Interim monitoring in sequential multiple assignment randomized trials

Liwen Wu  |  Junyao Wang  |  Abdus S. Wahed

Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence Liwen Wu, Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA.
Email: liw88@pitt.edu

Abstract
A sequential multiple assignment randomized trial (SMART) facilitates the comparison of multiple adaptive treatment strategies (ATSs) simultaneously. Previous studies have established a framework to test the homogeneity of multiple ATSs by a global Wald test through inverse probability weighting. SMARTs are generally lengthier than classical clinical trials due to the sequential nature of treatment randomization in multiple stages. Thus, it would be beneficial to add interim analyses allowing for an early stop if overwhelming efficacy is observed. We introduce group sequential methods to SMARTs to facilitate interim monitoring based on the multivariate chi-squared distribution. Simulation studies demonstrate that the proposed interim monitoring in SMART (IM-SMART) maintains the desired type I error and power with reduced expected sample size compared to the classical SMART. Finally, we illustrate our method by reanalyzing a SMART assessing the effects of cognitive behavioral and physical therapies in patients with knee osteoarthritis and comorbid subsyndromal depressive symptoms.

Keywords
adaptive treatment strategy, dynamic treatment regime, group sequential analysis, interim monitoring, sequential multiple assignment randomized trial

1 | INTRODUCTION

Adaptive treatment strategies (ATSs) are sequences of treatments personalized to patients given their own treatment history. Such strategies are popular in the management of chronic diseases, such as mental health disorders, where the treatment courses usually involve multiple stages due to the dynamic nature of disease progression (Lavori et al., 2000; Lavori and Dawson, 2004; Murphy, 2005; Kidwell, 2014). A sequential multiple assignment randomized trial (SMART) is a systematic and efficient vehicle for comparing multiple ATSs. In a SMART, after initial randomization and evaluation of an intermediate outcome, patients may be rerandomized to future treatments at the following stages. Several well-known applications of SMART include the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study for Schizophrenia (Schneider et al., 2003), the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study for major depression disorder (Rush et al., 2004), the Developmental and Augmented Intervention for Facilitating Expressive Language (CCNIA) trial in children with autism (Kasari et al., 2014), the Adaptive Implementation of Effective Programs Trial (ADEPT) for mood disorders (Kilbourne et al., 2014), and many recent oncology studies (Kidwell, 2014).

The challenge of evaluating ATSs from a SMART lies in taking into account the adaptive and sequential nature of treatment assignment. For continuous outcomes, Murphy (2005) first presented the strategy mean estimators based on the inverse probability weighting (IPW) method and proposed a corresponding sample size formula to test two ATSs from a SMART design. Dawson and Lavori (2010a, 2010b) proposed a semiparametric approach to...
estimate the ATS effects from a SMART and used the “variance inflation factor” to account for the loss of precision due to sequential randomization. Ko and Wahed (2012) demonstrated that the inverse probability weight-normalized (IPW) estimators are more efficient than the IPW estimators. They also provided a Wald test to compare multiple ATSs simultaneously. Based on the same testing procedure, Ogbagaber et al. (2016) further developed the power and sample size formulas for testing multiple ATSs. Methods for binary and survival outcomes have also been established in the past decade (Kidwell and Wahed, 2013; Kidwell et al., 2018).

A caveat of SMARTs is that, given the sequential nature of treatment decisions, many trials require more time to be completed than that for classical randomized trials. Compared to single-stage randomized controlled trials, patients are followed up through multiple stages in SMARTs, and the total study duration depends on the number of stages, definition of intermediate outcomes, as well as other clinical and ethical considerations that are equally important in traditional designs (Kidwell, 2014). One way to improve trial efficiency is to introduce interim analyses where early stops for efficacy and/or futility could be made by repeatedly testing the null hypothesis as data accumulate (Armitage, 1960). A large family of methods have been developed over the years to serve this purpose (Pocock, 1977; O’Brien and Fleming, 1979; Lan and DeMets, 1983; Wang and Tsiatis, 1987, among others), where the multiple looks into the data are facilitated by adjusting the nominal type I error at each interim look. The critical values are obtained from the joint distribution of the test statistics from repeated assessments.

Currently, there is no interim monitoring method in SMARTs for testing multiple ATSs simultaneously. Some statistical methods exist for interim analyses in multi-arm multistage (MAMS) clinical trials, a design similar to SMART but without the sequential randomization feature. Wason and Jaki (2012) proposed a “simultaneous” single-step stopping rule with an error spending function, while Urach and Posch (2016) optimized the multiplicity problem by incorporating sequential rejective tests. However, both methods are based on pairwise comparisons between different treatment strategies, or between different active treatments and a common control arm, instead of a global test of strategy effects. In the SMART setting, Ertefaie et al. (2016) proposed a promising resampling-based method to identify the subset of best ATSs while adjusting for the multiplicity as well as the correlation between strategies. However, it is not clear how this method can be generalized to repeated testing. A Bayesian method for interim monitoring and adaptation has been proposed in a relevant design called small n Sequential Multiple Assignment Randomized Trial (snSMART) that compares different treatments in a multistage trial with a limited sample size. The objective of such a design is, however, inherently different from SMART because it compares individual treatments rather than treatment sequences (Wei et al., 2018; Chao et al., 2020). Thus, the development of sequential testing procedures for global tests of ATSs in SMARTs is warranted from pragmatic considerations.

In this paper, we introduce the traditional group sequential methods for clinical trials to SMARTs. In Section 2, we describe the classical SMART design and the corresponding statistical inference. In Section 3, we introduce a method to perform interim monitoring in SMART (IM-SMART). The results from a simulation study are presented in Section 4. In Section 5, we apply our proposed method to a data set from the RAPID study assessing the effects of cognitive behavioral and physical therapies in patients with knee osteoarthritis and comorbid subsyndromal depressive symptoms to illustrate the benefits of including interim analysis in SMARTs. We conclude the paper with a discussion in Section 6.

2 | CLASSICAL SMART DESIGN AND INFESSION

2.1 | Notation

SMART designs vary depending on the intermediate response, the number of stages, and the number of treatments available at each stage (Ogbagaber et al., 2016). For clarity in describing our method, we use a SMART design similar to the Extending Treatment Effectiveness of Naltrexone (EXTEND) trial (Oslin, 2005; Lei et al., 2012), where patients are initially randomized to two treatments $A_j$, $j = 1, 2$, and then to two maintenance treatments $B_k$, $k = 1, 2$, if they responded, or to two alternative treatments $C_l$, $l = 1, 2$, if they failed to respond to the initial treatments (Figure 1). The objective is to estimate and compare overall strategy effects. In this trial, we have the opportunity to compare eight different treatment strategies $A_j B_k C_l$, $j, k, l = 1, 2$, where $A_j B_k C_l$ is defined as “treat with initial treatment $A_j$, then $B_k$ if respond to $A_j$, or $C_l$ if do not respond to $A_j$.”

