Published in final edited form as:

Clin Trials. 2010 ; 7(6): 643–652. doi:10.1177/1740774510376418.

Sample Size Calculations for Evaluating Treatment Policies in Multi-stage Designs

Ree Dawson, Ph.D. and

Frontier Science Technology and Research Foundation, 900 Commonwealth Ave., Boston MA 02215, U.S.A

Philip W. Lavori, Ph.D.

Department of Health Research and Policy, M/C 5405, Stanford CA 94305, U.S.A

Abstract

Background—Sequential Multiple Assignment Randomized (SMAR) designs are used to evaluate treatment policies, also known as adaptive treatment strategies (ATS). The determination of SMAR sample sizes is challenging because of the sequential and adaptive nature of ATS, and the multi-stage randomized assignment used to evaluate them.

Purpose—We derive sample size formulae appropriate for the nested structure of successive SMAR randomizations. This nesting gives rise to ATS that have overlapping data, and hence between-strategy covariance. We focus on the case when covariance is substantial enough to reduce sample size through improved inferential efficiency.

Methods—Our design calculations draw upon two distinct methodologies for SMAR trials, using the equality of the optimal semi-parametric and Bayesian predictive estimators of standard error. This ‘hybrid’ approach produces a generalization of the t-test power calculation that is carried out in terms of effect size and regression quantities familiar to the trialist.

Results—Simulation studies support the reasonableness of underlying assumptions, as well as the adequacy of the approximation to between-strategy covariance when it is substantial. Investigation of the sensitivity of formulae to misspecification shows that the greatest influence is due to changes in effect size, which is an *a priori* clinical judgment on the part of the trialist.

Limitations—We have restricted simulation investigation to SMAR studies of two and three stages, although the methods are fully general in that they apply to ‘K-stage’ trials.

Conclusions—Practical guidance is needed to allow the trialist to size a SMAR design using the derived methods. To this end, we define ATS to be ‘distinct’ when they differ by at least the (minimal) size of effect deemed to be clinically relevant. Simulation results suggest that the number of subjects needed to distinguish distinct strategies will be significantly reduced by adjustment for covariance only when small effects are of interest.

Keywords

Treatment policies; Adaptive treatment strategies; Multi-stage designs; sample size formulae

Introduction

Growing interest in the development of individualized treatments has led to a new generation of clinical trials and methodology aimed at evaluating treatment policies with full experimental rigor. The clinical trials are designed in accord with the sequential decision making underlying the policies. In these studies, successive courses of treatment are randomly and adaptively assigned over time, according to the individual subject's treatment and response history; each stage of randomization corresponds to a decision stage of the dynamic treatment policies under evaluation. Multi-stage randomization trials have been used in fields as diverse as psychiatric, cancer and AIDS research, and as early as the 1980s, for the evaluation of oncology treatment policies; see [1]. In the recent statistical literature, treatment policies have been described as adaptive treatment strategies (ATS) because treatment changes are tailored to the circumstances of the individual, including response to prior treatments [2]; the multi-stage randomization designs for evaluating ATS have been described as sequential, multiple assignment, randomized (SMAR) designs [3]. We adopt both terminologies in this paper.

To make ideas concrete, consider a two-stage generic ATS that specifies 'Start on medication A , change treatment to B if the patient's symptoms continue under A , otherwise maintain on C ', where A , B , and C are fixed for a particular strategy. Figure 1 depicts the SMAR trial to evaluate competitive choices for A , B , and C , using the state S to indicate whether symptoms persisted under A .

The SMAR structure illustrated in Figure 1 is well suited for the development of whole strategies. In particular, the design enables detection of interactive effects in treatment sequences that may be overlooked by the 'myopic' approach that evaluates each treatment decision with a separate single-stage trial [4]. For example, the medication A may be superior at symptom reduction (measured by S), but the optimal two-stage ATS may start with the alternative A^* because it enhances the effects of particular secondary treatments [3]. Generally, SMAR treatment alternatives are fully nested over time, according to clinical equipoise, to insure that the design encompasses the complete set of ATS relevant to determining the best strategy. This property distinguishes the SMAR design from those that randomize subjects at baseline to a subset of the strategies that would be determined sequentially by equipoise. For either approach, it is possible to evaluate the effect of initial (or a later stage) treatment by comparing whole ATS in the study that begin with different treatments [3,5]. The use of ATS for this purpose insures against fallacies associated with the myopic approach. In this paper, we focus on the comparison of whole strategies, as this is typically the primary goal of a SMAR trial.

Much of the development for SMAR trials has centered on methods for inference [6-8]. For reasons described above, evaluation is in terms of a 'final' outcome (measured after the last decision stage) that takes into account the sequence of intervening outcomes due to successively applied treatments [3]. Improvements to estimators have been proposed to provide gains in efficiency [9-12]. Efficiency is of particular importance to SMAR trials for two reasons. First, the original sample splits randomly and adaptively at each stage of the design, with subjects nested within strata defined by previous treatment and response history. Hence, the sequence of randomizations creates a monotone pattern of *ignorable* missingness for each ATS in the study [6]; improved efficiency helps address the resulting loss of statistical power. Second, the nested structure of treatment assignment gives rise to ATS that have overlapping treatments, e.g., two or more strategies may (adaptively) specify the same initial treatment [6]. The greater the overlap there is for pairs of strategies, the greater likelihood there is for diminished causal effects. Hence if two strategies with much overlap are to be compared, it is important to size the trial efficiently.

