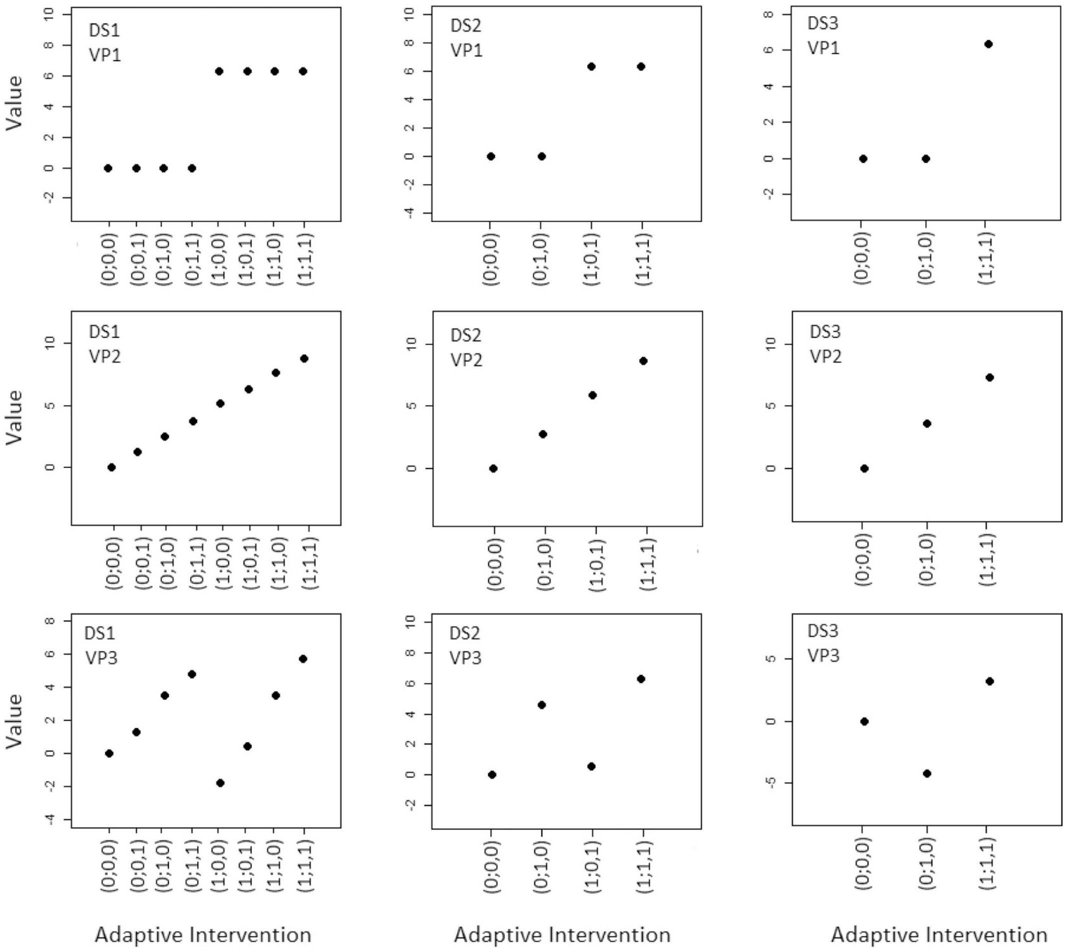

Fig. 2. Value patterns of AIs considered in the simulation.

0.10. For example, under VP1, $\beta_0 = \beta_2 = \cdots = \beta_7 = 0$ and $\beta_1 = 4.48$ and 6.33 yielded $\Delta = 0.05$ and 0.10, respectively. Table 3 provides the values of *β's* to generate all value patterns in the simulation studies. Details about how to choose $\boldsymbol{\beta}$ for each value pattern are provided in *Appendix B*.

3.3. Results

We first studied the actual type I error rate and empirical power of the proposed Wald test and compared it with a pairwise test adjusted for multiplicity. We applied this Wald test as the gate-keeping test for AI selection. The top row of Fig. 3 gives the type I error rates of the proposed Wald test at 5% nominal level under the 9 scenarios described in Section 3.1. The vertical axis is the actual type I error rate calculated as the proportion of simulated trials in which the Wald test led to the conclusion of significance among 5000 simulation replicates. This outcome scenario was generated by setting $\boldsymbol{\beta} = (0,0,0,0,0,0,0,0)^T$ in (7). Under each SMART design, the actual type I error rates of the proposed Wald test achieved the nominal level of 0.05. For comparison purposes, we also considered the pairwise testing procedures comparing AIs with Bonferroni's corrections, that is, using adjusted significance level for each individual test according to the number of comparisons under each DS (28, 6, and 3 for DS1, DS2, and DS3, respectively). Specifically, we would reject the H_0 in (1) if any pairwise test had a *P*-value less than 0.0018, 0.0083, and 0.0167 under DS1, DS2, and DS3, respectively. The simulation results indicated that the Bonferroni's correction was conservative, especially under DS1 where many comparisons were accounted for.

Rows 2–4 of Fig. 3 show the statistical powers under 3 value patterns (cf. Fig. 2) in 3 design structures (cf. Fig. 1B) given $\Delta = 0.05$. The vertical axis is the empirical power calculated by the proportion of simulated trials in which the testing procedure led to the conclusion of significance among 5000 simulation replicates. By comparing the theoretical and empirical powers of the proposed Wald test under different outcome scenarios, we verified that the asymptotic approximation discussed in Section 2.3 is accurate with a moderate sample size $n = 200$. It also displays that the proposed Wald test was generally more powerful than the Bonferroni's adjusted pairwise tests. In addition, the pairwise testing procedure had a sharp drop in power under DS1 when compared with the other design structures, likely due to the needs to adjust for many comparisons. While the Wald test also had lower power under DS1 than under DS2 and DS3, the drop was much less substantial. This demonstrated that an omnibus test was advantageous over a pairwise comparison procedure because the former attenuated the impact of a large number of AIs on the power of a SMART study.

Table 4 compares the proposed Wald test and an alternative test based on inverse probability weighted estimators (IPWE) described in Ogbagaber, Karp, and Wahed (2016) [7]. We extracted the scenarios and results of the IPWE-based omnibus test from Table I in Ogbagaber, Karp, and Wahed [7], and then calculated the corresponding sample size required by formula (6) in the *Methods* Section, and obtained the corresponding empirical power of the proposed Wald test using simulation. The proposed method generally required a smaller sample size while achieved comparable power given the IPWE-based omnibus test due to two reasons: first, our reference distribution was derived based on asymptotic theory of the MLEs; second, our method accounts for the fact that the asymptotic covariance matrix Σ of $\widehat{\Theta}$ is generally less than full rank, which has substantial impact when the design structure get more complicated.

Table 5 gives the results of selecting the best AI(s) using the proposed gate-keeping approach based on the Wald test across 9 SMART

Table 3
Values of $(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7)$ used in simulations.