To facilitate the comparison of the ATSs, we cast the problem in a counterfactual framework. Define $Y$ as the outcome of interest. Without loss of generality, we will treat $Y$ as a continuous variable in this paper. Let $Y_i(A_j B_k)$ be the potential outcome if patient $i$ received initial treatment $A_j$, responded, and received maintenance treatment $B_k$. Similarly, we define the potential outcome $Y_i(A_j C_l)$ for patient $i$ who received initial treatment $A_j$, did not respond, and received alternative treatment $C_l$. Let $R_{ji}$ denote the counterfactual response status for patient $i$.
Example of SMART design similar to the EXTEND trial (Lei et al., 2012). There are two initial treatments $A_j$, $j = 1, 2$, two maintenance treatments $B_k, k = 1, 2$ for patients who respond to their initial treatments, and two alternative treatments $C_l, l = 1, 2$ for nonresponders. There are in total eight different ATSs embedded: $A_jB_kC_l, j, k, l = 1, 2$

under treatment $j$ (i.e., $R_{jl} = 1$ if the patient responded to the initial treatment $A_j$, and 0 otherwise). The counterfactual outcome $Y$ for patient $i$ following strategy $A_jB_kC_l$ can be written as

\[ Y_i(A_jB_kC_l) = R_{jl}Y_i(A_jB_k) + (1 - R_{jl})Y_i(A_jC_l), \quad j, k, l = 1, 2. \quad (1) \]

Denote $\pi_j$ as the probability of response to the initial treatment $A_j$. The effect of ATS $A_jB_kC_l$ is then

\[ \mu_{jkl} = E(Y(A_jB_kC_l)) = \pi_j \mu_{A_jB_k} + (1 - \pi_j) \mu_{A_jC_l}, \quad j, k, l = 1, 2, \quad (2) \]

where $\mu_{A_jB_k} = E(Y(A_jB_k))$ and $\mu_{A_jC_l} = E(Y(A_jC_l))$. Inference procedure for these strategies will be described in the following subsections.

### 2.2 Observed data and estimators

The observed treatment assignments and responses for patient $i$ can be recorded as a vector $I_i(A), \{R_i, R_jI_j(B_k), (1 - R_j)I_j(C_l), Y_{ij}, j, k, l = 1, 2\}$, where each treatment indicator $I_j(\cdot)$ equals 1 if it corresponds to the actual treatments that patient $i$ was randomized to, and 0, otherwise; $R_i$ is the $i$th patient's response to their initial treatment. $R_{ij} = 1$, if the patient is a responder, and 0, otherwise; $Y_{ij}$ is the observed outcome for patient $i$ at the end of stage 2. As described in Ko and Wahed (2012) and Ogbagaber et al. (2016), the treatment effect of strategy $A_jB_kC_l$, $\mu_{jkl}$, can be estimated by IPW using the estimator

\[
\hat{\mu}_{jkl} = \frac{\sum_{i=1}^{n} I_i(A_i) \left\{ \frac{R_{jI_j}(B_k)}{p_k} + \frac{(1 - R_{jI_j})(C_l)}{q_l} \right\} Y_i}{\sum_{i=1}^{n} I_i(A_i) \left\{ \frac{R_{jI_j}(B_k)}{p_k} + \frac{(1 - R_{jI_j})(C_l)}{q_l} \right\}}, \quad (3)
\]

where $p_k$ is the known randomization probability for $B_k$, and $q_l$ is that for $C_l$. Under the usual assumptions of causal inference, more specifically, consistency, sequential randomization, and positivity assumptions, this estimator is shown to be consistent and can be twice as efficient as a similar estimator where the denominator is just the sample size $\sum_{i=1}^{n} I_i(A_j)$ (Ko and Wahed, 2012). Ogbagaber et al. (2016), following the theory of M-estimation, showed that the asymptotic variance of the above estimator is given by

\[
\text{var}(\hat{\mu}_{jkl}) = \frac{\sigma^2_{jkl}}{n} = \frac{\chi_j}{n} \left[ \pi_j \left( \frac{\sigma^2_{A_jB_k}}{p_k} + (1 - \pi_j)(\mu_{A_jB_k} - \mu_{A_jC_l})^2 \right) + \frac{1 - \pi_j}{q_l} \left( \frac{\sigma^2_{A_jC_l}}{q_l} + \frac{\sigma^2_{A_jB_k}}{p_k} + (1 - \pi_j)(\mu_{A_jB_k} - \mu_{A_jC_l})^2 \right) \right], \quad (4)
\]

where $\chi_j$ is the inverse of randomization probability to $A_j$, $\sigma^2_{A_jB_k} = \text{var}(Y(A_jB_k))$, and $\sigma^2_{A_jC_l} = \text{var}(Y(A_jC_l))$. The asymptotic covariance for two strategies sharing the same initial treatment is, for example,

\[
\text{cov}(\hat{\mu}_{111}, \hat{\mu}_{112}) = \frac{\sigma_{111,112}}{n} = \frac{\chi_1 \pi_1}{n} \times \left\{ \frac{\sigma^2_{A_jB_1}}{p_1} + (1 - \pi_1)(\mu_{A_jB_1} - \mu_{A_jC_1})(\mu_{A_jB_1} - \mu_{A_jC_2}) \right\}, \quad (5)
\]

The variance and the covariance in Equations (4) and (5) can be estimated by first estimating $\pi_i, \mu_{A_jB_k}, \mu_{A_jC_l}$, variances $\sigma^2_{A_jB_k}, \sigma^2_{A_jC_l}$, and the randomization probabilities by their empirical counterparts, and then plugging those estimates in the above formula (Ogbagaber et al., 2016). Alternatively, following Ko and Wahed (2012), the robust estimators for the variance in Equation (4) and the covariance in Equation (5) can be expressed as

\[
\text{vâr}(\hat{\mu}_{jkl}) = \frac{\sigma^2_{jkl}}{n} = \frac{1}{n(n - 1)} \sum_{i=1}^{n} I_i(A_i) \times \left\{ \left( \frac{R_{jI_j}(B_k)}{p_k} + \frac{(1 - R_{jI_j})(C_l)}{q_l} \right)(Y_i - \hat{\mu}_{jkl}) \right\}^2, \quad (6)
\]
and
\[ \text{cov}(\hat{\mu}_{jk1}, \hat{\mu}_{jk'l'}) = \frac{1}{n_j} = \frac{1}{n_j} \sum_{i=1}^{n} I_i(A_j) \]
\[ \times \left\{ \begin{array}{c}
\left( \frac{R_i I_i(B_k)}{\sigma_{kk'}} + \frac{(1-R_i) I_i(C_l)}{\sigma_{ll'}} \right) (Y_i - \hat{\mu}_{jkk'}) \\
\left( \frac{R_i I_i(B_{k'})}{\sigma_{kk'}} + \frac{(1-R_i) I_i(C_{l'})}{\sigma_{ll'}} \right) (Y_i - \hat{\mu}_{jk'l'})
\end{array} \right\} \]
respectively, where \( n_j = n_j / \hat{k}_j = \sum_{i=1}^{n} I_i(A_j) \). Notice that, by design, we have \( \text{cov}(\hat{\mu}_{jk1}, \hat{\mu}_{jk'l'}) = 0 \) for any \( k, l, k', l' \), since data from patients randomized to different initial treatments are independent of each other. It is also easy to see that \( \text{cov}(\hat{\mu}_{jk1}, \hat{\mu}_{jk'l'}) = 0 \) for any \( j, l, l' = 1, 2 \) and \( l \neq l' \) due to the sequential randomization assumption (Ko and Wahed, 2012).

### 2.3 Hypotheses testing

The global null hypothesis to test the homogeneity of all treatment strategies can be written in the form of a linear combination of strategy means, \( H_0 : \boldsymbol{C}\mu = \boldsymbol{0} \), where \( \mu = [\mu_{111}, \mu_{112}, \mu_{121}, \mu_{122}, \mu_{211}, \mu_{212}, \mu_{121}, \mu_{222}]^T \), and \( \boldsymbol{C} \) is a contrast matrix. An example could be \( \boldsymbol{C} = [1, -1 \ast I_7] \), where \( I_7 \) is a \( 7 \times 1 \) vector of ones, and \( I_7 \) is a \( 7 \times 7 \) identity matrix.