Less attention has been given to design methodology. A significant challenge is the development of methods for sample size determination because of the sequential and adaptive nature of both the strategies under study and the treatment assignment mechanism used to assign subjects to the ATS. Consider the generic ATS, and assume that the trial uses a final outcome Y , which summarizes symptom severity over time, to evaluate alternatives for A , B , and C . The unknown population parameters required for sample size calculations not only include those for the distribution of Y , but also for the distribution of the intervening state S . In particular, because the second randomization is adaptive, the distribution of Y will be stratified according to the *unknown* distribution of S (the responses to A). Adding another decision stage further splits the stratification, leading to many more unknowns than commonly needed for sample size calculations, as well as greater uncertainty about their values.

For specific contexts, simulation has been used to evaluate design requirements [7,13]. A sample size formula has also been developed for two-stage trials with time to event endpoints when two strategies specify the same initial treatment [14]. To provide a general method for trials with continuous outcomes, Murphy appeals to ‘working’ assumptions about treatment assignment to circumvent the need to specify parameters for the unknown distribution of state (response) history when determining the number of subjects for a SMAR trial. This allows a simple upper bound on the variance of the estimated mean for the ATS to be derived solely in terms of the population ‘within-strategy’ variance of Y ; simulations show that the sample size formula based on the bound may be conservative when treatments are assigned adaptively in terms of prior state history [3]. The approach has been extended to cover a range of research questions that can be addressed by the SMAR design [5]. As in Murphy’s original paper, the test statistics used to size SMAR trials to specific questions assume that strategy means have been estimated with semi-parametric marginal means (MM) models [15]. However, the MM variance estimator has been shown to be upwardly biased for clinically realistic scenarios, which may lead to excessive number of study subjects when used in sample size formulae [6,10].

Recently, Dawson and Lavori [16] developed an alternative method for determining sample sizes for SMAR trials with continuous outcomes, using *optimal* (efficient) semi-parametric theory as a basis for those calculations. The efficiency gains provided by optimal estimation effectively eliminate the problem of bias with the MM variance estimator. Moreover, by appealing to assumptions typical of design calculations, but adapted to the SMAR setting, the population within-strategy variance can be expressed in terms of effect size and regression quantities familiar to the trialist. This leads to a simple sample size formula for pairwise comparisons based on the t-test, and eliminates the need to specify intervening response rates or to assume worst case scenarios. Nonetheless, a key practical issue remains for the trialist intending to size a SMAR trial, which arises because of the nested nature of the randomization. Specifically, any overlap between a pair of ATS (created by sequential treatment assignment) not only diminishes their causal difference, but also introduces positive between-strategy covariance. Hence causal differences are getting smaller, while inference (that takes into account covariance) is getting more efficient. It is unclear what impact this has on observable effect sizes, and how design calculations should be carried out to efficiently size SMAR trials. In this article, we consider the role of covariance, and equivalently of overlap, in determining sample sizes for multi-stage designs.

Realization of a SMAR design

A tree is a canonic way to depict the nested structure of a SMAR realization (see Figure 1). For the simplest design, the first stage of randomization is represented by the initial node that divides into branches, one for each possible treatment option. Because the next

randomization is adaptive, each ‘treatment’ branch subdivides into further branches, one for each potential subsequent response. The tree continues to subdivide in the same way for successive stages of the design, along the treatment-response branch paths defined by previous randomizations. The set of strategies under evaluation determines the alternative treatment options and how they are adaptively applied, and consequently the SMAR tree or design. For example, the trialist would ‘dial’ in different choices for the placeholders A , B , and C in the example ATS in the Introduction, according to the strategies of *a priori* interest, which in turn would specify the treatment alternatives for a two-stage SMAR study. As for fixed treatment trials, the choices for A should satisfy clinical equipoise. The choices for B and C could vary according to the choices for A , and would satisfy *sequential* equipoise for the adaptive use of B and C , following the initial use of a particular choice for A .

For purposes of design and inference, the multi-stage design can be described sequentially in terms of the adaptive randomized treatment assignments. Consider a SMAR trial with three stages of randomization. Let S_1 be the baseline state of the subject, and let A_1 be initial treatment randomly assigned on the basis of S_1 . For stage two, let S_2 be the status of the subject measured at the start of the second stage and A_2 the treatment assigned by the second randomization according to $S_2 = (S_1, S_2)$ and A_1 . Analogously, S_3 is measured at the start of the third stage and A_3 is assigned by the last randomization according to $S_3 = (S_1, S_2, S_3)$ and $A_2 = (A_1, A_2)$. The sequence of randomizations can be expressed as (sequential) assignment to alternative decision rules, each of which determines treatment for the next stage of the study as a function of the state and treatment history to date. We write $d_k(S_k, A_{k-1})$ for the decision rule d_k that adaptively ‘outputs’ the choice for A_k ; the randomization probabilities for d_k , denoted $\{p_k(d_k|S_k, A_{k-1})\}$, are experimentally fixed by the design according to prior state-treatment history.