		$(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7)$	
Design structure	Value pattern	Δ = 0.05	Δ = 0.10
DS1	VP1	(0, 4.48, 0, 0, 0, 0, 0, 0)	(0, 6.33, 0, 0, 0, 0, 0, 0)
DS1	VP2	(0, 3.63, 0, 2.62, 0, 0, 0, 0)	(0, 5.13, 0, 3.70, 0, 0, 0, 0)
DS1	VP3	(0, 1.86, 0, 3.73, −9.32, 1.86, −0.93, 0)	(0, 2.64, 0, 5.82, −13.20, 2.64, −1.32, 0)
DS2	VP1	(0, 4.48, 0, 0, 0, 0, 0, 0)	(0, 6.33, 0, 0, 0, 0, 0, 0)
DS2	VP2	(0, 0, 0, 2.88, 12, 0, 0, 0)	(0, 0, 0, 4.13, 17.70, 0, 0, 0)
DS2	VP3	(0, −1.21, 0, 4.82, 4.82, 1.21, 0, 0)	(0, −1.72, 0, 6.87, 6.87, 1.72, 0, 0)
DS3	VP1	(0, 4.48, 0, 0, 0, 0, 0, 0)	(0, 6.33, 0, 0, 0, 0, 0, 0)
DS3	VP2	(0, 1.29, 0, 3.88, 0, 0, 0, 0)	(0, 1.82, 0, 5.47, 0, 0, 0, 0)
DS3	VP3	(0, 0, 0, −4.46, 0, 6.69, 0, 0)	(0, 0, 0, −6.36, 0, 9.54, 0, 0)

Details of design structures (DS) and value patterns (VP) are given in Sections 3.1 and 3.2.

designs under balanced randomization with moderate sample size (n = 200). When all the true AI values embedded in a SMART are equal (Δ = 0), the probability that the gate-keeping approach leading to a significant conclusion is close to the nominal level of 0.05. Importantly, we found that the probability of each AI being selected as the best given a certain design is close to uniform. That is to say, in a situation that no AI is truly better than the others, the gate-keeping approach does not select a certain AI with higher probability than any of the others.

When the AI values under a SMART design are different, as expected, AIs with higher true values were selected more often then those with lower values, and the selection accuracy improved as the effect size Δ became larger. Interestingly, under VP1 where an AI had either a value of 0 or a positive value, we found that the probability of selecting an AI with a value of 0 was negligible, which indicated wrong selection by the gate-keeping approach after rejecting H_0 is rare. Also, the probabilities of those AI with the positive value being selected are fairly

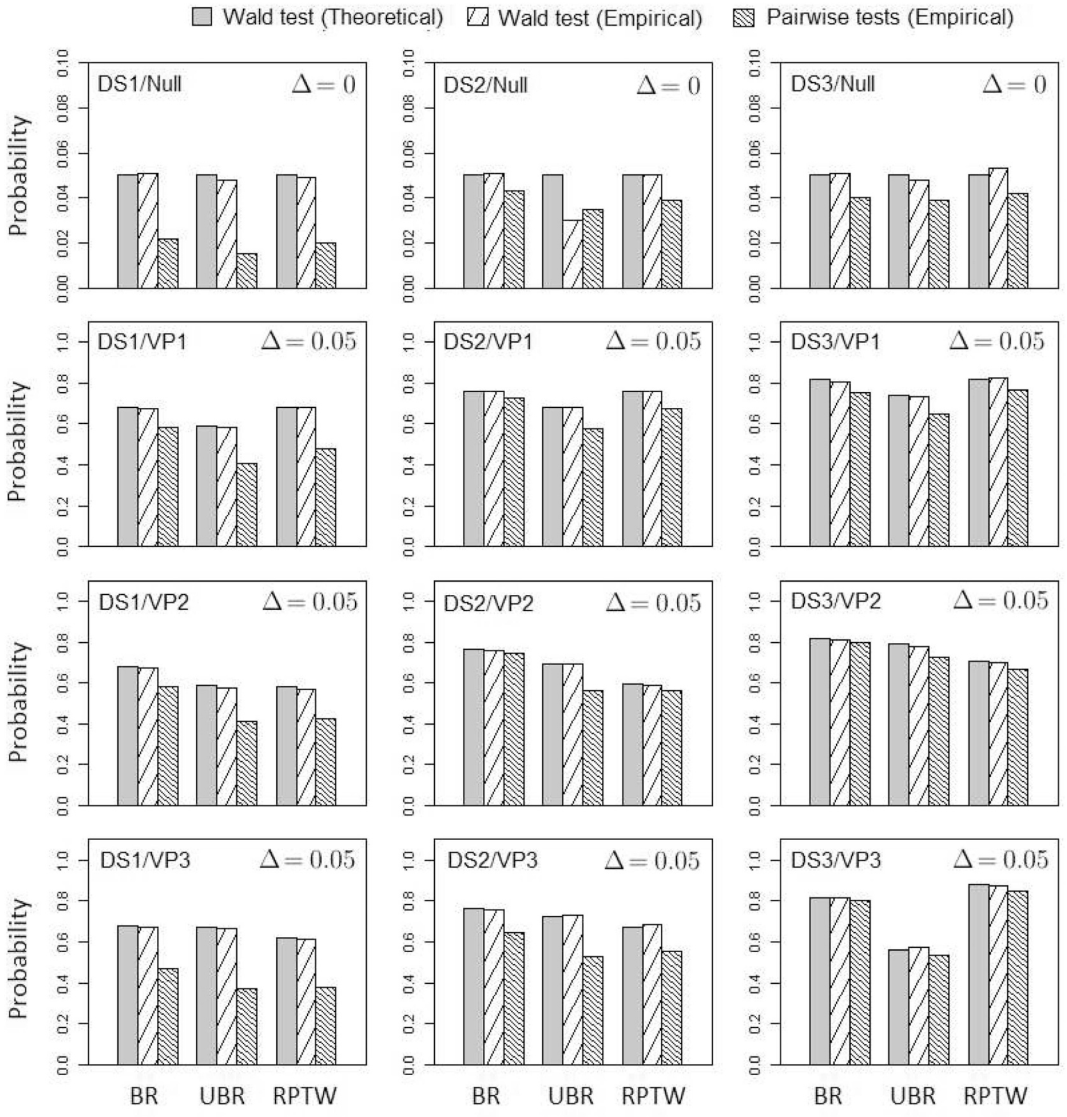

Fig. 3. Type I errors and statistical powers of proposed Wald test and pairwise tests adjusted for Bonferroni's correction.

Table 4
Comparison of the proposed Wald test versus the IPWE Wald test.