To test the null hypothesis, we define the following Wald-type test statistic
\[ T = n \hat{\mu}^T \boldsymbol{C}^T [\hat{\Sigma} \hat{\Sigma}^T]^g \hat{\Sigma} \hat{\mu}, \]
where \( \hat{\mu} \) is the estimated strategy mean vector and \( \hat{\Sigma} \) is an estimate of the variance–covariance matrix \( \Sigma \). More specifically, \( \hat{\Sigma} = \text{block diag}(\hat{\Sigma}_1, \hat{\Sigma}_2) \), where
\[
\hat{\Sigma}_j = \begin{bmatrix}
\hat{\sigma}_{j11} & \hat{\sigma}_{j11,j12} & \hat{\sigma}_{j11,j21} & 0 \\
\hat{\sigma}_{j11,j12} & \hat{\sigma}_{j12} & 0 & \hat{\sigma}_{j12,j22} \\
\hat{\sigma}_{j11,j21} & 0 & \hat{\sigma}_{j21} & \hat{\sigma}_{j21,j22} \\
0 & \hat{\sigma}_{j12,j22} & \hat{\sigma}_{j21,j22} & \hat{\sigma}_{j22} \end{bmatrix}
\]

With \( j, k, l = 1, 2 \), this is an \( 8 \times 8 \) matrix. However, given the dependency structure in \( \Sigma \) resulting from the shared paths between different strategies, \( \Sigma \) maybe singular. Thus, generalized inverse (g-inverse) is used in the definition.

**Theorem 1.** Let \( T \) be the test statistic defined in Equation (8). Under the global null hypothesis \( H_0 : \boldsymbol{C}\mu = \boldsymbol{0} \), \( T \) follows a central \( \chi^2 \) distribution with degrees of freedom \( \nu = \text{rank}(\Sigma \Sigma' \Sigma') \), where \( \Sigma_0 \) is the variance–covariance matrix evaluated under the null. Similarly, under a local alternative hypothesis \( H_a : \boldsymbol{C}\mu = \delta \), for some known constant vector \( \delta \), it follows a noncentral \( \chi^2 \) distribution with degrees of freedom \( \nu = \text{rank}(\Sigma \Sigma' \Sigma') \) and the noncentrality parameter \( \delta = n \delta^T [\Sigma \Sigma' \Sigma']^g \delta \), where \( \Sigma_a \) is the variance–covariance matrix evaluated under the alternative.

Theorem 1 can be seen as a special case of Theorem 5.12 of Boos and Stefanski (2013), which summarizes the results of Styan (1970). We include an outline of the proof in Web Appendix A. The test statistic is invariant to the choice of g-inverse and the specification of the contrast matrix \( \boldsymbol{C} \) as long as \( \boldsymbol{C} \) is of full row rank. When \( \Sigma_0 \) and \( \Sigma_a \) are of full rank, the degrees of freedom is equal to \( \text{rank}(\boldsymbol{C}) \) and the Wald-type statistic defined in Equation (8) coincides with the standard Wald statistic defined with the regular inverse. Given the dependency structure in \( \Sigma, T \) with g-inverse is more appropriate in a general SMART setting compared to the standard Wald statistic, and the critical value should be derived from a central \( \chi^2 \) distribution with degrees of freedom equals to \( \text{rank}(\Sigma \Sigma_0 \Sigma') \) instead of \( \text{rank}(\boldsymbol{C}) \). Failure to use the proper degrees of freedom could lead to a higher chance to reject the null hypothesis than the nominal level (i.e., overpowered), as observed in Ko and Wahed (2012), Ogbagaber et al. (2016), and Zhong (2018).

Even with the proper adjustment of the degrees of freedom for a singular \( \Sigma \), the distribution of the Wald-type statistic defined in Equation (8) might differ substantially from its asymptotic distribution, in particular when the sample size \( n \) is small. Thus, in addition to adjusting the degrees of freedom of the \( \chi^2 \) test, we also propose to multiply the estimator of the variance–covariance matrix with a variance inflation factor \( n^{-p} \), where \( p \) is the total number of parameters estimated in a trial. In the example SMART design with eight strategies in Figure 1, the parameters that need to be estimated include \( (\chi_1, p_1, q_1, \pi_j, \mu_{A_B}, \mu_{A_C}, \sigma_{A_B}, \sigma_{A_C}) \) for \( j, k, l = 1, 2 \), so that \( p = 21 \).

### 3 INTERIM MONITORING IN SMART

Our goal is to introduce the group sequential methods to SMARTs to allow interim monitoring of trial data. Suppose we have \( M \) number of planned analyses in total (i.e., \( M - 1 \) interim analyses). Let \( t_m, m = 1, \ldots, M \) denote the decision time of the \( m \)th analysis, \( n_m \) be the cumulative sample size up to time \( t_m \), \( \mu(t_m) \), and \( \Sigma(t_m) \) denote the corresponding strategy mean vector and the variance–covariance matrix. The test statistic at the \( m \)th analysis is
\[ T(t_m) = n_m \hat{\mu}(t_m)^T \Sigma(t_m) \Sigma(t_m)^T \delta C \hat{\mu}(t_m). \]
(10)
Efficacy boundaries $b_1, \ldots, b_M$ are defined such that the trial stops at the interim analysis $m^*$ if the global null hypothesis is not rejected at all interim looks $1, \ldots, m^* - 1$ and is rejected at the $(m^*)$th look $(T(t_{m^*}) > b_{m^*})$. Such design that maintains significance level $\alpha$ satisfies the closure testing procedure, that is,

$$\sum_{m^* = 1}^{M} \Pr \left( \bigcap_{m = 1}^{m^* - 1} T(t_m) \leq b_m \cap T(t_{m^*}) > b_{m^*} | H_0 \right) = \alpha. \tag{11}$$

If the joint distribution of $T(t_1), \ldots, T(t_M)$ is available, given a specific value of $M$, the overall type I error $\alpha$ can be split into $M$ components to calculate different efficacy boundaries such as the Pocock boundary (Pocock, 1977), the O’Brien Fleming Boundary (O’Brien and Fleming, 1979), or some other prespecified continuous error spending functions (Lan and DeMets, 1983). In a traditional group sequential framework, the joint probability in Equation (11) is usually evaluated recursively (Armitage et al., 1969). Such calculation requires knowledge about the correlation between test statistics $T(t_m)$ across interim looks. Since $T(t_m)$ is in a quadratic form, we consider the following decomposition.

**Theorem 2.** Let $Z(t_m)$ be a vector of length $\nu$, such that $T(t_m) = Z(t_m)^T Z(t_m)$ for $m = 1, \ldots, M$. Under the null hypothesis, treating $\Sigma$ as known, the stacked vector $Z = (Z(t_1), \ldots, Z(t_M))^T$ follows a multivariate normal distribution

$$Z \sim MN_{\nu \times M}(0, \Sigma \otimes I_\nu), \tag{12}$$

where $0$ is a vector of zeros, $\Sigma$ is an $M \times M$ matrix with diagonal elements $\Lambda_{ii} = 1$, and off-diagonal elements $\Lambda_{ij} = \sqrt{n_i/n_j}$, where $n_i$ is the cumulative sample size up to time $t_i$, $I_\nu$ is a $\nu \times \nu$ identity matrix, and $\otimes$ denotes the Kronecker product.

Then, $(T(t_1), \ldots, T(t_M))^T$ follows a multivariate $\chi^2$ distribution of Wishart type with the associated correlation matrix $\Lambda \otimes I_\nu$ (Dickhaus and Royen, 2015). The proof of Theorem 2 is given in Web Appendix B.

