

Statistical Methodology for a SMART Design  
in the Development of Adaptive Treatment Strategies

Technical Report Series #07-82  
The Methodology Center, The Pennsylvania State University

Alena I. Scott  
University of Michigan  
Institute for Social Research  
426 Thompson St.  
Ann Arbor, MI 48106-1248

&

Janet A. Levy  
Center for Clinical Trials Network  
National Institute on Drug Abuse  
6001 Executive Blvd. Mail Stop 9557  
Bethesda, Maryland 20892

&

Susan A. Murphy  
University of Michigan  
Institute for Social Research  
426 Thompson St.  
Ann Arbor, MI 48106-1248

**Corresponding Author:**

Susan A. Murphy  
Email: [samurphy@umich.edu](mailto:samurphy@umich.edu)  
Fax: 734-763-4676  
Phone: 734-763-5046

**Acknowledgements:**

This research was supported by NIH grants: R21 DA019800, K02 DA15674, and P50 DA10075.

## ABSTRACT

The treatment of drug addiction presents a challenge to clinicians for many reasons; heterogeneity in response to treatment, the chronic nature of the disease as well as the high probability of relapse after a response to treatment. As a result clinicians must make decisions regarding the sequencing of treatments over time. Sequences of treatments which are guided by a patient's responses to prior treatments have been termed "adaptive treatment strategies". Specialized experimental designs (SMART or sequential multiple assignment randomize trials) have been proposed as a way to develop such sequences of treatments. Using an example of an adaptive treatment strategy for the treatment of prescription opioid dependence, we demonstrate how such a strategy would be refined through a SMART trial. We specify four research hypotheses that the SMART design might answer; two concerning the efficacy of individual treatments within a sequence, and two concerning the efficacy of the sequences themselves. For each hypothesis, we present a test statistic and sample size formula. Two of the sample size formulae are newly developed; one for supporting hypotheses about prespecified sequences of treatments beginning with different initial treatments, and another supporting the estimation of the best treatment strategy among those tested. We present the results of a set of simulations to evaluate the robustness of the newly developed sample size formulae in the presence of violations of their assumptions. Both formulae performed well in the presence of mild to moderate violations to their assumptions. In conclusion, we make recommendations for future methodological research and highlight the promise of SMART trials to untangle the factors affecting relapse and treatment withdrawal in drug addiction treatment.

# Statistical Methodology for a SMART Design in the Development of Adaptive Treatment Strategies

## 1. INTRODUCTION

The past two decades have brought new pharmacotherapies as well as psychotherapies to the field of drug addiction.<sup>1-4</sup> Despite this progress, the treatment of drug addiction in clinical practice remains a matter of trial and error. Some reasons for this difficulty are as follows. First, to date, no one treatment has been found that works well for most patients; that is, patients are heterogeneous in response to any one treatment. Second, as many authors have pointed out<sup>5,6</sup>, addiction is often a chronic condition with symptoms waxing and waning over time. Third, relapse is common, regardless of the type of treatment. Therefore, the clinician is faced with first finding a sequence of treatments that work to initially stabilize the patient, and next deciding which types of treatments will prevent relapse in the longer term. Here, we present a statistical methodology (as well as an evaluation of that methodology) for the design and analysis of simple sequential multiple assignment randomized trials (SMART). These experimental designs support the investigation of a sequence of treatments in a principled way. For clarity, the SMART design is discussed in the context of the development of a treatment strategy for prescription opioid dependence. Additionally, we introduce and evaluate two new methods for sizing SMART trials. We conclude that with this statistical methodology, SMART designs are now ready for widespread use.

Treatment strategies that are shaped by the individual patient characteristics or patient's responses to prior treatments are called "adaptive treatment strategies"<sup>7-14</sup>. Here is an example of an adaptive treatment strategy for opioid dependence, modeled after a trial currently in progress within the Clinical Trials Network of the National Institute on Drug Abuse<sup>15</sup>.

Example: First, provide all patients with a four week course of buprenorphine/naloxone plus individual drug counseling (IDC)<sup>16</sup>. If the patient remains abstinent<sup>A</sup> from opioid use during those four weeks, provide 12 additional weeks of relapse prevention therapy (RPT). If at any time during the four weeks, the patient meets the criterion for non-abstinence a second, longer buprenorphine/naloxone facilitated detoxification is provided. The second detoxification lasts 12 weeks and is accompanied by cognitive behavioral therapy (CBT).