				Required sample size		Actual power	
$Pr\ (X = 1 \mid U = 0)$	$Pr\ (X = 1 \mid U = 1)$	$Pr\ (V = 1 \mid U, X = 1)$	Nominal Power	IPWE	Proposed	IPWE	Proposed
0.5	0.5	0.5	0.80	70	63	0.84	0.84
0.5	0.5	0.7	0.80	79	70	0.85	0.83
0.5	0.5	0.5	0.90	89	80	0.92	0.95
0.5	0.5	0.8	0.90	120	107	0.92	0.95
0.5	0.2	0.5	0.80	83	75	0.83	0.82
0.5	0.2	0.7	0.80	92	81	0.83	0.84
0.5	0.2	0.5	0.90	106	96	0.90	0.94
0.5	0.2	0.8	0.90	134	117	0.92	0.94
0.7	0.5	0.5	0.80	62	56	0.85	0.84
0.7	0.5	0.7	0.80	71	63	0.85	0.85
0.7	0.5	0.5	0.90	79	71	0.92	0.95
0.7	0.5	0.7	0.90	91	81	0.92	0.94
0.2	0.7	0.5	0.80	72	65	0.84	0.83
0.2	0.7	0.7	0.80	82	70	0.84	0.84
0.2	0.7	0.5	0.90	92	83	0.91	0.94
0.2	0.7	0.7	0.90	104	90	0.92	0.92

close, which indicates that when there are multiple best AIs existing, the gate-keeping approach will not make a recommendation in favor of any intervention.

4. Application

Cheung, Chakraborty and Davidson (2015) analyzed data in the CODIACS trial with an objective to determine which adaptive intervention should be selected in a depression treatment program based on the reduction of depression at 6 months post baseline [8]. The depression level was measured as Beck Depression Inventory (BDI) and the intervention leading to higher BDI reduction was regarded as more effective. Each intervention for depression management would adapt to an initial response at 8 weeks defined as no increase in BDI. Precisely, the value of an intervention in this application was the expected reduction of BDI reduction at 6 months. Table 1 gives the MLEs of all eight AI values calculated based on the data of 108 patients, along with the standard errors.

Suppose the goal of our study is to select the best AI for depression management to be assessed in a confirmation trial in which the selected AI will be compared with the standard care. We applied the proposed two-stage gate-keeping approach for this selection. At Step 1, we conducted the Wald test under Hypothesis (1) with total number of AIs $G = 8$ and obtained the test statistics $Q = 36.0$ per Formula (2). As the null distribution of the test statistic was a chi-squared distribution with 5 degrees of

Table 5
The distribution of selected AI by the gate-keeping method after the Wald test (at 5% level) under balanced randomization and a total sample size of $n = 200$.

	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS1-Null		DS1-VP1		DS1-VP2		DS1-VP3		DS1-VP1		DS1-VP2		DS1-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,0)	0.00	0.006	0.00	0.000	0.00	0.000	0.00	0.001	0.00	0.000	0.00	0.000	0.00	0.000
(0;0,1)	0.00	0.006	0.00	0.000	0.87	0.000	0.93	0.005	0.00	0.000	1.23	0.000	1.32	0.001
(0;1,0)	0.00	0.006	0.00	0.000	1.75	0.001	2.49	0.045	0.00	0.000	2.47	0.000	3.52	0.033
(0;1,1)	0.00	0.006	0.00	0.000	2.62	0.007	3.42	0.209	0.00	0.000	3.70	0.001	4.84	0.300
(1;0,0)	0.00	0.006	4.48	0.169	3.63	0.018	−1.24	0.000	6.33	0.241	5.13	0.009	−1.76	0.000
(1;0,1)	0.00	0.006	4.48	0.168	4.50	0.068	0.31	0.000	6.33	0.251	6.36	0.053	0.44	0.000
(1;1,0)	0.00	0.006	4.48	0.166	5.38	0.119	2.48	0.027	6.33	0.231	7.60	0.120	3.52	0.014
(1;1,1)	0.00	0.006	4.48	0.167	6.25	0.458	4.04	0.385	6.33	0.228	8.83	0.759	5.72	0.598

	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS2-Null		DS2-VP1		DS2-VP2		DS2-VP3		DS2-VP1		DS2-VP2		DS2-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,1)	0.00	0.012	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000
(0;1,1)	0.00	0.012	0.00	0.000	1.92	0.003	3.21	0.181	0.00	0.000	2.77	0.000	4.59	0.170
(1;0,1)	0.00	0.013	4.48	0.380	4.00	0.076	0.40	0.000	6.33	0.488	5.90	0.042	0.57	0.000
(1;1,1)	0.00	0.013	4.48	0.381	5.92	0.678	4.42	0.576	6.33	0.489	8.67	0.932	6.31	0.806

	$\Delta = 0.00$		$\Delta = 0.05$						$\Delta = 0.10$					
	DS3-Null		DS3-VP1		DS3-VP2		DS3-VP3		DS3-VP1		DS3-VP2		DS3-VP3	
AI	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.	Value	Prob.
(0;0,1)	0.00	0.0016	0.00	0.000	0.00	0.000	0.00	0.057	0.00	0.000	0.00	0.000	0.00	0.023
(0;1,1)	0.00	0.0016	0.00	0.001	2.59	0.034	−2.96	0.000	0.00	0.000	3.65	0.012	−4.24	0.000
(1;1,1)	0.00	0.0019	4.48	0.807	5.17	0.776	2.22	0.761	6.33	0.985	7.30	0.972	3.18	0.962

freedom according to Formula (3), we got $P < 0.001$ and concluded that the values of AIs were significantly different at 5% level. Consequently, we continued to Step 2 and selected the AI that started with problem-solving therapy followed by medication (AI with $g = 5$) as the recommendation for further investigation. Alternatively, we could use pairwise testing approach to compare each AI against the observed best intervention (AI with $g = 5$): one comparison had a P value less than 0.05 (AI with $g = 2$ and $P = 0.049$); however, multiplicity adjustment would require a P value less than 0.0018 according to Bonferroni's method. Thus, the pairwise testing approach would have failed to declare the overall difference among the AIs and no AI would be selected for further study.

5. Discussions

We have proposed an omnibus test for comparing several AIs in a SMART, and derived a sample size determination procedure based on the test. Traditional hypothesis tests of comparing AIs embedded in a SMART focused on pairwise comparisons of AIs. Murphy proposed a hypothesis test for comparing two non-overlapping AIs in a SMART [1]. In this comparison, data collected from patients followed two AIs are statistically independent, and thus the sample size can be determined in a similar fashion to a two-sample *t*-test. Oetting and colleagues [10] further discussed Murphy's sample size formula and proposed an algorithm to select the best AI embedded in a SMART, assuming the best two AIs can be correctly identified at the design stage. Dawson and Lavori [11] provided a sample size formula based on the nested structure estimation for SMART with continuous outcomes. The covariance between two AIs was obtained by combining the stage-specific variance inflation factor (VIF) and sequence-specific variance. To deal with two AIs that are overlapped in the early stages, Dawson and Lavori further proposed a conservative approach to estimate the VIF by stage-specific regression [12]. Hypothesis tests for comparing two AIs with survival outcomes were studied in Feng and Wahed [13], and Li and Murphy [14].