For a three-stage SMAR trial, each observable sequence $\{d_1, d_2, d_3\}$ corresponds to an ATS, which we denote as $\mathbf{d}$, if the domain for each successive decision rule includes the state-treatment histories produced by previous rules in the sequence. This condition insures that the ATS is a well-defined policy for adaptively determining the ‘next’ treatment. A two-stage trial similarly evaluates sequences $\{d_1, d_2\}$ that conform to ATS. For example, the generic ATS in the introduction consists of two decision rules: $d_1(s_1) \equiv d_1$ specifies the initial medication A , $d_2(S = 1, A)$ adaptively specifies the ‘switch’ medication B when symptoms continue, according to the choice for initial medication, where $A \equiv A_1$ and $S \equiv S_2$ is an indicator of failure to remit; $d_2(S = 0, A)$ analogously specifies the ‘maintenance’ medication C when symptoms abate. Note that alternative decision rules, and hence ATS, are determined by the choices for A , B , C that are ‘dialed’ in by the trialist.

To evaluate competitive ATS, the SMAR design includes a primary outcome Y , obtained after the last stage of randomization, assumed in this paper to be continuous. We judge the performance of an ATS $\mathbf{d}$ by $\mu_{\mathbf{d}}$, the population mean of Y that would be observed if all subjects are treated according to $\mathbf{d}$.

Basis for sample size formulae

Under appropriate experimental conditions detailed in the Appendix (e.g., sequentially blocked randomization), the optimal semi-parametric estimator of $\mu_{\mathbf{d}}$ obtained from SMAR data, denoted $\hat{\mu}_{\mathbf{d}}$, and the estimator for the variance of $\hat{\mu}_{\mathbf{d}}$ are equal to the analogous (Bayesian) predictive estimators [16]. We draw upon both methodologies to develop a hybrid approach to SMAR design calculations that accounts for between-strategy covariance.

The estimator of strategy means on which the sample size formulae are based can be expressed in terms of stage-specific, stratified quantities as:

$$\hat{\mu}_d = \sum_{s_3} \varphi_3(s_3) m_3(s_3) \quad (1)$$

where $m_3(s_3)$ is the sample mean of final responses among subjects sequentially randomized to d through the final stage and having state values $S_3 = s_3 \equiv (s_1, s_2, s_3)$; the sample mean $m_3(s_3)$ is weighted by the proportion of subjects with state history s_3 obtained under d :

$$\varphi_3(s_3) = \prod_{k=1}^3 f_k(s_k) \quad (2)$$

where $f_k(s_k)$ is the sample conditional response rate for $S_k = s_k$, given assignment to d through $k-1$ and $S_{k-1} = s_{k-1}$. The estimator (1) is a version of Robin's G-Computational Algorithm, derived through iterated conditional expectations [4,6,10]. It can be viewed as sequential non-parametric adjustment for the intervening states that determine missingness created by the SMAR randomizations, thereby guaranteeing consistency. The estimator for the variance of $\hat{\mu}_d$ can be obtained non-parametrically from $V(U_{opt})$, the variance of the most efficient influence function U_{opt} , in order to achieve the semi-parametric efficiency bound

[16]. In this case, the within-strategy standard error is equal to $\sqrt{\hat{v}_{opt}}$, where $\hat{v}_{opt} = n^{-1} \hat{V}(U_{opt})$ and n is the number of subjects in the trial.

It is possible to derive a formula for population within-strategy variance suitable for design calculations, using $V(U_{opt})$ and a simplifying assumption [16]. At each stage, consider the residual obtained by the regression of the final outcome Y_d (obtained under d) on the earlier states. We assume the residual variance at each stage is homogeneous across state history. This condition is a sequential version of the type of homogeneity of variance assumption often used for sample size determination. When subjects are allocated with equal probability to treatment alternatives, which themselves are equal in number at every decision point of stage k , the randomization probabilities $p_k(d_k|S_k, A_{k-1}) \equiv p_k$, and $V(U_{opt})$ can be re-expressed as:

$$\sigma_Y^2 P_3^{-1} [1 - (1 - p_1 p_2 p_3) R_1^2 - (1 - p_2 p_3) R_2^2 - (1 - p_3) R_3^2] \quad (3)$$

where σ_Y^2 is the variance of Y_d , $P_3 = p_1 p_2 p_3$, $R_r^2 = (1 - \sigma_r^2 / \sigma_Y^2)$ is the coefficient of determination for the regression of Y_d on the states $S_{d,3}$ (obtained under d), and R_k^2 denotes the increment in the coefficient of determination when $S_{d,k}$ is added to the regression of Y_d on $S_{d,k-1}$.

We refer to the multiplier of σ_Y^2 in (3) as the 'variance inflation factor' (VIF) due to the SMAR design. It accounts for the loss of precision due to missingness created by successive randomizations, relative to a trial that would allocate all subjects to d . It also makes explicit the efficiency gains due to semi-parametric optimality, as the first term in (3) corresponds to the MM variance estimator [16].

When randomization probabilities for d depend on state history, (3) becomes:

$$\sigma_y^2 [E_d(P_3^{-1}) - E_d((1-p_1p_2p_3)P_3^{-1})R_1^2 - E_d((1-p_2p_3)P_3^{-1})R_2^2 - E_d((1-p_3)P_3^{-1})R_3^2] \quad (4)$$

where the expectation $E_d()$ is calculated under the distribution of S_3 when all treatments are assigned according to the strategy d [16].