### 3.1 Efficacy boundaries via multivariate $\chi^2$ distribution

Given the information proportion at each interim analysis $m, m = 1, \ldots, M$, we can use the distribution of $Z = (Z(t_1), \ldots, Z(t_M))^T$ outlined in Theorem 2 to numerically generate the multivariate distribution of the test statistics $(T(t_1), \ldots, T(t_M))^T$ under the null hypothesis. For example, suppose the eight ATSs from the SMART in Figure 1 is to be compared in an IM-SMART design that has two interim analyses planned at 1/3 (33.3\%) and 2/3 (66.7\%) information proportions of the total sample size followed by a final analysis at the end of the trial ($M = 3$). Suppose under the null hypothesis, each test statistic $T(t_m)$ follows a $\chi^2$ distribution with 5 degrees of freedom. The stacked normalized vector $Z = (Z(t_1), Z(t_2), Z(t_3))^T$ follows a zero-mean multivariate normal distribution with the variance–covariance matrix

$$\begin{bmatrix}
1 & \sqrt{1/2} & \sqrt{1/3} \\
\sqrt{1/2} & 1 & \sqrt{2/3} \\
\sqrt{1/3} & \sqrt{2/3} & 1
\end{bmatrix} \otimes I_5. \tag{13}
$$

To obtain the trivariate $\chi^2$ distribution, one can numerically generate a sufficiently large sample of $Z$ and compute the empirical joint distribution of $T(t_m) = Z(t_m)^T Z(t_m)$ for $m = 1, 2, 3$. To demonstrate, we generated a sample of 100,000 replications from the multivariate normal distribution above using the R package “MASS,” and calculated the empirical trivariate $\chi^2$ distribution. Then, a fine grid search was performed to find the efficacy boundaries satisfying Equation (11) given the overall type I error $\alpha$. When $\alpha = 0.05$, for the Pocock-type boundaries, we have $b_1 = b_2 = b_3 = 14.46$, and the O’Brien Fleming-type boundaries take values $b_1 = 23.28, b_2 = 19.00, b_3 = 13.44$. Other types of boundaries under this setting or with more interim looks can be obtained similarly.

While straightforward, the above approach for finding boundaries might be computationally intensive in some cases. Thus, it is appealing to have alternative solutions. Dickhaus and Royen (2015) explored methods to approximate the multivariate $\chi^2$ distribution with limited dimensions. For example, for $M = 2$, the bivariate $\chi^2$ distribution can be approximated by

$$\Pr(T(t_1) < x_1, T(t_2) < x_2; \Lambda_{12}) = \sum_{n=0}^{\infty} a_n (\lambda_1^2, \ldots, \lambda_\nu^2) G_{\nu/2+n}^{(n)} \left( \frac{x_1}{2} \right) G_{\nu/2+n}^{(n)} \left( \frac{x_2}{2} \right), \tag{14}$$

where $\Lambda_{12} = \text{cov}(Z(t_1), Z(t_2))$, canonical correlations $\lambda_1, \ldots, \lambda_\nu$ are the roots of eigenvalues for $\Lambda_{12} \Lambda_{12}^T$, $a_n = a_n (\lambda_1^2, \ldots, \lambda_\nu^2)$ is a set of polynomials defined recursively as

$$a_0 = 1, \quad a_{n+1} = \frac{1}{2n+1} \sum_{\mu=0}^{n} a_{n-\mu} \sum_{\mu=1}^{\nu} \lambda_\mu (\lambda_{\mu+1}) \tag{15}$$
TABLE 1 Efficacy boundaries for single interim analysis planned at different information proportions and the final analysis, obtained by the approximation method, $M = 2, \nu = 5, \alpha = 0.05$

| Info. prop. | Pocock-type boundaries | OBF-type boundaries |
|-------------|------------------------|---------------------|
|             | Interim     | Final    | Interim | Final |
| 0.20        | 12.72       | 12.72    | 24.78   | 11.08  |
| 0.30        | 12.66       | 12.66    | 20.28   | 11.11  |
| 0.40        | 12.59       | 12.59    | 17.68   | 11.18  |
| 0.50        | 12.50       | 12.50    | 15.94   | 11.27  |
| 0.60        | 12.39       | 12.39    | 14.68   | 11.37  |
| 0.70        | 12.26       | 12.26    | 13.72   | 11.48  |
| 0.80        | 12.08       | 12.08    | 12.91   | 11.55  |
| 0.90        | 11.85       | 11.85    | 12.17   | 11.55  |

\[ \text{TABLE 2 Strategy means under alternative hypothesis for different scenarios in the simulation study, calculated by plugging in the simulation setups to Equation (2)} \]

|                | \( \mu_{111} \) | \( \mu_{112} \) | \( \mu_{121} \) | \( \mu_{122} \) | \( \mu_{211} \) | \( \mu_{212} \) | \( \mu_{221} \) | \( \mu_{222} \) |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Alt(1)         | 17.5           | 15.0           | 21.0           | 18.0           | 17.5           | 15.0           | 21.0           | 18.5           |
| Alt(2)         | 19.0           | 15.0           | 20.4           | 16.4           | 17.5           | 15.0           | 21.0           | 18.5           |
| Alt(3)         | 16.5           | 15.0           | 21.4           | 19.0           | 17.5           | 15.0           | 21.0           | 18.5           |
| Alt(4)         | 19.0           | 15.0           | 20.4           | 16.4           | 16.5           | 15.0           | 21.4           | 19.9           |

where \( \hat{S}_{m^*} \) is estimated by Ertefaie’s resampling method at each interim look whenever the global null hypothesis is rejected.

4 | SIMULATION

To evaluate the operating characteristics of the proposed IM-SMART design, we conducted a variety of simulations. The simulation scenarios were designed based on the SMART described in Figure 1. Under each scenario, we first randomly assigned each patient to the initial treatment \( A_j \) with probability \( \kappa_j \) for \( j = 1, 2 \). The response \( R_i \) for patient \( i \) was generated from a Bernoulli distribution with probability of success \( \pi_j \), if the patient was randomly assigned to treatment \( A_j, j = 1, 2 \). Given the initial assignment and the response, we then simulated the second-stage treatment assignment with probability \( p_1 \) for treatment \( B_1 \) among responders, and with probability \( q_1 \) for treatment \( C_1 \) among nonresponders. After obtaining the realization of treatment paths for each patient in the simulated data set, we generated the counterfactual outcome variables \( Y(A_j B_k) \) and \( Y(A_j C_l) \) from normal distributions with means \( \mu_{A_j B_k} \) and \( \mu_{A_j C_l} \), and standard deviations \( \sigma_{A_j B_k} \) and \( \sigma_{A_j C_l} \), respectively, for \( j, k, l = 1, 2 \). The observed outcome \( Y_i \) for the \( i \)th patient in the sample is then taken as the counterfactual outcome consistent with the patient’s treatment path.

We followed Ko and Wahed (2012) and Ogbagaber et al. (2016) for specification of the simulation parameters. Specifically, under the null hypothesis, we set \( \mu_{A_j B_k} = \mu_{A_j C_l} = 15, \forall j, k, l = 1, 2 \); while under the alternative hypothesis, we set \( \mu_{A_j B_k} = \mu_{A_j C_l} = 15, \mu_{A_j B_2} = 22, \text{and} \mu_{A_j C_1} = 20, \forall j = 1, 2 \). For the true values of variances, we set \( \sigma_{A_j B_k} = 12 \) and \( \sigma_{A_j C_l} = 10, \forall j, k, l = 1, 2 \). The four alternative scenarios we explored are established by varying the first-stage response rate \( \pi_j \) \((\pi_1, \pi_2) = \{(0.5, 0.5), (0.2, 0.5), (0.7, 0.5), (0.2, 0.7)\}\). The strategy means of the four different alternative scenarios can be calculated by plugging in these settings to Equation (2) and they are shown in Table 2. For example, under alternative scenario 1, the strategy mean for ATS \( A_1 B_1 C_1 \), \( \mu_{111} = \pi_1 * \mu_{A_1 B_1} + (1- \pi_1) * \mu_{A_1 C_1} = 0.5 * 15 + 0.5 * \)

and \( G_{\nu/2+n}^{(n)}(\cdot) \) is the \( n \)th derivatives of a Gamma density with parameters \( \nu/2 + n \) (shape) and 1 (rate). Applying this approximation of bivariate \( \chi^2 \) distribution to the left-hand side of the closure testing procedure in Equation (11), we can obtain the efficacy boundaries for \( M = 2 \) with different information proportions at the interim analysis by a grid search. Table 1 shows the boundaries produced by this approximation.