In addition, we have explored applying the proposed test as a gate-keeping method for AI selection. While the literature of ranking and selection procedures can be traced back to 1970's [15–17], these methods have not been widely used in clinical trials, where selection of treatment may occur only when there is an adequate level of statistical evidence concerning the differences among the treatments. A gate-keeping test is a common approach to evaluate the extent of statistical evidence. For example, Dunnett [18] proposed a gate-keeping procedure in the presence of a control. Using this method, when and only when the null hypothesis of the ANOVA F test is rejected, pairs of difference will be identified by comparing each treatment with the control group. More recently, gate-keeping procedures have been proposed to handle multiple hierarchical objectives in clinical trials comparing non-adaptive interventions. Westfall and Krishen [19] proposed a serial gate-keeping procedure, where a family of null hypotheses defines a serial gate-keeper, that is, it requires rejecting all the null hypotheses before making further tests. Demitrienko, Offen, and Westfall [20] proposed a parallel gate-keeping paradigm whereby a family of null hypothesis defines parallel gate-keepers, so that rejecting at least one null will suffice for further comparisons. These two gate-keeping approaches was later unified into a general tree gate-keeping procedure [21]. As far as the authors are aware, this is the first proposal of using a gate-keeping procedure in SMART. It would be interesting to further pursue applying the above-mentioned hierarchical gate-keeping methods in SMART. Numerous non-gate-keeping methods for treatment selections have also been suggested. Whitehead proposed a Bayesian selection trial design, by which several experimental treatments are first evaluated, and the most promising one will then be compared to a standard treatment [22]. Thall and colleagues proposed a two-stage procedure to identifying the best of the experimental interventions and determining whether it is superior to a control with an objective to minimize the expected total sample size under the null [23]. Cheung [24] proposed a class of sequential selection boundaries for multi-armed clinical trial for selecting a treatment in comparison with a control. A contribution of this article is the extension of the selection paradigm concept to the context of adaptive interventions in a SMART. As a result of the selection paradigm, one can substantially reduce the sample size of a SMART by powering the study based on a gate-keeping test, and thus improving the feasibility of conducting a SMART. The simulation study shows that the power of the proposed omnibus test is affected by the number of embedded AIs to a lesser extent than pairwise comparison with multiplicity adjustments. As the “curse of dimensionality” is a major concern in evaluating AIs embedded in a SMART, especially if we consider more than two stages and multiple response categories, performing such an omnibus test as a gate-keeping test is a reasonable approach in light of feasibility.

From a practical viewpoint, the proposed method facilitates clear clinical decisions at the end of a trial. Specifically, in this article, we consider an approach whereby an AI is selected upon rejecting the null of no difference. We note that the goal of a selection trial is not to select the best intervention with high probability, but rather select a superior intervention in that it is not a “bad” one [25]: The two objectives coincide in scenarios where no AI falls in the “indifference zone”. Indifference zone is a notion developed in the ranking and selection literature, and generally refers to a region of the parameter space where selection properties are not explicitly calibrated due to insufficient separation of the parameters (values) of interest; and therefore, the inferential procedure is indifferent to the selection decision. For example, under VP1 when the effect size Δ is 0.10, where there are two possible AI values (0 vs. 6.33), it is apparent that AIs having a value of 6.33 is the best, and there is no ambiguity as to what would constitute a correct selection (assuming a difference of 6.33 is of clinical significance). In contrast, in VP2 or VP3, the best AI is separated by the second-best AI by a small difference (VP2 8.83 vs. 7.60; VP3 5.62 vs. 4.84). It may not be of sufficient practical relevance to power the study to differentiate the two at the cost of a large sample size, and it is not an incorrect decision to select the second-best AI, although it has a slightly smaller value than the best AI; that is, the second-best AI falls in the indifference zone. See further discussion and examples of indifference zone in Bechhofer, Santner, and Goldsman [26] and Cheung [27].

The proposed omnibus test can be coupled with other clinical decision rules such as identifying inferior interventions, as long as these rules are pre-specified. In this article, we focus on selecting the best AI after the null hypothesis of the gate-keeping test is rejected. Depending on the trial's objective, the omnibus test can also be used in conjunction with AI elimination instead of selection. In this case, we would be interested in keeping the probability of correctly elimination of an inferior AI with high probability. To illustrate, the re-analysis of CODIACS (cf. Table 1) shows that several AIs ($g = 3, 7$) had estimated values close to the observed best ($g = 5$), whereas some were clearly inferior to these promising AIs. In order to identify AIs for elimination, one might perform unadjusted pairwise tests against the observed best—when the null hypothesis of the gate-keeping test in (1) is rejected: interventions that had significantly different values than the observed best would be eliminated. In the CODIACS re-analysis, the AI with $g = 2$ would have been declared inferior according to this procedure.

We have explored the distributional results of MLEs for the AI values under general SMART designs. Interestingly, we noted that the limiting covariance of the MLE is less than full rank, which we believe is true also for other estimators (e.g., IPWE), because each AI is a linear combination of potentially overlapping treatment sequences embedded in a SMART, as explained in Section 2.2. This is a key result that allows us to establish an efficient omnibus test with a null reference distribution with the degrees of freedom $\nu < G - 1$ and propose the advanced sample size calculation method for designing a SMART. Without this result, one might naturally conjecture a null distribution with $G - 1$ degrees of freedom, which would lead to a conservative test.

A potential limitation of using the MLE is that it will require the full specification of the model, and it may be perceived as restrictive in the application. We, however, note that under normality, the MLE is asymptotically identical to the IPWE, which suggests a certain degree of robustness of the MLE, at least for continuous outcomes. Furthermore, the proposed gate-keeping procedure is not tied to MLE. Ogbagaber, Karp, and Wahed (2016), for example, construct an omnibus test based on IPWE [7]. As long as we can obtain a consistent estimator for the AI value and the asymptotic variance-covariance of these estimators, we will be able to apply the gate-keeping method. These are certain topics for further study. Having said that, we note that the results in this article are derived under rather general conditions on the distribution of final primary outcome, with the exponential family being the most prominent example that the theory is applicable to. While we have focused on evaluating the proposed method with continuous outcome, simulation studies using binary outcome data (not reported here) show similar performance. Thus, the specific procedure studied in this article shall have applications in very broad settings.

Funding

The work was partially supported by the NIH R01 grant MH109496 and the NIH R21 grant MH108999.

Acknowledgment

We thank the Associate Editor and two referees for their comments on this article.