A patient whose treatment is consistent with this strategy experiences one of two *sequences* of two treatments, as depicted in Figure 1. The two sequences are:

1. Four week buprenorphine/naloxone detoxification plus IDC followed by 12 week RPT (if abstinent for initial four weeks),

---

<sup>A</sup> Abstinance might be operationalized using a criterion based on self report or opioid use, urine screens and adherence to medication.

2. Up to a four week buprenorphine/naloxone detoxification plus IDC followed by 12 week buprenorphine/naloxone detoxification plus CBT (if not abstinent during the initial four weeks).

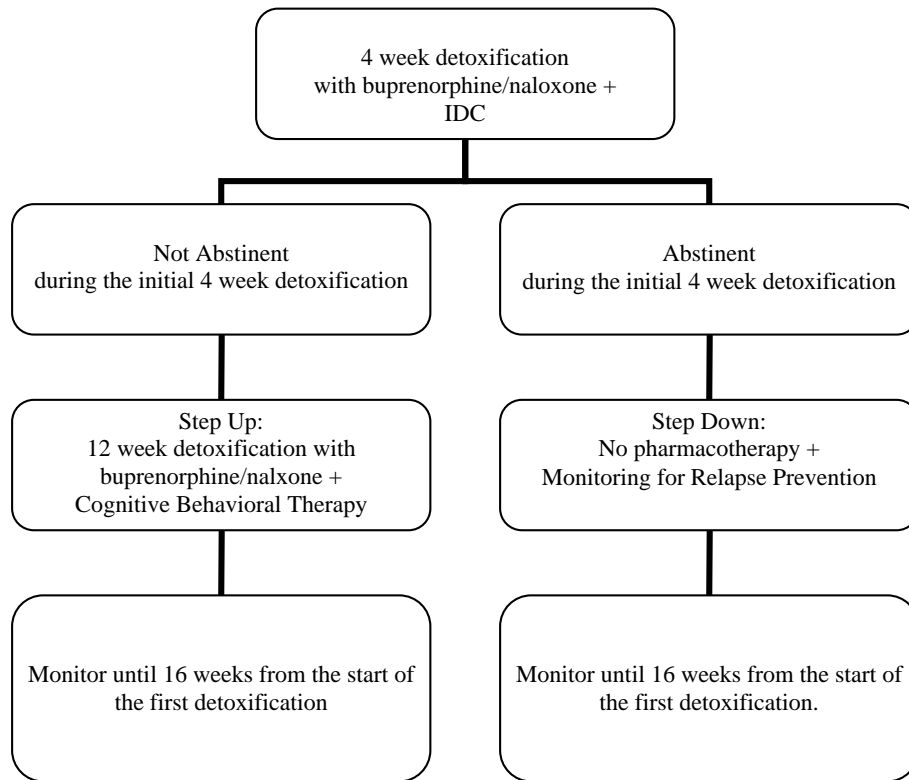


Figure 1. An adaptive treatment strategy for prescription opioid dependence

In the next section, several research questions useful in guiding the development of a hypothetical adaptive treatment strategy for prescription opioid dependence are discussed. Following that, we present an experimental design (a SMART design) to answer these questions. We present statistical methodology for evaluating data from a particular SMART design and a comprehensive discussion and evaluation of these statistical considerations in the fourth and fifth sections. In the final section, we present a summary and conclusions and discuss suggested areas of future research.

## 2. RESEARCH QUESTIONS TO REFINE AN ADAPTIVE TREATMENT STRATEGY

In order to refine the adaptive treatment strategy in Figure 1, we might ask if we could begin with a less intensive psychotherapy, for example standard medical management (MM)<sup>17</sup>, which focuses primarily on medication compliance. This less burdensome treatment might be effective for a large majority of patients. That is, which initial accompanying psychosocial therapy, IDC or MM, produces the best long term outcome

in the context of options for further treatment? A second question focuses on the psychotherapy accompanying the longer 12 week detoxification. That is, for patients who do not remain abstinent during the initial 4 week detoxification, and are provided a 12 week regimen of buprenorphine/naloxone, which accompanying psychosocial treatment produces the best long term outcome: IDC or CBT? The long term outcome might be a cumulative measure of the number of days the subject remained abstinent (as confirmed by a combination of urine screens and self report) over the entire 16 week trial.