3.2 | Identifying the best ATS(s)

Rejection of the global null hypothesis at either an interim or the final analysis in an IM-SMART design implies that there is at least one ATS significantly different from another ATS. One might be interested in identifying the most effective ATS(s) in the trial. Taking into account the correlation between treatment estimates and multiplicity, we employ the method proposed by Ertefaie et al. (2016) to perform a post hoc test to identify the best strategies whenever the global null is rejected. Suppose among all eight strategies in the SMART setting aforementioned, there is a true set of best ATSs, \( S \), with the maximum (or the best) outcome, and let \( \hat{S}_m \) denote a set of ATSs that are not statistically different from \( S \) based on the data accumulated by interim look at time \( t_m, m = 1, ..., M \). Thus, for an \( M \)-look group sequential SMART design, the overall probability to identify the best ATSs is

\[
\text{best.select} = \sum_{m=1}^{M} \Pr \left( \hat{S}_m \leq \hat{S}_m^* \right) \leq \sum_{m=1}^{M} \left( \int_{t_m}^{t_m^*} T^g(t_m) \leq b_m \cap T^g(t_m^*) > b_m^* \right),
\]

(16)
TABLE 3  False rejection rates for the proposed IM-SMART design under null scenarios with an interim analysis taking place at information proportion of 50%

| π₁ | π₂ | p₁ | n max | Rej at interim | Rej at final | Empirical type I error | n = n_max | Rej at interim | Rej at final | Empirical type I error |
|----|----|----|-------|--------------|-------------|------------------------|-----------|--------------|-------------|------------------------|
| 0.5 | 0.5 | 0.5 | 252 | 0.021 | 0.025 | 0.045 | 0.024 | 0.022 | 0.045 |
| 0.5 | 0.5 | 0.5 | 317 | 0.039 | 0.025 | 0.063 | 0.035 | 0.024 | 0.058 |
| 0.2 | 0.5 | 0.5 | 263 | 0.024 | 0.023 | 0.046 | 0.025 | 0.024 | 0.048 |
| 0.2 | 0.5 | 0.5 | 309 | 0.031 | 0.021 | 0.051 | 0.028 | 0.023 | 0.050 |
| 0.7 | 0.5 | 0.5 | 244 | 0.023 | 0.017 | 0.040 | 0.028 | 0.022 | 0.049 |
| 0.7 | 0.5 | 0.7 | 273 | 0.029 | 0.021 | 0.049 | 0.023 | 0.026 | 0.048 |
| 0.2 | 0.7 | 0.5 | 254 | 0.019 | 0.022 | 0.041 | 0.025 | 0.025 | 0.049 |
| 0.2 | 0.7 | 0.7 | 278 | 0.031 | 0.023 | 0.053 | 0.024 | 0.019 | 0.043 |
| OBF-type boundary |
| 0.5 | 0.5 | 0.5 | 228 | 0.007 | 0.040 | 0.046 | 0.007 | 0.044 | 0.051 |
| 0.5 | 0.5 | 0.8 | 288 | 0.010 | 0.052 | 0.062 | 0.009 | 0.044 | 0.053 |
| 0.2 | 0.5 | 0.5 | 239 | 0.003 | 0.039 | 0.041 | 0.005 | 0.042 | 0.047 |
| 0.2 | 0.5 | 0.8 | 280 | 0.009 | 0.044 | 0.052 | 0.007 | 0.047 | 0.054 |
| 0.7 | 0.5 | 0.5 | 221 | 0.005 | 0.035 | 0.040 | 0.007 | 0.050 | 0.057 |
| 0.7 | 0.5 | 0.7 | 248 | 0.008 | 0.041 | 0.048 | 0.009 | 0.042 | 0.050 |
| 0.2 | 0.7 | 0.5 | 230 | 0.003 | 0.036 | 0.039 | 0.008 | 0.040 | 0.047 |
| 0.2 | 0.7 | 0.7 | 252 | 0.007 | 0.039 | 0.045 | 0.008 | 0.043 | 0.051 |

Note: Different null scenarios are defined by varying the response rates for initial treatments $A_j, \pi_j$, for $j = 1, 2$, and the probability to rerandomize responders to treatment $B_1, p_1$, in the second stage. The empirical type I error is evaluated under two sample sizes for the final analysis, $n = \{n_{max}, 500\}$, and over 5000 simulated trials. Rej at interim: the probability of rejection at the interim analysis; Rej at final: the probability of rejection at the final analysis condition on not rejecting at the interim analysis; $n_{max}$: the maximum sample size required to achieve 90% power at the final analysis.

20 = 17.5. We also varied the second-stage randomization probability $p_1$ for responders ($p_1 = \{0.5, 0.7, 0.8\}$) under each of the scenarios. For an easier comparison across scenarios, we assumed equal randomization for initial treatments (i.e., $\pi_1 = \pi_2 = 2$) and second-stage treatments for nonresponders (i.e., $q_1 = q_2 = 0.5$).

We planned a single interim analysis when the information proportion is 0.5, which means the interim analysis is triggered at half of the maximum sample size. However, these results can be extended to other information proportions and more than one interim analysis (see Section 3.2). Using Equations (3), (6), and (7), we estimated the strategy means and the variance–covariance matrix for each sample. We would like to note here that under the null hypothesis, $\Sigma_0$ is singular and has rank 6 for all scenarios. Thus, with a contrast matrix $C = [1_7, -1 \ast I_7]$, the global Wald test statistic $T$ follows a $\chi^2$ distribution with 5 degrees of freedom. For the interim and final decisions, we used the efficacy boundaries shown in Table 1 at 0.5 information proportion. Under each scenario, we simulated 5000 trials and evaluated the empirical type I error (proportion of samples for which the null hypothesis is rejected) and the power (rejection rate under the alternative scenarios).

The results under various null scenarios are shown in Table 3. For each scenario, we evaluated the overall type I error rate under two sample sizes: $n = n_{max}$, which is the estimated maximum sample size to achieve 90% power at the final analysis, and a larger sample size $n = 500$, to assess the asymptotic behavior of the test statistic. The type I error is generally well-controlled around the nominal rate of $\alpha = 0.05$ for almost all null scenarios at $n = n_{max}$. For the larger sample size, the empirical type I error closely approximates the nominal level.

Table 4 summarizes the performance of the group sequential analysis with sample size $n_{max}$ under the alternative scenarios. In addition to the empirical rejection rates at the interim, the conditional rejection rate at the final analysis (if the global null hypothesis is not rejected at the interim), and the empirical power, Table 4 also presents the average sample size at which the trial stops ($E(n)$) and the sample size required to achieve 90% power in the absence of an interim analysis ($n_{classical}$) according to
TABLE 4 Power of the proposed IM-SMART design under alternative scenarios with an interim analysis taking place at information proportion of 50%