Appendix

Appendix A

Proof of Asymptotic Distribution of $\widehat{\Theta}$

The log-likelihood function based on $\{(U_z, X_z, V_z, Y_z); z = 1, \ldots, n\}$ is

$$\log L(p_{ij}, \phi_{ijk}, \tau_{ijk}) = \sum_{z=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} I(U_z = T_i, X_z = R_{ij}, V_z = S_{ijk}) \log f(y_z \mid \phi_{ijk}, \tau_{ijk}) + \sum_{z=1}^{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} I(U_z = T_i, X_z = R_{ij}, V_z = S_{ijk}) \log p_{ij} + \text{constant}.$$

The MLEs can be derived by solving the score equations based on the first and second derivatives of the log-likelihood function. Specifically, the MLE for p_{ij} is

$$\widehat{p}_{ij} = \frac{\sum_{z=1}^{n}\sum_{k=1}^{K_{ij}} I(U_z = T_i, X_z = R_{ij}, V_z = S_{ijk})}{\sum_{z=1}^{n}\sum_{j'=1}^{J_i}\sum_{k=1}^{K_{ij'}} I(U_z = T_i, X_z = R_{ij'}, V_z = S_{ij'k})},$$

whereas the MLEs for ϕ_{ijk} and τ_{ijk}, denoted as $\widehat{\phi}_{ijk}$ and $\widehat{\tau}_{ijk}$ respectively, generally have no closed form expression.

For an AI indexed by i, its value $\theta_{i;\, k_{i1},\, \ldots,\, k_{iJi}}$ is determined by

$$\mathbf{p}_i = \begin{pmatrix} p_{i1} \\ \vdots \\ p_{iJ_i} \end{pmatrix}, \quad \boldsymbol{\phi}_{ij} = \begin{pmatrix} \phi_{ij1} \\ \vdots \\ \phi_{ijK_{ij}} \end{pmatrix}, \quad \boldsymbol{\phi}_i = \begin{pmatrix} \boldsymbol{\phi}_{i1} \\ \vdots \\ \boldsymbol{\phi}_{iJ_i} \end{pmatrix},$$

for which we use $\widehat{\mathbf{p}}_i$, $\widehat{\boldsymbol{\phi}}_i$ to denote their MLEs. Then, let $\boldsymbol{\theta}_i$ be the vector of $\theta_{i;\, k_{i1},\, \ldots,\, k_{iJi}}$'s, arranged in the lexicographical order in $(k_{i1}, \ldots, k_{iJ_i})$. We express $\boldsymbol{\theta}_i$ in two equivalent forms as

$$\theta_i = \mathbf{A}_i\Lambda_i(\mathbf{p}_i)\boldsymbol{\phi}_i = \mathbf{A}_i\Gamma_i(\boldsymbol{\phi}_i)\mathbf{p}_i, \tag{A.1}$$

where

$$\mathbf{A}_i = (\mathbf{I}_{K_{i1}} \otimes \mathbf{1}_{K_{i2}} \otimes \cdots \otimes \mathbf{1}_{K_{iJ_i}} \mid \mathbf{1}_{K_{i1}} \otimes \mathbf{I}_{K_{i2}} \otimes \cdots \otimes \mathbf{1}_{K_{iJ_i}} \mid \cdots \mid \mathbf{1}_{K_{i1}} \otimes \cdots \otimes \mathbf{1}_{K_{i(J_i-1)}} \otimes \mathbf{I}_{K_{iJ_i}}) \tag{A.2}$$

is a $G_i \times m_i$ matrix with ⊗ denoting the Kronecker product, $G_i = \Pi_{j=1}{}^{J_i}K_{ij}$, and $m_i = \Sigma_{j=1}{}^{J_i}K_{ij}$; also, $\mathbf{I}_k$ denotes the $k \times k$ identity matrix, $\mathbf{1}_k$ the $k \times 1$ matrix of 1's, and $\Lambda_i(\mathbf{p}_i) = \text{bdiag}\{p_{ij}\mathbf{I}_{K_{ij}}; j = 1, \ldots, J_i\}$ is an $m_i \times m_i$ block diagonal matrix and $\Gamma_i(\boldsymbol{\phi}) = \text{bdiag}\{\boldsymbol{\phi}_{ij}; j = 1, \ldots, J_i\}$ is a $G_i \times J_i$ block diagonal matrix with "bdiag{·}" denoting a block diagonal matrix.

The two expressions of MLE of $\boldsymbol{\theta}_i$ in (A.1) can be respectively expressed as

$$\widehat{\theta}_i = \mathbf{A}_i\Lambda(\widehat{\mathbf{p}}_i)\widehat{\boldsymbol{\phi}}_i = \mathbf{A}_i\Gamma_i(\widehat{\boldsymbol{\phi}}_i)\widehat{\mathbf{p}}_i.$$

Now define $\Sigma_{\mathbf{p}_i} = \pi_i^{-1}(\text{diag}\{\mathbf{p}_i\} - \mathbf{p}_i\mathbf{p}_i^T)$, $\Sigma_{\boldsymbol{\phi}_i} = \text{bdiag}\left\{\Sigma_{\phi_{i1}}, \ldots, \Sigma_{\phi_{iJ_i}}\right\}$,

$\Sigma_{\boldsymbol{\phi}_{ij}} = (\pi_i p_{ij} \pi_{ijk})^{-1}\text{bdiag}\{\sigma^2(\boldsymbol{\phi}_{ijk}, \boldsymbol{\tau}_{ijk}); k = 1, \ldots, K_{ij}\}$,

and $\sigma^2(\phi_{ijk}, \tau_{ijk}) = (i_{\phi_{ijk}\phi_{ijk}} - i_{\phi_{ijk}\tau_{ijk}}{}^T \cdot i_{\tau_{ijk}\tau_{ijk}}{}^{-1} \cdot i_{\phi_{ijk}\tau_{ijk}})^{-1}$, where

$$\begin{pmatrix} i_{\phi_{ijk}\phi_{ijk}} & i_{\phi_{ijk}\tau_{ijk}}^T \\ i_{\phi_{ijk}\tau_{ijk}} & i_{\tau_{ijk}\tau_{ijk}} \end{pmatrix}$$

is the block Fisher's information matrix of distribution $f(y \mid \phi_{ijk}, \tau_{ijk})$.

Let $\Sigma = \text{bdiag}\{\Sigma_1, \ldots, \Sigma_I\}$, where $\Sigma_i = \mathbf{A}_i(\Gamma_i(\boldsymbol{\phi}_i)\Sigma_{\mathbf{p}_i}\Gamma_i(\boldsymbol{\phi}_i)^T + \Lambda_i(\mathbf{p}_i)\Sigma_{\boldsymbol{\phi}_i}\Lambda_i(\mathbf{p}_i))\mathbf{A}_i^T$. Assume that $f(y_z \mid \phi_{ijk}, \tau_{ijk})$ satisfies the regularity conditions as specified in Theorem 5.39 of van der Vaart (1998) [9]. We first prove the asymptotic distribution of $\widehat{\Theta}$ in Section 2.2. Noticing that under the standard regularity conditions, we have

$$\sqrt{n}\begin{pmatrix} \widehat{\mathbf{p}}_i - \mathbf{p}_i \\ \widehat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{p}_i} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\boldsymbol{\phi}_i} \end{pmatrix}\right),$$