On the other hand, instead of focusing our research questions on a particular treatment in the strategy, we may be interested in the possible adaptive treatment strategies. A set of simple strategies is presented below in Table 1. In some settings, CBT may be the preferred psychosocial therapy to use with longer detoxifications; in this case we might want to compare strategies A and C. These two strategies both use CBT in the second detoxification for those who are not abstinent in the first. However, strategy A begins the sequence with the more intensive IDC while strategy C begins with the less intensive MM. Alternatively, we may simply wish to identify which of the four strategies, A, B, C or D results in the highest mean outcome.

The different research questions are summarized in Table 2 below.

Table 1. Potential strategies to consider for the treatment of prescription opioid dependence

Initial Psychosocial Treatment (combined with 4 week detoxification)	Response to Initial Treatment	Second Treatment
<b>Strategy #A: Begin with IDC, if non-response provide CBT, if response provide RPT.</b>		
IDC	Not abstinent	12 week detoxification plus CBT
IDC	Abstinent	Relapse Prevention
<b>Strategy #B: Begin with IDC, if non-response provide IDC if response provide RPT.</b>		
IDC	Not abstinent	12 week detoxification plus IDC
IDC	Abstinent	Relapse Prevention
<b>Strategy #C: Begin with MM, if non-response provide CBT, if response provide RPT</b>		
MM	Not abstinent	12 week detoxification plus CBT
MM	Abstinent	Relapse Prevention
<b>Strategy #D: Begin with MM, if non-response provide IDC, if response provide RPT.</b>		
MM	Not abstinent	12 week detoxification plus IDC
MM	Abstinent	Relapse Prevention

Table 2. Four research questions to guide the development of adaptive treatment strategies

Type	Research Question	Null Hypothesis
Two analyses that concern components of adaptive treatment strategies		
1	What is the effect of initial treatment assignment on long term outcome given specified treatments provided in the interim?	The mean long term outcome for all patients assigned to IDC initially will be equal to the mean long term outcome of all patients assigned to MM initially.
2	Considering only patients whose disorder did not respond to the initial treatment, what is the best subsequent psychosocial treatment in the context of a 12 week buprenorphine detoxification: IDC vs. CBT?	Considering only patients whose disorder did not respond to the initial treatment, the average long term outcome for those provided subsequently with CBT will be the same as the long term average of those provided with IDC.
Two possible analyses that concern whole adaptive treatment strategies		
3	What is the difference in long term outcomes between the strategies that provide patients who are not abstinent with 12 weeks of buprenorphine/naloxone plus CBT, when the psychosocial treatment accompanying the initial 4 week detoxification is IDC versus when it is MM?	The mean long term outcome for all those exposed to the strategy that begins with IDC, then treats patients who are not abstinent with CBT will be equal to the long term mean outcome of all exposed to the analogous strategy that begins with MM.
4	Which treatment strategy produces the best outcome?	This is an estimation problem not a hypothesis testing problem.

### 3. EXPERIMENTAL DESIGNS TO SUPPORT THE DEVELOPMENT OF ADAPTIVE TREATMENT STRATEGIES

In the previous section, we posed four different research questions in the development of a hypothetical adaptive treatment strategy for prescription opioid dependence. Two refer to comparisons of individual components (e.g. treatments) of the strategy and two refer to comparisons of whole strategies. Note that traditional experimental trials typically evaluate a single treatment with no specification and/or control of preceding or subsequent treatments. In contrast the SMART design provides data that can be used to assess the usefulness of each treatment in a sequence and can also be used to contrast whole strategies. Further rationale for the SMART trial can be found in Murphy et al<sup>10,11</sup>. A SMART design that can be used to address the four questions posed in the previous section is presented in Figure 2.