| $\pi_1$ | $\pi_2$ | $p_1$ | Rej at interim | Rej at final | Empirical power | Best. select | $E(n)$ | $n_{\text{classical}}$ | $n_{\text{max}}$ |
|--------|--------|------|---------------|--------------|----------------|--------------|--------|-----------------|-----------------|
|        |        |      |               |              |                |              |        |                 |                 |
| **Pocock-type boundary** | | | | | | | | | |
| Alt(1) | 0.5 | 0.5 | 0.5 | 0.43 | 0.78 | 0.88 | 0.86 | 198 | 225 | 252 |
|        | 0.5 | 0.5 | 0.8 | 0.52 | 0.79 | 0.88 | 0.90 | 235 | 283 | 317 |
| Alt(2) | 0.2 | 0.5 | 0.5 | 0.45 | 0.78 | 0.88 | 0.86 | 204 | 235 | 263 |
|        | 0.2 | 0.5 | 0.8 | 0.50 | 0.80 | 0.88 | 0.90 | 232 | 276 | 309 |
| Alt(3) | 0.7 | 0.5 | 0.5 | 0.43 | 0.78 | 0.87 | 0.87 | 192 | 218 | 244 |
|        | 0.7 | 0.5 | 0.7 | 0.47 | 0.79 | 0.88 | 0.90 | 208 | 244 | 273 |
| Alt(4) | 0.2 | 0.7 | 0.5 | 0.43 | 0.79 | 0.88 | 0.88 | 199 | 226 | 254 |
|        | 0.2 | 0.7 | 0.7 | 0.47 | 0.79 | 0.88 | 0.88 | 213 | 248 | 278 |
| **OBF-type boundary** | | | | | | | | | |
| Alt(1) | 0.5 | 0.5 | 0.5 | 0.22 | 0.84 | 0.88 | 0.86 | 203 | 225 | 228 |
|        | 0.5 | 0.5 | 0.8 | 0.28 | 0.83 | 0.88 | 0.86 | 248 | 283 | 288 |
| Alt(2) | 0.2 | 0.5 | 0.5 | 0.21 | 0.84 | 0.87 | 0.87 | 214 | 235 | 239 |
|        | 0.2 | 0.5 | 0.8 | 0.31 | 0.88 | 0.92 | 0.92 | 237 | 276 | 280 |
| Alt(3) | 0.7 | 0.5 | 0.5 | 0.20 | 0.83 | 0.86 | 0.86 | 199 | 218 | 221 |
|        | 0.7 | 0.5 | 0.7 | 0.23 | 0.82 | 0.87 | 0.86 | 220 | 244 | 248 |
| Alt(4) | 0.2 | 0.7 | 0.5 | 0.22 | 0.84 | 0.88 | 0.87 | 205 | 226 | 230 |
|        | 0.2 | 0.7 | 0.7 | 0.23 | 0.83 | 0.87 | 0.87 | 223 | 248 | 252 |

Note: Different alternative scenarios are defined in Table 2 by varying the response rates for initial treatments $A_j$, $\pi_j$, for $j = 1, 2$, and the probability to rerandomize responders to treatment $B$, $p_1$, in the second stage. The empirical power is evaluated by the percentage of any rejection of the global null hypothesis either at the interim or the final analysis, and Best.select is the probability of correctly selecting the best ATs. A total of 5000 trials are simulated under each scenario. Rej at interim: the probability of rejection at the interim analysis; Rej at final: the probability of rejection at the final analysis condition on not rejecting at the interim analysis; $E(n)$: the average sample size at which the trial stops; $n_{\text{classical}}$: the sample size required to achieve 90% power for a classical SMART design; $n_{\text{max}}$: the maximum sample size required to achieve 90% power at the final analysis adjusted for interim look.

Ogbagaber et al. (2016). For the Pocock-type boundaries, under all different scenarios, the rejection rate at interim is 50% or lower. The overall empirical power is close to the designed level of 90%. The estimated expected sample sizes (the sample size at which the trial stops) for all scenarios are considerably lower than that of a classical SMART design with no interim analysis. In general, the IM-SMART design saves about 15–20% of the total sample size on average. The performance of the OBF-type boundaries shows a similar trend. Since OBF boundaries are more conservative at the interim analysis, the rate of early stopping, in this case, is less than that under Pocock-type boundaries. The estimated expected sample size is generally larger than that under Pocock-type boundaries, but it still leads to about a 10% reduction in the expected total sample size from that of an equivalent classic SMART.

The Best.select column in Table 4 presents the probability of correctly selecting the best treatment strategies (with the maximum outcome) under each alternative scenario via the procedure described in Section 3.2. The probability is very close to the empirical power, implying that the IM-SMART design is powerful in identifying the best ATs.

5 | ANALYSIS OF RAPID STUDY

In this section, we reanalyze the RAPID (Reducing pain; Preventing depression) SMART trial (Karp et al., 2019) using IM-SMART approach. The goal of this trial was to evaluate the effects of cognitive behavioral therapy (CBT) and physical therapy (PT) in patients with knee osteoarthritis and comorbid subsyndromal depressive symptoms. It is worth mentioning that any interim analysis should be planned a priori, and this application is for demonstration purposes only. Moreover, this study is relatively small-sized compared to the sample sizes considered in Section 4, and it was not designed for the global Wald test we apply here. Thus, the test might be underpowered. Nevertheless, this application signifies the benefit of adding interim analyses to SMARTs.

In the RAPID study, 99 patients were randomized to CBT ($n = 39$), PT ($n = 40$), and enhanced usual care (EUC, $n = 20$). After 8 weeks of the initial treatment, patients’ responses were assessed by improvement in the PGIC (patient-reported global impression of change) score in activity limitations, symptoms, emotions, and overall
quality of life. A result of at least “much better” for the P-GIC score is defined as a response. Nonresponders were rerandomized to either having an additional 4 weeks of their initial treatment or switching to the other non-EUC treatment.

We will analyze the RAPID data in two different ways to demonstrate how our method could be applied to various SMART settings. The first analysis will exclude the EUC arm so that the design becomes very similar to the one considered in the method section. The second analysis will analyze the data including the EUC arm. In the first case, there are four ATSs in this SMART, namely:

1. Start on 8 weeks of CBT, if no response, continue CBT for 4 more weeks (CBT–CBT);
2. Start on 8 weeks of CBT, if no response, switch to PT for 8 weeks (CBT–PT);
3. Start on 8 weeks of PT, if no response, switch to CBT for 8 weeks (PT–CBT);
4. Start on 8 weeks of PT, if no response, continue PT for 4 more weeks (PT–PT).

All patients were followed for 12 months regardless of their initial treatment assignment or response status. We considered three of the continuous outcomes measured at the last follow-up visit in this application: a measure of pain (PAIN, numeric rating score), depression (using PHQ-9), and anxiety (using GAD-7) (Karp et al., 2019). Excluding patients whose outcomes of interest are missing, we have 65 patients in the analysis set (Figure 2).

The question of interest is whether these three aforementioned outcome measures differ across these four ATSs, and whether the differences, if any, can be detected early with IM-SMART design. For each of the outcomes, the global null hypothesis is that all of the four ATSs have the same strategy means. Under the null hypothesis, the global Wald-type test statistic follows a $\chi^2$ distribution with 3 degrees of freedom. To mimic the process of conducting an interim analysis planned a priori, we assumed the enrollment of patients and the evaluation of outcomes follow the order of their study ID. Suppose the interim analysis is conducted after 46 patients were evaluated for their outcomes (i.e., at 70% of total information), the interim and the final critical values according to the Pocock-type boundaries are both 8.83, which is obtained following the steps described in Section 3.1. Table 5 shows the results for the interim and the final analyses for each of the outcomes of interest.

At the interim analysis, the test statistics for outcome PHQ-9 is below the critical value, so patient enrollment
would have continued beyond the interim if PHQ-9 was the primary outcome. The final analysis based on the whole sample agrees with the interim result that there is no significant difference between treatment strategies based on the PHQ-9 measure. The results for GAD-7 are in the other direction. The interim test statistic for GAD-7 is larger than the critical value, implying an early stop for efficacy if GAD-7 was our primary outcome. The final test statistic agrees with the interim decision. Had interim analysis been included in the trial design, it would have led to a maximum of 30% reduction in the total sample size. The results for the PAIN score follow a similar pattern as that for the PHQ-9 score.