Sample size: Pairwise comparison with no overlap

To illustrate SMAR sample size calculations, we consider the two-stage ATS in the Introduction, and suppose there is interest in comparing two versions, denoted d and d^* , which specify different choices for the initial medication A. In this case, the first decision rule is not adaptive, and the two strategies do not share any baseline data in a SMAR trial (see Figure 1). Furthermore, once distinct initial treatments are specified, the two ATS as a whole are distinct. Consequently, there isn't any overlap between the two strategies, and no between-strategy covariance to consider when testing the null hypothesis that $\mu_d = \mu_{d^*}$.

Let ES be the standardized effect size of interest specified by the trialist. To generalize the usual t-test formula for sample size, we pool the marginal outcome variance of d and d^* to obtain σ_p^2 ; consequently $ES = \frac{\Delta}{\sigma_p}$, where $\Delta = \mu_d - \mu_{d^*}$ is the difference in population means. Similarly, let VIF_p be the 'pooled' (averaged) variance inflation factor, where the VIF in (3) reduces to:

$$P_2^{-1} [1 - (1 - p_2)R_T^2]$$

in the absence of using a baseline state for the first randomization. For illustration, set the nominal level of power to be achieved to 0.80, with the level of the test = 0.05. Then the sample size is calculated as $7.9 * 2 * \frac{VIF_p}{ES^2}$. If the final outcome is strongly related to the intermediate state S_2 for both strategies, e.g., $R_T^2 = 0.7$ for d and d^* , then $VIF_p = 4[1 - 0.5(0.7)] = 2.6$, noting that $p_1 = p_2 = 1/2$ in Figure 1. Adding another stage to the trial increases VIF_p to $8[1 - 0.75(0.2) - 0.5(0.5)] = 4.8$, assuming $R_2^2 = 0.2$, $R_3^2 = 0.5$, $p_3 = 1/2$.

Sample size: Pairwise comparison with overlap

The t-test generalization ignores the role of between-strategy covariance in SMAR inference, which may lead to an excessive number of subjects. To derive an adjustment to sample size when covariance is likely to be substantial, we consider the case when two ATS agree except for the last decision rule. As detailed in the Appendix, we make assumptions that imply $\mu_d = \mu_{d^*}$, despite differences in final treatment, in order to approximate between-strategy covariance. Under this scenario, the reduced sample size becomes:

$$n^* = n(1 - R_p^2) \frac{P_3^{-1}}{VIF_p} ; \quad P_3 = \prod_{k=1}^3 p_k \quad (5)$$

where n is the sample size calculated without regard to covariance, and R_p^2 is the mean of the R_T^2 values for the two strategies. The adjustment factor $\frac{n^*}{n}$ approximates the ratio of the

(pooled) variance that takes into account between-strategy covariance, to the (pooled) variance that does not, as desired.

The assumed scenario is generally not realistic, and will be ‘anti-conservative’, in that the reduced sample size will be too small, when causal differences are constant conditional on state history S_3 . To improve the adjustment (5), we use ES^2 as a crude upper bound for the relative error in total variance due to the covariance approximation; this gives $n^{**} = n^* + ES^2n$. The underlying rationale for the correction suggests that n^{**} will be conservative, i.e., provide more than the nominal level of power for the pairwise comparison. See the Appendix for a complete exposition.

As a simple illustration of the adjustments for between-strategy covariance, assume now that d and d^* of the previous section specify the same initial treatment. Then $n^* = n(1 - 0.7)4/2.6 = .46n$, and $n^{**} = (ES^2 + 0.46)n$. Similarly, $n^{**} = (ES^2 + 0.5)n$ if a third stage is added to the trial.

Sample size: SMAR trial set up

To develop a strategy for determining SMAR sample size requirements, we posit the following set up. The trialist specifies *a priori* the *ES* of clinical relevance, i.e., effects smaller than that are not worth detecting. The appropriate sample size is one that insures (i) any pairwise comparison arising from the trial will be fully powered if effects are at least *ES*; (ii) resources are not ‘wasted’ on comparisons smaller than *ES*. Conceptually, it's useful to think of pairs of strategies as either distinct (having effects at least *ES*) or not; the required sample size will be the maximum needed for any comparison of distinct strategies.

As discussed in the Introduction, a key question is what role between-strategy covariance might play in sample size calculations. Any overlap in treatment differences when *ES* is sufficiently large may preclude a ‘distinct’ causal difference, and covariance can likely be ignored. However, for small enough effect sizes, two strategies could be distinct, despite common treatments; in this case, covariance might be substantial enough to require consideration. To make this concrete, consider a SMAR trial with K stages in which treatment under an ATS is either uniformly effective or uniformly ineffective across state histories, at any particular stage of the study. For two strategies in the trial, let δ (formally a function of the pair) be the stage at which their treatments diverge. For example, $\delta = 1$ implies there are no common treatments for a pair of ATS; $\delta = K$ implies all but the last treatments are common. Assume now that as δ varies from 1 to K , there are at least two strategies divergent at δ , such that one is uniformly effective thereafter, while the other is uniformly ineffective. This condition insures that the potential for distinct strategies is as great as possible, even as pairwise covariances increase in magnitude. As constructed, the choice of *ES* determines the ‘ δ threshold’ after which strategies fail to be distinct.