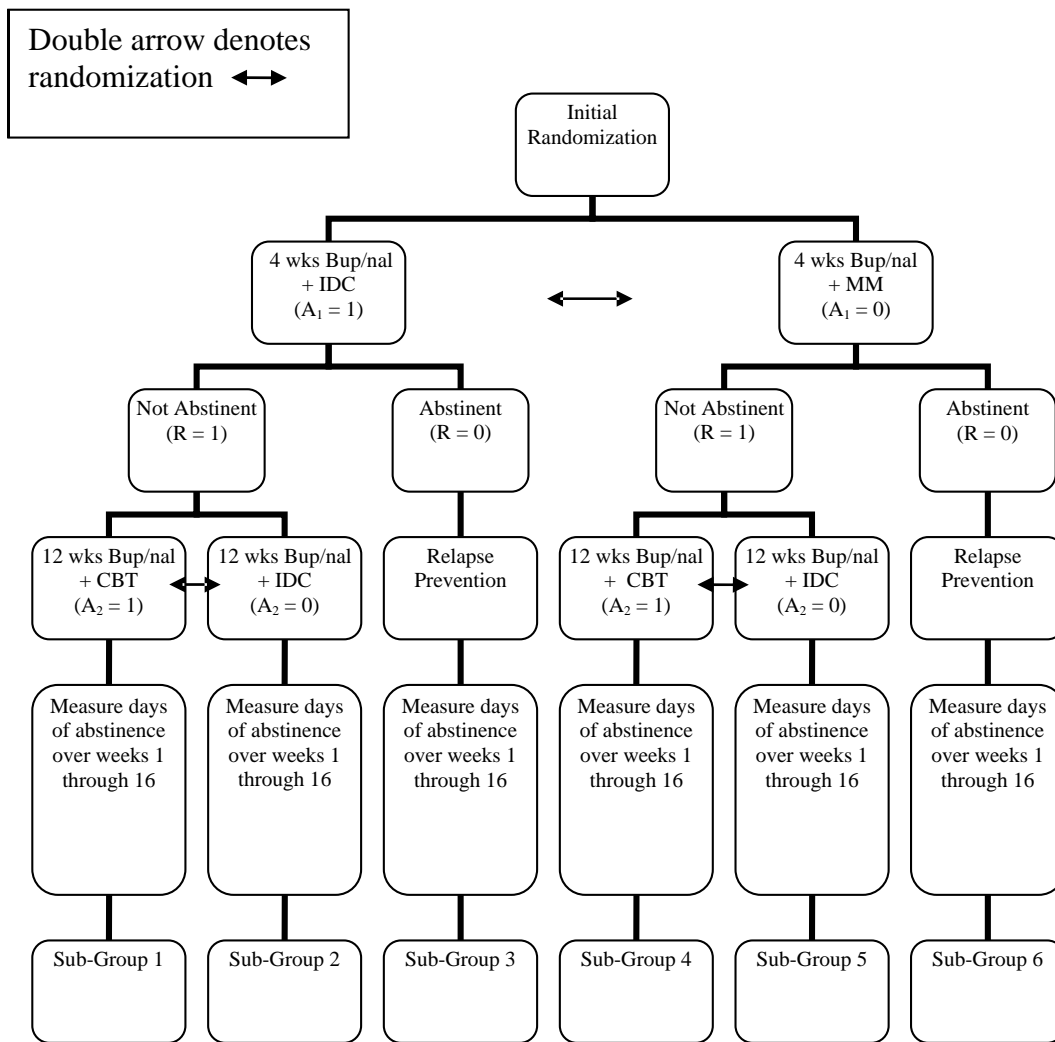


Figure 2. SMART trial to develop adaptive treatment strategies for prescription opioid dependence.

Note that because SMART specifies sequences of treatments, it allows us to determine the effectiveness of one of the components in the presence of either preceding or subsequent treatments. Second, it uses randomization to support causal inferences about the effectiveness of a component to produce different long term outcomes. Questions concerning the effectiveness of MM vs. IDC as the first psychosocial treatment, with subsequent psychosocial treatments consisting of either IDC or CBT for those who are not abstinent and RPT for those who are abstinent are answered by comparing the outcomes of groups 1 through 3 with those of groups 4 through 6. This is the main effect of the initial psychosocial treatment, as depicted in the dummy coding shown in Table 3. Similarly, questions concerning the effectiveness of IDC vs. CBT as a second line treatment (when preceded by either MM or IDC) are asked by pooling outcome data from groups 1 and 4, and comparing it to the pooled outcome data of groups 2 and 5. This is

the main effect of the second psychosocial treatment among those not abstinent during the initial four week detoxification as shown in Table 3.