where $\Sigma_{\mathbf{p}_i}$ and $\Sigma_{\boldsymbol{\phi}_i}$ are given above. By the delta-method and using the two equivalent expressions of $\boldsymbol{\theta}_i$ in (A.1),

$$\begin{aligned}&\sqrt{n}\,(\widehat{\theta}_i - \theta_i)\\ &= (\mathbf{A}_i\Gamma_i(\boldsymbol{\phi}_i) \mid \mathbf{A}_i\Lambda(\mathbf{p}_i))\sqrt{n}\begin{pmatrix}\widehat{\mathbf{p}}_i - \mathbf{p}_i\\ \widehat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i\end{pmatrix} + \mathbf{O}_p(1)\end{aligned}$$

$$\xrightarrow{d} N\,(\mathbf{0}, \mathbf{A}_i(\Gamma_i(\boldsymbol{\phi}_i)\Sigma_{\mathbf{p}_i}\Gamma_i(\boldsymbol{\phi}_i)^T + \Lambda_i(\mathbf{p}_i)\Sigma_{\boldsymbol{\phi}}\Lambda_i(\mathbf{p}_i))\mathbf{A}_i^T) = N\,(\mathbf{0}, \Sigma_{\theta_i}).$$

We establish two lemmas:

Lemma 1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $k \times k$ *real symmetric matrices. Assume* $\mathbf{A}$ *is positive definite and* $\mathbf{B}$ *is positive semi-definite. Then,*

$$\text{rank}(\mathbf{A} + \mathbf{B}) = \text{rank}(\mathbf{A}) = k.$$

Proof: For the positive semi-definite matrix $\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$, there exist an orthogonal matrix $\mathbf{C}$ such that

$$\mathbf{C}\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)\mathbf{C}^T = \text{diag}\{\lambda_1, \cdots, \lambda_k\},$$

where $\lambda_i \geq 0$, $i = 1, \ldots, k$, are the eigenvalues of $\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$. Therefore,

$$\begin{aligned}\text{rank}(\mathbf{A} + \mathbf{B}) &= \text{rank}\left(\mathbf{I}_k + \mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)\\ &= \text{rank}\left(\mathbf{I}_k + \mathbf{C}\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\mathbf{C}^T\right)\\ &= \text{rank}(\text{diag}\{1 + \lambda_1, \ldots, 1 + \lambda_k\}) = k.\end{aligned}$$

Lemma 2. *Let* $\mathbf{A}_i$ *be defined as in Eq.* (A.2). *Then,*

$$\text{rank}(\mathbf{A}_i) = \sum_{j=1}^{J_i} K_{ij} - J_i + 1 = m_i - J_i + 1.$$

Proof: We apply the principle of mathematical induction to J_i. For such purpose, we write $\mathbf{A}_i$ as $\mathbf{A}_i(K_{i1}, \ldots, K_{iJ_i})$. If $J_i = 2$, then

$$\mathbf{A}_i(K_{i1}, K_{i2}) = (\mathbf{I}_{K_{i1}} \otimes \mathbf{1}_{K_{i1}} \mid \mathbf{1}_{K_{i1}} \otimes \mathbf{I}_{K_{i2}}),$$

which, after some elementary operations for block matrices, becomes

$$\begin{pmatrix}\mathbf{0} & \mathbf{0} & \mathbf{I}_{K_{i2}}\\ \mathbf{0} & \mathbf{I}_{K_{i1}-1} \otimes \mathbf{1}_{K_{i2}} & \mathbf{0}\end{pmatrix}.$$

Thus

$$\text{rank}(\mathbf{A}_i(K_{i1}, K_{i2})) = \text{rank}(\mathbf{I}_{K_{i2}}) + \text{rank}(\mathbf{I}_{K_{i1}-1} \otimes \mathbf{1}_{K_{i2}}) = K_{i2} + K_{i1} - 1.$$

Suppose the conclusion holds for J_i, consider now the case of $J_i + 1$. Denote

$$K_i' = \sum_{j=2}^{J_i+1} K_{ij}.$$

Since after some elementary operations for block matrices

$$\mathbf{A}_i(K_{i1}, \ldots, K_{iJ_i}, K_{i,J_i+1}) = (\mathbf{I}_{K_{i1}} \otimes \mathbf{1}_{K_i'} \mid \mathbf{1}_{K_{i1}} \otimes \mathbf{A}_i(K_{i1}, \ldots, K_{i,J_i+1}))$$

becomes

$$\begin{pmatrix}\mathbf{0} & \mathbf{0} & \mathbf{A}_i(K_{i2}, \ldots, K_{i,J_i+1})\\ \mathbf{0} & \mathbf{I}_{K_{i1}-1} \otimes \mathbf{1}_{K_i'} & \mathbf{0}\end{pmatrix},$$

we have

$$\begin{aligned}\text{rank}(\mathbf{A}_i(K_{i1}, \ldots, K_{i,J_i+1})) &= \text{rank}(\mathbf{A}_i(K_{i2}, \ldots, K_{i,J_i+1})) + \text{rank}(I_{K_{i1}-1} \otimes \mathbf{1}_{K_i'})\\ &= K_i' - J_i + 1 + K_{i1} - 1\\ &= \sum_{j=1}^{J_i+1} K_{ij} - (J_i + 1) + 1,\end{aligned}$$

which proves the claim for $J_i + 1$.

By Lemma 1,

$$\text{rank}(\Lambda_i(\mathbf{p}_i)\Sigma_{\boldsymbol{\phi}_i}\Lambda_i(\mathbf{p}_i) + \Gamma_i(\boldsymbol{\phi}_i)\Sigma_{\mathbf{p}_i}\Gamma_i(\boldsymbol{\phi}_i)) = m_i,$$

hence of full rank. Then, by Lemma 2,

$$\begin{aligned}\mathrm{rank}(\Sigma_{\theta_i}) &= \mathrm{rank}(\mathbf{A}_i(\Lambda_i(\mathbf{p}_i)\Sigma_{\boldsymbol{\phi}_i}\Lambda_i(\mathbf{p}_i) + \Gamma_i(\boldsymbol{\phi}_i)\Sigma_{\mathbf{p}_i}\Gamma_i(\boldsymbol{\phi}_i))\mathbf{A}_i^T)\\ &= \mathrm{rank}(\mathbf{A}_i) = \sum_{j=1}^{J_i} K_{ij} - J_i + 1.\end{aligned}$$

Proof of Degrees of Freedom ν

We have proved that $Q \overset{d}{\to} \chi_\nu^2$ under the null hypothesis of (1). By a contiguity argument, under the local alternatives $\{\Theta_n\}$ which satisfies

$$\lim_{n\to\infty} \ n(\mathbf{C}\Theta_n)^T(\mathbf{C}\Sigma\mathbf{C}^T)^-(\mathbf{C}\Theta_n) = \lambda^* > 0,$$