Simulations

We conducted simulations to evaluate the proposed formulae for sample size requirements, to assess the role of covariance in design calculations, and to understand the impact of misspecification. To operationalize the SMAR trial described above, we adopted a variant on the simulation scheme used to evaluate the performance of predictive and semi-parametric estimators [6,10]. The central feature of the scheme is a transition matrix (TM), which is used to generate states over time for a particular strategy; the distribution of strategy-specific states is governed by the choice of entries for TM. To explicitly allow for simulated effects due to the final treatment A_K (for a K -stage trial), we include a ‘final’ state S_{K+1} , not necessarily measured during a real trial. The outcome for strategy d is generated as

a regression on state history, with normal error: $Y_d = \mathbf{S}_{d,K+1}^T \beta + e$, $e \sim N(0, \sigma_e^2)$. Hence, the assumption that residual variance is homogeneous across state history fails at each stage. The vector of regression coefficients β is the same for each strategy, thereby insuring that all observables follow a consistent population regression model. Thus, differences in the transition matrices give rise to causal differences underlying ES .

We set $K = 2, 3$; β is varied to generate different scenarios for R_k^2 and R_T^2 values. In all cases, $\sigma_e^2 = 1$ and the intercept $\beta_0 = 0.5$, which is the coefficient for $S_0 \equiv 1$. States are assumed to be binary at each stage of the SMAR trial: $S_k = 1, 2$, with the higher value indicative of poorer response, and equiprobable values for baseline S_1 . The transition matrices TM_G and TM_B represent, respectively, hypothetical but plausible effective ('good') and ineffective ('bad') stage-specific responses:

$$\text{TM}_G = \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix} \quad \text{TM}_B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}$$

where $\text{TM}_{ij} = \Pr(j|i)$ for $\text{TM} = \text{TM}_G, \text{TM}_B$. We label strategies according to the sequence of transition matrices used to generate the outcomes (including S_{K+1}). For example, GGG describes a strategy with a tendency toward good responses at all three stages for $K = 3$, while GBB describes a strategy that has a tendency toward bad responses for the last two stages. As long as two strategies agree on their treatments, they must have the same TM. We can thus consider pairs of strategies whose differences in TM sequence is due to the divergence of treatments. Accordingly, GGG and GBB implies $\delta = 1$; GGG and GGB implies $\delta = 2$. Analogously for $K = 2$: GG and GB have common treatments through stage 1, but treatments diverge thereafter ($\delta = 1$).

Performance of sample size formulae

Table 1 summarizes 1000 replications of each of the scenarios. The results show that the original sample size formula (ignoring covariance) achieves the nominal level of power of 0.80 for $\delta < K$, but is overly conservative when strategies diverge at the last point; see columns 1-3 for $K = 3$ and columns 1-2 for $K = 2$. Table 1 also confirms that the approximation to covariance derived for the case $\delta = K$ (GGG vs. GGB; GG vs. GB) used by n^* is 'anti-conservative', as expected. However, the additional adjustment to sample size provided by n^{**} is reasonable, and with two exceptions is conservative but within the simulation standard error of 0.013.

Some of the scenarios in Table 1 were modified by setting $\sigma_e = 2^{1/2}$ or 2, or by using TM_B prior to δ ; similar results to those reported here were obtained for unadjusted and adjusted sample sizes.

Sensitivity of sample size formulae

The performance of the sample size formulae detailed in Table 1 depends on the trialist knowing the 'true' population quantities underlying the simulations. To study the impact of misspecification error, we selected pairs of regression coefficients from Table 1, one assumed to be true, designated β , and one assumed to be in error, designated β_e . Table 2 shows the changes in required sample size when calculations use β_e instead of β . The corresponding changes in VIF ($\mathbf{VIF}$ vs. $\mathbf{VIF}_e$) and ES ($\mathbf{ES}$ vs. $\mathbf{ES}_e$) are also presented in order to assess the sensitivity of sample size to their misspecification, implicitly via β_e .

Rows 4-6, 9-10 show substantial changes in VIF due to ‘error’, and that the calculated sample sizes are excessive when VIF_e is specified instead of **VIF**. In each case, the changes in VIF are due to differences in the R_T^2 values for the two coefficient vectors, rather than to misspecification of the individual R_k^2 ; forcing equality of the R_T^2 values, while keeping the proportionate increments $\{R_k^2/R_T^2\}$ unchanged, makes the VIF ratios in those rows approximately equal to one (thereby eliminating any effects due to error). Table 2 also shows that sample size is sensitive to small changes in ES when effects are small, and that the impact is greater than that due to misspecified R_T^2 values.

The results in Table 2 suggest that sample size is somewhat robust to misspecification errors in the $\{R_k^2/R_T^2\}$. For example, the proportionate increments generated for the strategy BB ($K = 2, \delta = 1$) are (0.43, 0.57) when $\beta = (1, 1, 3)$, and (0.24, 0.76) when $\beta_e = (1, 2, 2)$. Yet once R_T^2 values are equalized, while keeping $\{R_k^2/R_T^2\}$ fixed, the VIF ratio is approximately one; there is no influence on calculated sample size due to halving the ‘true’ R_1^2/R_T^2 value of 0.43 to the ‘erroneous’ value of 0.24. Expression (3) provides some insight: earlier increments have smaller multipliers because of how randomization probabilities enter into the VIF. Moreover, earlier increments are likely smaller given that later treatment would typically be more important to the final outcome, as constructed here. This suggests an intrinsic robustness, even though the trialist may be more uncertain about the relative importance of early treatment.