Table 3. Dummy coding to reflect treatments potentially tested by the SMART design

Sub-Group Number	Initial Psychosocial Treatment  1 = IDC 0 = MM	Response to Treatments  1 = Not abstinent 0 = Abstinent	Second Psychosocial Treatment for Patients who Relapsed  1 = CBT 0 = IDC
1	1	1	1
2	1	1	0
3	1	0	0
4	0	1	1
5	0	1	0
6	0	0	0

The SMART design can also be used to answer the two questions concerning entire strategies. Strategy A can be compared to strategy C by appropriately pooling outcomes from sub-groups 1 and 3 to form an average outcome for strategy A and appropriately pooling outcomes from sub-groups 4 and 6 to form an average outcome for strategy C (see the next section for formulae). We could also follow a similar process to form the average outcome for each of the four strategies and then the best strategy overall will be the strategy that is associated with the highest average outcome.

#### 4. TEST STATISTICS AND SAMPLE SIZE FORMULAE

In this section, we provide the test statistics and sample size formulae for the four different types of research questions summarized in Table 3. As discussed previously, Analyses 1 and 2 concern the components of an adaptive treatment strategy, and Analyses 3 and 4 concern strategies as a whole. We introduce new methods for sizing SMART trials that have Analysis 3 or 4 as the main goal. The formulae below are for the SMART design in which there are two initial treatment options, then two treatment options for non-responders and one treatment option for responders at the second time point. In conversations with researchers across the mental health field, we have found this design to be of the greatest interest; these designs are similar to the designs employed by STAR\*D<sup>18,19</sup> and CATIE<sup>20</sup>. Also note that this design is balanced; that is, the two treatment options for non-responders are the same regardless of initial treatment.

We use the following notation.  $A_1$  is the indicator for the initial treatment, R denotes the response to the initial treatment (non-response = 1 and response = 0),  $A_2$  is the treatment indicator for non-responders, and Y denotes a continuous outcome. The two treatment options for non-responders to treatment  $A_1=1$  are the same as the two treatment options for non-responders to  $A_1=0$ . For simplicity, suppose the patients are randomized equally

to the two treatment options at each level; i.e. the probability that a patient receives initial treatment  $A_1=1$  is 0.5.

#### 4.1 Statistics for Different Analyses

The test statistics for Analyses 1-3 are presented in Table 4; the method for performing Analysis 4 is also given in Table 4. Note that while Analyses 1-3 are hypotheses tests, Analysis 4 is not a hypothesis test. The test statistics for Analyses 1 and 2 are the standard test statistics for a two group comparison with large samples<sup>21</sup> and are not unique to the SMART design. The estimator of a strategy mean, used in both Analysis 3 and Analysis 4, as well as the test statistic for Analysis 3 are given in Murphy<sup>9</sup>. In large samples the three test statistics corresponding to Analyses 1-3 are normally distributed (with mean zero under the null hypothesis of no effect). In Tables 4 and 5, we present simplified versions of these equations; these versions are for strategies with two initial treatment options and two secondary treatment options for non-responders.