$Q \overset{d}{\to} \chi_\nu^2(\lambda^*)$. We now verify that the degrees of freedom formula (3). Let $G = \Sigma_{i=1}{}^I G_i$ and $m = \Sigma_{i=1}{}^I m_i$. Define an $G \times m$ matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathrm{bdiag}\{\mathbf{A}_i; i = 1,\ldots, I\}.$$

Without loss of generality, consider an $(G-1) \times G$ contrast matrix

$$\mathbf{C} = (\mathbf{1}_{G-1} \,|{-I_{G-1}}).$$

By subtracting the first row from the remaining $(G-1)$ rows in $\mathbf{A}$, and then subtracting the first column from the remaining columns (all of these are elementary operations), $\mathbf{A}$ is converted to

$$\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{0} & \mathbf{B}\end{pmatrix},$$

and check that $(\mathbf{0} \mid \mathbf{B}) = \mathbf{CA}$ holds. Then,

$$\mathrm{rank}(\mathbf{A}) = 1 + \mathrm{rank}(\mathbf{B}) = 1 + \mathrm{rank}(\mathbf{CA}).$$

Therefore, the degrees of freedom of ${\chi_\nu}^2$ test is

$$\nu = \mathrm{rank}(\mathbf{CA}) = \mathrm{rank}(\mathbf{A}) - 1 = \sum_{i=1}^{I}\sum_{j=1}^{J_i} K_{ij} - \sum_{i=1}^{I} J_i + I - 1.$$

Appendix B

Specification of ϕ_{ijk}'s in the Simulation

We provide an example of generating the sequence-specific mean outcome ϕ_{ijk} in the simulations under design structure 1 (DS1) and balanced randomization scheme (BR) with value pattern 1 (VP1). There are 8 possible treatment sequences in this setting and the sequence-specific means ϕ_{ijk}'s can be expressed as a set of linear functions of $\boldsymbol{\beta}$ as follows,

$$\begin{aligned}\phi_{111} &= \beta_0\\ \phi_{112} &= \beta_0 + \beta_3\\ \phi_{121} &= \beta_0 + \beta_2\\ \phi_{122} &= \beta_0 + \beta_2 + \beta_3 + \beta_6\\ \phi_{211} &= \beta_0 + \beta_1\\ \phi_{212} &= \beta_0 + \beta_1 + \beta_3 + \beta_5\\ \phi_{221} &= \beta_0 + \beta_1 + \beta_2 + \beta_4\\ \phi_{222} &= \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7.\end{aligned}$$

The value of an AI in this setting is

$$\theta_{i;k_{i1},k_{i2}} = p_{i1}\phi_{i1k_{i1}} + p_{i2}\phi_{i2k_{i2}},$$

where $k_{i1} \in \{0,1\}$ and $k_{i2} \in \{0,1\}$ for $i = 1, 2$. Thus, the targeted AI values are

$$\Theta_{\boldsymbol{\beta}}^* = \begin{pmatrix}\theta_{1;1,1}\\ \theta_{1;1,2}\\ \theta_{1;2,1}\\ \theta_{1;2,2}\\ \theta_{2;1,1}\\ \theta_{2;1,2}\\ \theta_{2;2,1}\\ \theta_{2;2,2}\end{pmatrix} = \begin{pmatrix}\beta_0 + \frac{1}{3}\beta_2\\ \beta_0 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 + \frac{1}{3}\beta_6\\ \beta_0 + \frac{1}{3}\beta_2 + \frac{2}{3}\beta_3\\ \beta_0 + \frac{1}{3}\beta_2 + \beta_3 + \frac{1}{3}\beta_6\\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_4\\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 + \frac{1}{3}\beta_4 + \frac{1}{3}\beta_5 + \frac{1}{3}\beta_6 + \frac{1}{3}\beta_7\\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \frac{2}{3}\beta_3 + \frac{1}{3}\beta_4 + \frac{2}{3}\beta_5\\ \beta_0 + \beta_1 + \frac{1}{3}\beta_2 + \beta_3 + \frac{1}{3}\beta_4 + \beta_5 + \frac{1}{3}\beta_6 + \frac{1}{3}\beta_7\end{pmatrix}.$$

We add the subscript $\boldsymbol{\beta}$ to $\Theta^*_{\boldsymbol{\beta}}$ in the above formula to indicate that the value of $\Theta^*_{\boldsymbol{\beta}}$ given (p_{i1}, p_{i2}) only depends on the values of $\boldsymbol{\beta}$. With VP1, we have

$$\theta_{1;1,1} = \theta_{1;1,2} = \theta_{1;2,1} = \theta_{1;2,2} < \theta_{2;1,1} = \theta_{2;1,2} = \theta_{2;2,1} = \theta_{2;2,2}.$$

Thus, we know that any set of $\boldsymbol{\beta}$ satisfying $\beta_1 > 0$ and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ can be used to build a SMART with VP1 and under DS1. We proceed to calculate the covariance between two AI values as

$$Cov\left(\widehat{\theta}_{i;k_{i1},k_{i2}}, \widehat{\theta}_{i';k_{i'1},k_{i'2}}\right) = \frac{p_{i1}p_{i2}(\phi_{ik_{i2}} - \phi_{ik_{i1}})(\phi_{i'k_{i'2}} - \phi_{i'k_{i'1}})I(T_i = T_{i'})}{\pi_i} + \frac{p_{i1}\sigma^2 I(S_{i1k_{i1}} = S_{i'1k_{i'1}})}{\pi_i \pi_{i1k_{i1}}} + \frac{p_{i2}\sigma^2 I(S_{i2k_{i2}} = S_{i'2k_{i'2}})}{\pi_i \pi_{i2k_{i2}}},$$

where $i, i', k_{i1}, k_{i2}, k_{i'1}, k_{i'2} = 1, 2$; $I(E) = 1$ when event E occurs and $I(E) = 0$ otherwise. The values of $I(.)$'s depend on the relationship between the two given AIs $d_{i;\ k_{i1},\ k_{i2}}$ and $d_{i';\ k_{i'1},\ k_{i'2}}$. For example, when the two AIs are completely overlapped, we have $I(T_i = T_{i'}) = I(S_{i1k_{i1}} = S_{i'1k_{i'1}}) = I(S_{i2k_{i2}} = S_{i'2k_{i'2}}) = 1$ so that the above formula is the variance of an AI. When both AIs adopt the same Stage-1 treatment but different treatments for either responders or non-responders at Stage 2, we have $I(T_i = T_{i'}) = 1$ and $I(S_{i1k_{i1}} = S_{i'1k_{i'1}}) = I(S_{i2k_{i2}} = S_{i'2k_{i'2}}) = 0$. By this means, we can write Σ^* in (5) as a function of $\{\pi_i, \pi_{ijk}, p_{i1}, p_{i2}, \phi_{ijk}, \sigma_{ijk}\}$, where $\pi_i = \pi_{ijk} = 0.5$, $(p_{i1}, p_{i2}) = \left(\frac{2}{3}, \frac{1}{3}\right)$ and $\sigma_{ijk} = 10$, for $i, j, k = 1, 2$. The value of Σ^* now only depends on $\boldsymbol{\beta}$. Let $\Delta = 0.05$, consider equation