For the GAD-7 score, since the global null hypothesis is rejected at the interim and the final analyses, a post hoc analysis is triggered. The goal of this post hoc analysis is to find the best ATSs (i.e., with lowest GAD-7) among the four. Using the multiple comparison procedure proposed by Ertefaie et al. (2016), this is equivalent to identifying strategies with index $i$ where $\mu_i \leq \min(\mu_j - c_1 \hat{\sigma}_{ij})$ for all $j \neq i$, where $\mu_i$ is the strategy mean with index $i$, $\hat{\sigma}_{ij}$ is the pooled variance between the two strategies that are being compared, and $c_1$ is a constant, following Ertefaie’s resampling method, can be estimated to account for the correlation between treatment strategies while maintaining overall type I error. For the interim readouts, the estimated values are $c_{CBT-CBT} = 1.908$, $c_{CBT-PT} = 1.598$, $c_{PT-CBT} = 1.726$, $c_{PT-PT} = 1.766$, and strategies CBT–CBT and PT–PT are identified as the best ATSs. For the final analysis, $c_{CBT-CBT} = 1.674$, $c_{CBT-PT} = 1.715$, $c_{PT-CBT} = 1.679$, $c_{PT-PT} = 1.826$. The best ATSs selected are the same as the interim analysis.

In the second analysis, we include the EUC as a standalone strategy, and altogether we will compare five strategies. Since patients on the EUC arm were not examined for the intermediate endpoint and were not rerandomized during the study, we estimated the mean treatment effect and the variance for the EUC arm by the observed sample mean and variance. Under the null hypothesis, the global Wald-type test statistics follow a $\chi^2$ distribution with 4 degrees of freedom. Similar to the first analysis, a hypothetical interim look is planned at 70% of total information, and the interim and final critical values (Pocock type) of 10.59 are used. The results are similar to the first analysis (Web Table 1).

Similar to the first analysis, a post hoc analysis would be triggered for the GAD-7 score. Since a control arm (EUC) is included in this study, the goal of this post hoc analysis is to test whether any of the four ATSs performs better than EUC. A strategy has significantly lower GAD-7 compared to EUC when $(\hat{\mu}_{EUC} - \mu_i) / \hat{\sigma}_i \geq c$, where $\hat{\mu}_i$ is the estimated strategy effect for each ATSs other than EUC, $i$ is an arbitrary index, $\hat{\mu}_{EUC}$ is the estimated treatment effect of the EUC arm, $\hat{\sigma}_i$ is the pooled standard error between EUC and the strategy indexed with $i$. In this analysis, $c$ takes values 1.714 and 1.546 for the interim and final analyses, respectively. At both the interim and final analysis, strategies CBT–CBT and PT–PT show a significantly lower GAD-7 score compared to the EUC arm. Overall, this application highlights the favorable resource-saving benefit of IM-SMART design, and the flexibility of the proposed method when extended to various settings.

### TABLE 5 Reanalysis of RAPID data, where an interim analysis is assumed to take place at 70% of total information

| Strategies  | PHQ-9 | GAD-7 | PAIN |
|-------------|-------|-------|------|
|             | Interim mean (var) | Final mean (var) | Interim mean (var) | Final mean (var) | Interim mean (var) | Final mean (var) |
| CBT-CBT     | 2.61 (0.37) | 2.62 (0.50) | 1.59 (0.34) | 1.63 (0.29) | 4.77 (0.73) | 5.42 (0.69) |
| CBT-PT      | 3.43 (1.44) | 3.23 (0.79) | 3.52 (0.73) | 3.28 (0.42) | 5.90 (2.17) | 6.24 (1.19) |
| PT-CBT      | 4.01 (0.67) | 3.43 (0.35) | 2.53 (0.47) | 2.45 (0.28) | 5.54 (1.21) | 5.38 (0.85) |
| PT-PT       | 4.42 (0.85) | 3.64 (0.54) | 2.11 (0.20) | 1.30 (0.12) | 6.04 (1.09) | 5.27 (0.75) |
| $\chi^2_{d.f. = 3}$ | 2.804 | 1.039 | 8.955 | 11.090 | 1.121 | 0.551 |
| Decision    | Continue | Do not reject | Stop | Reject | Continue | Do not reject |

Note: The strategy means and variances are estimated by Equations (3) and (6). The decisions are made by comparing the test statistic $\chi^2_{d.f. = 3}$ with a Pocock-type critical value $b_1 = b_2 = 8.83$ for each of the outcomes of interest.

### DISCUSSION

SMART is an efficient way to evaluate multiple ATSs, but it is generally more resource-intensive than traditional designs. We introduced interim monitoring to SMARTs and used simulations to show that the proposed IM-SMART design can maintain the overall power and type I error while reducing the expected sample size. This resource-saving benefit of IM-SMART was also illustrated by reanalyzing the RAPID data.

The key feature of our proposed IM-SMART design is that it allows early stopping at the interim analysis when the early efficacy signal is sufficiently strong. In the
traditional group sequential framework, it is also common to include futility boundaries at interim analyses, where one can claim early stop if there is strong evidence in favor of the null hypothesis or the direction is opposite to the alternative. From an ethical perspective, it would be less beneficial to enroll more patients in the inferior treatment arms or, sometimes, to continue the trial. The IM-SMART design only considers efficacy boundaries for the global null hypothesis because SMARTs are usually designed to compare strategies combining treatments that have been approved individually for safety. Thus, a futility analysis for safety may not be warranted in this setting.

In this paper, we utilized two commonly used boundaries, the Pocock and O’Brian Fleming types, to demonstrate our idea of group sequential testing in the SMART setting. The Pocock type has constant boundaries across interim looks, while the O’Brian Fleming type is more stringent in early stopping. Both types of boundaries are commonly used, and the simulation results are consistent with the trend described in traditional group sequential designs. The proposed method can also be extended to other types of boundaries that are discussed in detail in Jennison and Turnbull (1999).

Our proposed method is developed based on a global test between all treatment strategies and the null hypothesis is written as \( H_0 : \bm{C} \mu = \bm{0} \). In this form, it is assumed that we are interested in a linear contrast. In some cases, researchers may also want to investigate some nonlinear contrast between treatment strategies, where the null hypothesis can be written as \( H_0 : \bm{h}(\mu) = \bm{0} \). As discussed in Boos and Stefanski (2013), the test statistic follows a similar asymptotic distribution under a straightforward application of the Delta method. However, earlier studies also point out that, with finite sample size, the numerical outcomes of such test may be substantially different for different forms of the nonlinear function \( \bm{h} \) that are algebraically equivalent under the same null hypothesis (Gregory and Veall, 1985; Lafontaine and White, 1986). Moreover, researchers may be more interested to claim early stop only if there is an ATS that is superior to all the others or a common control, rather than the general global test considered in this paper. We proposed to perform pairwise comparisons as a post hoc test. However, a more direct answer, for example, restricted inference on a “one-sided” alternative hypothesis, may be of interest. In addition, if researchers are interested in deeply tailored ATSs as the secondary aim of the study, the aim could be compromised by early stopping. Future work is needed to investigate how the proposed method can be extended to accommodate these considerations.

An important practical limitation of such group sequential analysis is the information and time gap between the proposed number of patients to be enrolled before the interim analysis and the time when their outcomes are available for evaluation. Continuing enrollment during this gap is sometimes known as trial “overrun” (Whitehead, 1992) if the interim decision is early stopping. The challenge of overrun is inherited from the traditional group sequential design, and it might be problematic when time-to-event outcomes that take a long time to observe are considered in the trial. Since follow-up is generally longer in SMARTs to accommodate multiple stages of randomization and follow-up, this might pose a practical challenge during implementation. In our simulation study and the reanalysis of the RAPID study, we assumed that enrollment is suspended until the responses are evaluated in all patients enrolled before the interim analysis. In practice, this may not be feasible because of the timeline and budget constraints. There are several statistical procedures proposed in the literature to account for the overrun data in the traditional group sequential framework, including Whitehead’s (1992) deletion method and the combination test approach (Hall et al., 2008; Hampson and Jennison, 2013). Incorporating this consideration in estimating the strategy effects at the final analysis and identifying the best timing for interim analysis in the SMART setting can be explored in future research.