Less obvious to discern from Table 2 is the influence of the last treatment A_K (which produces outcome S_{K+1} in the simulations) on misspecification error via R_T^2 . The large VIF ratios in rows 4-6, 9-10, discussed above, occur primarily because of the differential effects due to A_K that occurs in the second and fourth (β, β_e) pair (compare the regression vectors in terms of their last two coefficients). Diminished effects generated by β_e for the last treatment imply that previous treatment history has a relatively stronger relationship to the outcome Y ; hence the R_T^2 and VIF values for the β_e vectors are higher than those for the corresponding β vectors. The import for the trialist is to have sources for plausible R_T^2 values, or be willing to ‘bet’ that the last treatment has pronounced effect.

Sample size: SMAR trial with unequal randomization probabilities

The simulation set up provides a way to calculate sample sizes when randomization probabilities depend on the history of the individual subject. For example, suppose that the two-stage ATS in the Introduction keeps patients on initial treatment if symptoms abate; subjects who remit will not be randomized again in the SMAR trial. Assume further that the trialist sets ES to be moderate, and uses the comparison of strategies GG to BG to size the study, where the second ‘G’ of GG or BG only applies to subjects requiring a second randomization (i.e., $S_2 = 2$ for an individual subject). In this case, calculations are similar to previous ones for comparing d and d^* , except that (4) must be used instead of (3). For this example, $E_d(P_2^{-1})$ is:

$$E_d(P_2^{-1}) = 2\Pr[S_2=(1, 1)] + 4\Pr[S_2=(1, 2)] + 2\Pr[S_2=(2, 1)] + 4\Pr[S_2=(2, 2)] \quad (6)$$

noting that $P_2^{-1}=2$ for a single randomization. For GG, $E_d(P_2^{-1})=2(0.35)+4(0.15)+2(0.25)+4(0.25)=2.74$, and similarly $E_d(P_2^{-1})=2.7$ for BG, given equiprobable S_1 . Also, $E_d((1-p_2)P_2^{-1})=2(0.15)+2(0.25)=0.8$ and 0.9 for GG and BG, respectively, since $1-p_2=0$ if $S_2=1$. Assuming $R_t^2=0.7$ as before, $VIF_p=2.13$, compared to 2.6 when $p_2 \equiv 1/2$.

Analogous calculations could be performed for when ES is deemed to be large (e.g., GG vs. BB) or small (e.g., GG vs. GB), taking into account between-strategy covariance for the latter case. For a particular specification for ES , simulation could be used to find the pair of ATS with the smallest ‘distinct’ effect, in order to determine the required sample size.

Discussion

In this paper, we draw upon two distinct methodologies for SMAR trials to develop a hybrid approach to design calculations. The simulation results support the reasonableness of the sequential homogeneity of variance assumption used to map the population optimal variance into familiar regression quantities, as well the adequacy of the approximation to between-strategy covariance when overlap between strategies is substantial. The regression framework for the calculations allows the trialist to ‘guess’ or specify the strength of the relationship of final outcomes to state history, as well as the relative importance of each stage of the study, in order to determine sample sizes for pairwise comparisons. Furthermore, the simulation set up itself can be used by the trialist to ‘firm up’ guesses for variance inflation factors and effect sizes, as well as gauge whether adjustment for between-strategy covariance is needed, using clinical appraisal of what constitutes a ‘good’ and ‘not so good’ response to treatment.

We also used simulation to investigate the sensitivity of sample size formulae to misspecification. The results show that the greatest influence is due to ES , which is an *a priori* clinical judgment on the part of the trialist. The results further suggest an intrinsic robustness to misspecifying regression quantities, particularly the relative contribution of each SMAR stage to the variance inflation factor. Moreover, sensitivity to specification of R_t^2 is diminished when the last stage of treatment has a pronounced effect on final outcome, in comparison to earlier stages, as might typify SMAR realizations.

We designate strategies as distinct or not, according to an *a priori* specified effect size of clinical relevance, to develop a strategy for determining sample sizes reflective of the SMAR design. Patients are randomized sequentially to a set of nested treatment options, which not only gives rise to a multiplicity of strategies to be evaluated, but also insures ‘partial’ redundancy among competitors (induced by common treatment history). In addition, differences among ATS are likely to be attenuated by their own sequential structure, e.g., early gains in efficacy are prone to erosion over time. The significance of this to the trialist is that many strategies may be ‘nearly’ as effective (or not), and that it is necessary to define ‘nearly’ in a clinically meaningful way before the start of the study. The notion of effect size offers one way to specify a neighborhood of indifference for SMAR trials that is accessible to clinicians and researchers alike. Our simulation results suggest that the number of subjects needed to distinguish distinct strategies will be significantly reduced by adjustment for covariance only when small effects are of interest.

We remark that presentation has been limited to two- and three-stage SMAR designs, because of the practical focus of the paper. All methodological results extend readily to the K -stage study [16].

Acknowledgments

Supported by National Institute of Mental Health Grant No. R01-MH51481 to Stanford University.

Appendix

Throughout, we assume the observed proportions assigned to treatments at each stage coincide with the fixed randomization probabilities of the design; such coincidence occurs asymptotically by the law of large numbers and might be achieved in a study using sequentially blocked randomization [6,10]. When this condition holds, the predictive and optimal semi-parametric estimators of $\hat{\mu}_d$ and its standard error are equal [16].