$$(\mathbf{C}\Theta^*_{\boldsymbol{\beta}})^T(\mathbf{C}\Sigma^*\mathbf{C}^T)^-(\mathbf{C}\Theta^*_{\boldsymbol{\beta}}) = 0.05, \tag{B.1}$$

where the contrast matrix

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

By solving $(B.1)$, we obtain a set of

$$\boldsymbol{\beta} = (0,4.48,0,0,0,0,0,0),$$

which will be used to simulate the SMART data under DS1 and BR, with VP1 and $\Delta = 0.05$. In each simulated SMART data based on $\boldsymbol{\beta} = (0,4.48,0,0,0,0,0,0)^T$, there are 8 possible sequences and the sequences-specific mean vector is $(0,0,0,0,4.48,4.48,4.48,4.48)^T$. The solution to equation $(B.1)$ is not unique. However, for the purpose of simulation, any set of solution $\boldsymbol{\beta}$ to equation $(B.1)$ can be used. We fix $\beta_0 = 0$ in all simulations.

References

[1] S.A. Murphy, An experimental design for the development of adaptive treatment strategies, Stat. Med. 24 (10) (2005) 1455–1481 https://doi.org/10.1002/sim.2022.

[2] J.M. Robins, A new approach to causal inference in mortality studies with sustained exposure periods-application to control to the health worker survivor effect, Math. Comput. Model. 7 (1986) 1393–1512, https://doi.org/10.1016/0270-0255(86)90088-6.

[3] S.A. Murphy, M.J. van der Laan, J.M. Robin, CPPRG, Marginal mean models for dynamic regimes, J. Am. Stat. Assoc. 96 (2001) 1410–1423, https://doi.org/10.1198/016214501753382327.

[4] Y. Hochberg, A.C. Tamhane, Multiple comparison procedures, John Wiley and Sons, New York, NY, 1987, https://doi.org/10.1002/9780470316672.

[5] J. Tukey, Comparing individual means in the analysis of variance, Biometrics 5 (2) (1949) 99–114, https://doi.org/10.2307/3001913.

[6] J.C. Hsu, Simultaneous confidence intervals for all distances from the best, Ann. Stat. 9 (5) (1981) 1026–1034, https://doi.org/10.1214/aos/1176345582.

[7] S.B. Ogbagaber, J. Karp, A.S. Wahed, Design of sequentially randomized trials for testing adaptive treatment strategies, Stat. Med. 35 (6) (2016) 840–858, https://doi.org/10.1002/sim.6747.

[8] Y.K. Cheung, B. Chakraborty, K.W. Davidson, Sequential multiple assignment randomized trials (SMART) with adaptive randomization for quality improvement in depression treatment program, Biometrics 71 (2) (2005) 450–459, https://doi.org/10.1111/biom.12258.

[9] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, New York, NY, 1998, https://doi.org/10.1017/CBO9780511802256.

[10] A.L. Oetting, J.A. Levy, R.D. Weiss, S.A. Murphy, Statistical methodology for a SMART design in the development of adaptive treatment strategies, Causality and Psychopathology: Finding the Determinant of Disorders and their Cures, American Psychiatric Publishing, Arlington, VA, 2008, pp. 179–205.

[11] R. Dawson, P.W. Lavori, Sample size calculations for evaluating treatment policies in multi-stage design, Clinical Trials. 7 (6) (2010) 643–652, https://doi.org/10.1177/1740774510376418.

[12] R. Dawson, P.W. Lavori, Efficient design and inference for multistage randomized trials of individualized treatment policies, Biostatistics. 13 (1) (2012) 142–152, https://doi.org/10.1093/biostatistics/kxr016.

[13] W. Feng, A.S. Wahed, A supremum log rank test for comparing adaptive treatment strategies and corresponding sample size formula, Biometrika 95 (3) (2008) 695–707, https://doi.org/10.1093/biomet/asn025.

[14] Z. Li, S.A. Murphy, Sample size formulae for two-stage randomized trials with survival outcomes, Biometrika 98 (3) (2011) 503–518, https://doi.org/10.1093/biomet/asr019.

[15] J.D. Gibbons, I. Olkin, M. Sobel, Selection and Ordering Population: A New Statistical Methodology, Wiley, New York, 1979, https://doi.org/10.1137/1.9781611971101.

[16] S.S. Gupta, S. Panchapakesan, Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations, Wiley, New York, 1979, https://doi.org/10.1137/1.9780898719161.

[17] T.J. Santner, A.C. Tamhane, Design of Experiments: Ranking and Selection, Marcel Dekker, New York, 1984.

[18] C.W. Dunnett, Selection of the best treatment in comparison to a control with application to a medical trial, Design of Experiments: Ranking and Selection, Santner TJ, Tamhane AC, Marcel Dekker, New York, 1984, pp. 47–66.

[19] P.H. Westfall, A. Krishen, Optimally weighted, fixed sequence and gatekeeper multiple testing procedures, J. Stat. Plan. Infer. 99 (1) (2001) 25–40, https://doi.org/10.1016/S0378-3758(01)00077-5.

[20] A. Dmitrienko, W.W. Offen, P.H. Westfall, Gatekeeping strategies for clinical trials that do not require all primary effects to be significant, Stat. Med. 22 (15) (2003) 2387–2400, https://doi.org/10.1002/sim.1526.

[21] A. Dmitrienko, B.L. Wiens, A.C. Tamhane, X. Wang, Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives, Stat. Med. 26 (12) (2007) 2465–2478, https://doi.org/10.1002/sim.2716.

[22] J. Whitehead, Sample sizes for phase II and phase III clinical trial: a n integrated approach, Stat. Med. 5 (5) (1986) 459–464, https://doi.org/10.1002/sim.4780050510.

[23] P.F. Thall, R. Simon, S.S. Ellenberg, Two-stage selection and testing designs for comparative clinical trials, Biometrika 75 (2) (1988) 303–310, https://doi.org/10.2307/2336178.

[24] Y.K. Cheung, Simple sequential boundaries for treatment selection in multi-armed randomized clinical trials with a control, Biometrics 64 (3) (2008) 940–949, https://doi.org/10.1111/j.1541-0420.2007.00929.x.

[25] R.E. Bechhofer, A single-sample multiple decision procedure for ranking means of normal population with known variances, Ann. Math. Statist. 25 (1954) 16–39, https://doi.org/10.1214/aoms/1177728785.

[26] R.E. Bechhofer, T.J. Santner, D.M. Goldsman, Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparison, John Wiley and Sons, Hoboken, New Jersey, 1995.

[27] Y.K. Cheung, Sequential implementation of stepwise procedures for identifying the maximum tolerated dose, JASA 102 (480) (2007) 1448–1461, https://doi.org/10.1198/016214507000000699.