Our proposed method is based on a specific type of a two-stage SMART design. However, it can be easily adapted to other SMART designs, for example, where only non-responders will be rerandomized at the second stage, as shown in our application to the RAPID study. Moreover, the method is developed based on a continuous outcome. Future research can generalize this IM-SMART design to include other types of outcomes such as binary or survival endpoints. The approximation of efficacy boundaries for more than two interim looks is also of interest, where the challenge is to obtain a closed-form boundary from a multivariate \( \chi^2 \) distribution.

**ACKNOWLEDGMENTS**

We thank the Center for Research Computing at the University of Pittsburgh for providing computational support. We thank Dr. Jordan Karp for providing the RAPID study data.

**DATA AVAILABILITY STATEMENT**

The RAPID (Reducing Pain; Preventing Depression) study data used in this paper were obtained directly from the study PI (Jordan Karp, MD; karplj@arizona.edu). This data set is not publicly available but may be requested from the study PI. A simulated data set is available with this paper at the Biometrics website on Wiley Online Library.
REFERENCES

Armitage, P. (1960) Sequential Medical Trials. Oxford: Blackwell Scientific Publications.

Armitage, P., McPherson, C. and Rowe, B. (1969) Repeated significance tests on accumulating data. Journal of the Royal Statistical Society: Series A (General), 132, 235–244.

Boos, D.D. and Stefanski, L.A. (2013) Essential Statistical Inference: Theory and Methods, Volume 120. New York: Springer Science & Business Media.

Chao, Y.-C., Braun, T.M., Tamura, R.N., Kidwell, K.M. (2020) A Bayesian group sequential small n sequential multiple-assignment randomized trial. Journal of the Royal Statistical Society Series C, 69, 663–680.

Dawson, R. and Lavori, P.W. (2010a) Efficient design and inference for multi-stage randomized trials of individualized treatment policies. COBRA Preprint Series, p. 69.

Dawson, R. and Lavori, P.W. (2010b) Sample size calculations for evaluating treatment policies in multi-stage designs. Clinical Trials, 7, 643–652.

Dickhaus, T. and Royen, T. (2015) A survey on multivariate chi-square distributions and their applications in testing multiple hypotheses. Statistics, 49, 427–454.

Ertefaie, A., Wu, T., Lynch, K.G. and Nahum-Shani, I. (2016) Identifying a set that contains the best dynamic treatment regimes. Biostatistics, 17, 135–148.

Gregory, A.W. and Veall, M.R. (1985) Formulating Wald tests of nonlinear restrictions. Econometrica: Journal of the Econometric Society, 53(6), 1465–1468.

Hall, W. and Ding, K. (2008) Sequential tests and estimates after overrunning based on p-value combination. In: Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh. Beachwood, OH: Institute of Mathematical Statistics, pp. 33–45.

Hampson, L.V. and Jennison, C. (2013) Group sequential tests for delayed responses (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75, 3–54.

Jennison, C. and Turnbull, B.W. (1999) Group Sequential Methods with Applications to Clinical Trials. Boca Raton: CRC Press.

Karp, J.F., Zhang, J., Wahed, A.S., Anderson, S., Dew, M.A., Fitzgerald, G.K., et al. (2019) Improving patient reported outcomes and preventing depression and anxiety in older adults with knee osteoarthritis: results of a sequenced multiple assignment randomized trial (SMART) study. American Journal of Geriatric Psychiatry, 27, 1035–1045.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., et al. (2014) Communication interventions for minimally verbal children with autism: a sequential multiple assignment randomized trial. Journal of the American Academy of Child & Adolescent Psychiatry, 53, 635–646.

Kidwell, K.M. (2014) SMART designs in cancer research: past, present, and future. Clinical Trials, 11, 445–456.

Kidwell, K.M., Seewald, N.J., Tran, Q., Kasari, C. and Almirall, D. (2018) Design and analysis considerations for comparing dynamic treatment regimens with binary outcomes from sequential multiple assignment randomized trials. Journal of Applied Statistics, 45, 1628–1651.

Kidwell, K.M. and Wahed, A.S. (2013) Weighted log-rank statistic to compare shared-path adaptive treatment strategies. Biostatistics, 14, 299–312.

Kilbourne, A.M., Almirall, D., Eisenberg, D., Waxmonsky, J., Goodrich, D.E., Fortney, J.C., et al. (2014) Protocol: adaptive implementation of effective programs trial (ADEPT): cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. Implementation Science, 9, 1–14.

Ko, J.H. and Wahed, A.S. (2012) Up-front versus sequential randomizations for inference on adaptive treatment strategies. Statistics in Medicine, 31, 812–830.

Lafontaine, F. and White, K.J. (1986) Obtaining any Wald statistic you want. Economics Letters, 21, 35–40.

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

Lavori, P.W. and Dawson, R. (2004) Dynamic treatment regimes: practical design considerations. Clinical Trials, 1, 9–20.

Lavori, P.W., Dawson, R. and Rush, A.J. (2000) Flexible treatment strategies in chronic disease: clinical and research implications. Biological Psychiatry, 48, 605–614.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D. and Murphy, S.A. (2012) A “SMART” design for building individualized treatment sequences. Annual Review of Clinical Psychology, 8, 21–48.

Murphy, S.A. (2005) An experimental design for the development of adaptive treatment strategies. Statistics in Medicine, 24, 1455–1481.

O’Brien, P.C. and Fleming, T.R. (1979) A multiple testing procedure for clinical trials. Biometrics, 35(3), 549–556.

Ogbagaber, S.B., Karp, J. and Wahed, A.S. (2016) Design of sequentially randomized trials for testing adaptive treatment strategies. Statistics in Medicine, 35, 840–858.

Oslin, D.W. (2005) Managing alcoholism in people who do not respond to naltrexone (EXTEND). Available at: https://clinicaltrials.gov/ct2/show/NCT00115037.

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

Rush, A.J., Fava, M., Wisniewski, S.R., Lavori, P.W., Trivedi, M.H., Sackeim, H.A., et al. (2004) Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Controlled Clinical Trials, 25, 119–142.

Schneider, L.S., Ismail, M.S., Dagerman, K., Davis, S., Olin, J., McManus, D., et al. (2003) Clinical antipsychotic trials of intervention effectiveness (CATIE): Alzheimer’s disease trial. Schizophrenia Bulletin, 29, 57–72.

Styan, G.P. (1970) Notes on the distribution of quadratic forms in singular normal variables. Biometrika, 57, 567–572.

Urach, S. and Posch, M. (2016) Multi-arm group sequential designs with a simultaneous stopping rule. Statistics in Medicine, 35, 5536–5550.

Wang, S.K. and Tsiatis, A.A. (1987) Approximately optimal one-parameter boundaries for group sequential trials. Biometrics, 43(1), 193–199.

Wason, J.M. and Jaki, T. (2012) Optimal design of multi-arm multi-stage trials. Statistics in Medicine, 31, 4269–4279.

Wei, B., Braun, T.M., Tamura, R.N. and Kidwell, K.M. (2018) A Bayesian analysis of small n sequential multiple assignment randomized trials (snSMARTs). Statistics in Medicine, 37, 3723–3732.
Whitehead, J. (1992) Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials*, 13, 106–121.
Zhong, X. (2018) Design and analysis of sequential multiple assignment randomized trial for comparing multiple adaptive interventions. PhD thesis, Mailman School of Public Health, Columbia University.

**SUPPORTING INFORMATION**
Web Appendices referenced in Sections 2, 3, 5, and source code are available with this paper at the Biometrics website on Wiley Online Library.

**How to cite this article:** Wu L, Wang J, Wahed AS. (2023) Interim monitoring in sequential multiple assignment randomized trials. *Biometrics*, 79, 368–380. https://doi.org/10.1111/biom.13562