To adjust design calculations for between-strategy covariance, we decompose the predictive variance for $\hat{\mu}_d$, denoted $\hat{v}_{PR}$ into ‘naïve’ and ‘penalty’ components of variance:

$$\hat{v}_{PR} = \hat{v}_n + \hat{v}_p$$

where (suppressing dependence on state history)

$$\hat{v}_n = \sum_{s_3} \varphi_3^2 \hat{v}(m_3); \quad \hat{v}_p = \sum_{s_3} m_3 m_3' \widehat{\text{cov}}(\varphi_3, \varphi_3); \quad (7)$$

and $\hat{v}(m_3) \equiv \hat{v}(m_3(s_3))$ is the sample variance of $m_3(s_3)$. (When the argument s_3 is suppressed, we mark the function instead, e.g., $\varphi_3 = \varphi_3(s_3)$.) The first term $\hat{v}_n$ is the ‘naïve’ variance estimate that assumes the coefficients of $m_3(s_3)$ in (1) are known *a priori*, and $\hat{v}_p$ is the ‘penalty’ paid for estimating them via (2) [6,10].

Let v_p be the population counterpart to $\hat{v}_p$. Assume (i) A_2 is the same for a pair of strategies, given state history S_2 , and (ii) for each value of s_3 , the population counterpart to $m_3(s_3)$ denoted $\mu_3(s_3)$ is the same for both strategies. Then the ‘penalty’ v_p is equal to the population between-strategy covariance, which we denote as *cov* [6]. Furthermore, given the assumed setup for design calculations, the ‘within’ quantities for each strategy of the pair satisfy:

$$\frac{v_{opt} - v_p}{v_{opt}} = \frac{\sigma_K^2 P_3^{-1}}{\sigma_Y^2 VIF} \quad (8)$$

where v_{opt} is the population optimal within-strategy variance. (Note that nP_3 times $\hat{v}_n$ is the non-parametric alternative to the estimated residual regression variance; the population counterparts are the same.) This provides the basis for n^* in (5). Because the final treatment is assumed to differ for at least some subjects, agreement on all $\mu_3(s_3)$, will usually fail, in which case v_p provides an approximation to covariance. Accordingly, $\frac{v_{opt} - v_p}{v_{opt}}$ approximates the ratio of the (pooled) variance that takes into account between-strategy covariance, to the (pooled) variance that does not. Both (8) and (5) generalize when randomization probabilities depend on individual history, by replacing P_3^{-1} with $E_d[P_3^{-1}]$ and using (4) for *VIF*.

Assumption (ii) is generally not realistic, as it implies with (i) that the population means of the two strategies agree, despite differences in final treatment. Furthermore, if the second assumption fails to hold *and* causal effects are constant across subgroups indexed by state history, the adjustment is ‘anti-conservative’ in that the reduced sample size will be too small. This occurs because the between-strategy covariance will be overstated by the amount

$\Delta^2 \sum_{s_3, s_3'} cv_3$, where $cv_3 \equiv cv_3(s_3, s_3')$ is the population counterpart to $\widehat{cov}(\varphi_3, \varphi_3') \equiv \widehat{cov}(\varphi_3(s_3), \varphi_3(s_3'))$; as before, Δ is the difference in population means. To see this, fix s_3, s_3' and let μ_3, μ_3' be the corresponding population subgroup means. The term corresponding to s_3, s_3' in $v_p + \tilde{v}_p - 2cov$ is:

$$cv_3 (\mu_3 \mu_3' + \tilde{\mu}_3 \tilde{\mu}_3' - 2\mu_3 \tilde{\mu}_3') = cv_3 \Delta(\mu_3 - \tilde{\mu}_3') \quad (9)$$

where we use $\tilde{\cdot}$ to distinguish strategy-specific quantities. If $s_3 = s_3'$, then (9) is equal to $cv_3 \Delta^2(\mu_3 - \tilde{\mu}_3' \equiv \Delta$ by assumption). Otherwise, it is necessary to pair terms for (s_3, s_3') and (s_3', s_3) to obtain the result, noting that $\Delta(\mu_3 - \tilde{\mu}_3') + \Delta(\mu_3' - \tilde{\mu}_3) = 2\Delta^2$.

To get n^{**} , we use ES^2 as a ‘crude’ upper bound for the relative error in total variance due to the covariance approximation:

$$\frac{\Delta^2 \sum_{s_3, s_3'} cv_3}{v_{OPT,p}} = \frac{\Delta^2 \sum_{s_3, s_3'} cv_3}{\sigma_p^2 VIF_p} = ES^2 \frac{\sum_{s_3, s_3'} cv_3}{VIF_p}$$

where $v_{OPT,p}$ is obtained by pooling v_{OPT} across strategies. The intuition is that the covariance due to estimating the $\phi_3(s_3)$ will be less than the variance inflation factor, as one pertains to the ‘penalty’ component of total variance and the other to total variance. This suggests that $n^{**} = n^* + ES^2 n$ will be conservative.

References

1. Dawson R, Green AI, Drake RE, McGlashan TH, Schanzer B, Lavori PW. Developing and testing adaptive treatment strategies using substance-induced psychosis as an example. *Psychopharmacology Bulletin* 2008;41:51–67. [PubMed: 18779776]
2. Lavori PW, Dawson R. Adaptive treatment strategies in chronic disease. *Annual Review of Medicine* 2008;59:443–453.
3. Murphy S. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 2005;24:1455–1481. [PubMed: 15586395]
4. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. *Clinical Trials* 2004;1:9–20. [PubMed: 16281458]
5. Oetting, AI.; Levy, JA.; Weiss, RD.; Murphy, SA. Statistical Methodology for a SMART Design in the Development of Adaptive Treatment Strategies. In: Shrout, PE., editor. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*. Arlington VA: American Psychiatric Publishing, Inc.; 2009.
6. Dawson R, Lavori PW. Sequential causal inference: Application to randomized trials of adaptive treatment strategies. *Statistics in Medicine* 2008;27:1626–45. [PubMed: 17914714]
7. Thall PF, Millikan R, Sung HG. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine* 2000;19:1011–1028. [PubMed: 10790677]

8. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* 2002;58:48–57. [PubMed: 11890326]
9. Bembom O, van der Laan MJ. Analyzing sequentially randomized trials based on causal effect models for realistic individualized treatment rules. *Statistics in Medicine* 2008;27:3689–3716. published online. [PubMed: 18407580]
10. Lavori PW, Dawson R. Improving the efficiency of estimation in randomized trials of adaptive treatment strategies. *Clinical Trials* 2007;4:297–308. [PubMed: 17848491]
11. Lokhnygina Y, Helterbrand JD. Cox regression methods for two-stage randomization designs. *Biometrics* 2007;63:422–428. [PubMed: 17425633]
12. Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* 2004;60:124–133. [PubMed: 15032782]
13. Wolbers M, Helterbrand JD. Two-stage randomization designs in drug development. *Statistics in Medicine*. 2009 published online.
14. Feng W, Wahed AS. Sample size for two-stage studies with maintenance therapy. *Statistics in Medicine* 2009;28:2028–2041. [PubMed: 19382105]
15. Murphy SM, van der Laan MJ, Robins JM. *Journal of the American Statistical Association* 2001;96:1410–1423. [PubMed: 20019887]
16. Dawson R, Lavori PW. Efficient design and inference for multi-stage randomized trials of individualized treatment policies. 2009 In submission.

Figure 1. Two-stage SMAR design for evaluating two-stage strategies. There are two options for initial medication, A and A^* . The alternatives for second medication depend on initial medication and response to it. For example, responders to A^* are randomized to either C^* or C^* with equal probability.

Table 1

Performance of Sample Size Formulas for Nominal Power = 0.80

K = 3									
GGG vs.	BBB	GBB	GBB	GBB	GBB	GBB	GBB	GBB	GBB
Formula used	n	n	n	n	n	n	n	n	n
$\beta = (1.2, 3.4)$									
Achieved power	0.808	0.815	0.892	0.745	0.785/0.825				
Sample size	359	595	1928	1361	1446/1585				
$\beta = (1.2, 3.3)$									
Achieved power	0.818	0.817	0.931	0.780	0.803				
Sample size	346	607	2678	1705	1782				
$\beta = (1.2, 2.4)$									
Achieved power	0.792	0.816	0.844	0.759	0.775/0.811				
Sample size	398	703	1711	1362	1453/1585				
$\beta = (1.2, 2.3)$									
Achieved power	0.797	0.812	0.888	0.786	0.803				
Sample size	380	720	2280	1680	1762				
K = 2									
GG vs.	BB	GB	GB	GB	GB	GB	GB	GB	GB
Formula used	n	n	n	n	n	n	n	n	n
$\beta = (1.2, 3)$									
Achieved power	0.796	0.892,	0.763	0.800					
Sample size	362	1121	887	936					
$\beta = (1.2, 2)$									
Achieved power	0.814	0.886	0.742	0.798					
Sample size	400	1580	1148	1194					
$\beta = (1.1, 3)$									
Achieved power	0.806	0.826	0.802	0.809					
Sample size	472	944	845	899					

Table 2

Impact of Misspecification Error on Calculated Sample Size

β vs. β_e	$\frac{n_e}{n}$ (ratio)	$\frac{VIF_e}{VIF}$ (ratio)	$\frac{ES^2}{ES_e^2}$ (ratio)
(1,2,3,4) vs. (1,2,2,3) , $\delta = 1$	380/359 (1.06)	5.31/5.47 (0.82)	(.49/.47) ² (1.09)
$\delta = 2$	720/595 (1.21)	5.27/5.44 (0.97)	(.38/.34) ² (1.25)
$\delta = 3$	2280/1928 (1.23)	5.21/5.38 (0.97)	(.21/.19) ² (1.22)
(1,2,3,3) vs. (1,2,2,4) , $\delta = 1$	398/346 (1.15)	5.80/4.99 (1.16)	(.48/.48) ² (1.00)
$\delta = 2$	703/607 (1.16)	5.77/4.98 (1.16)	(.36/.36) ² (1.00)
$\delta = 3$	1711/2678 (0.69)	5.73/4.90 (1.17)	(.17/.23) ² (0.55)
(1,2,3) vs. (1,2,2) , $\delta = 1$	400/362 (1.10)	2.93/3.14 (0.93)	(.37/.34) ² (1.18)
$\delta = 2$	1580/1121 (1.41)	2.89/3.13 (0.92)	(.21/.17) ² (1.52)
(1,2,2) vs. (1,1,3) , $\delta = 1$	472/400 (1.18)	3.45/2.93 (1.18)	(.34/.34) ² (1.00)
$\delta = 2$	944/1580 (0.60)	3.44/2.89 (1.19)	(.17/.24) ² (0.52)