Table 4. Test statistics for each of the possible hypotheses

Type of Analysis	Test Statistic
<b>1<sup>(1)</sup></b>	$Z = \frac{(\bar{Y}_{A_1=1} - \bar{Y}_{A_1=0})}{\sqrt{\frac{S^2_{A_1=1}}{N_{A_1=1}} + \frac{S^2_{A_1=0}}{N_{A_1=0}}}}$ <p>where <math>N_{A_1=i}</math> denotes the number of subjects who received <math>i</math> as the initial treatment</p>
<b>2<sup>(1)</sup></b>	$Z = \frac{(\bar{Y}_{R=1, A_2=1} - \bar{Y}_{R=1, A_2=0})}{\sqrt{\frac{S^2_{R=1, A_2=1}}{N_{R=1, A_2=1}} + \frac{S^2_{R=1, A_2=0}}{N_{R=1, A_2=0}}}}$ <p>where <math>N_{R=1, A_2=i}</math> denotes the number of non-responders who received <math>i</math> as the second treatment</p>
<b>3<sup>(2)</sup></b>	$Z = \frac{\sqrt{N}(\hat{\mu}_{A_1=1, A_2=a_2} - \hat{\mu}_{A_1=0, A_2=b_2})}{\sqrt{\hat{\tau}^2_{A_1=1, A_2=a_2} + \hat{\tau}^2_{A_1=0, A_2=b_2}}}$ <p>where <math>N</math> is the <b>total</b> number of subjects, and <math>a_2</math> and <math>b_2</math> are the second treatments in the two prespecified strategies being compared.</p>
<b>4</b>	Choose largest of $\hat{\mu}_{A_1=1, A_2=1}, \hat{\mu}_{A_1=0, A_2=1}, \hat{\mu}_{A_1=1, A_2=0}, \hat{\mu}_{A_1=0, A_2=0}$

<sup>(1)</sup>  $\bar{Y}$  and  $S^2$  the sample mean and the sample variance; the subscript on  $N$  denotes the group of subjects

<sup>(2)</sup> See Table 5 for a definition of  $\hat{\mu}$  and  $\hat{\tau}^2$ .

Table 5. Estimators for strategy means and estimators for variance of estimator of strategy means.

Data for  $i^{\text{th}}$  patient is of the form  $(A_{1i}, R_i, A_{2i}, Y_i)$ , where  $A_{1i}, R_i, A_{2i}$ , and  $Y_i$  are defined as in Section 4, and  $N$  is the *total* sample size.

Strategy sequence (a1, a2)	$\hat{\mu}_{A1=a1, A2=a2}$ : Estimator for strategy mean
(1, 1)	$\frac{\frac{1}{N} \sum_{i=1}^N Y_i * A_{1i} * (R_i * A_{2i} + (1 - R_i))}{\frac{1}{N} \sum_{i=1}^N 0.5 * (R_i * 0.5 + (1 - R_i))}$
(1, 0)	$\frac{\frac{1}{N} \sum_{i=1}^N Y_i * A_{1i} * (R_i * (1 - A_{2i}) + (1 - R_i))}{\frac{1}{N} \sum_{i=1}^N 0.5 * (R_i * 0.5 + (1 - R_i))}$
(0, 1)	$\frac{\frac{1}{N} \sum_{i=1}^N Y_i * (1 - A_{1i}) * (R_i * A_{2i} + (1 - R_i))}{\frac{1}{N} \sum_{i=1}^N 0.5 * (R_i * 0.5 + (1 - R_i))}$
(0, 0)	$\frac{\frac{1}{N} \sum_{i=1}^N Y_i * (1 - A_{1i}) * (R_i * (1 - A_{2i}) + (1 - R_i))}{\frac{1}{N} \sum_{i=1}^N 0.5 * (R_i * 0.5 + (1 - R_i))}$
Strategy sequence (a1, a2)	$\hat{\tau}^2_{A1=a1, A2=a2}$ : N*Estimator for variance of above estimator of strategy mean
(1, 1)	$\left( \frac{1}{N} \sum_{i=1}^N \frac{(Y_i - \hat{\mu}_{11}) * A_{1i} * (R_i * A_{2i} + (1 - R_i))}{0.5 * (R_i * 0.5 + (1 - R_i))} \right)^2$
(1, 0)	$\left( \frac{1}{N} \sum_{i=1}^N \frac{(Y_i - \hat{\mu}_{10}) * A_{1i} * (R_i * (1 - A_{2i}) + (1 - R_i))}{0.5 * (R_i * 0.5 + (1 - R_i))} \right)^2$
(0, 1)	$\left( \frac{1}{N} \sum_{i=1}^N \frac{(Y_i - \hat{\mu}_{01}) * (1 - A_{1i}) * (R_i * A_{2i} + (1 - R_i))}{0.5 * (R_i * 0.5 + (1 - R_i))} \right)^2$
(0, 0)	$\left( \frac{1}{N} \sum_{i=1}^N \frac{(Y_i - \hat{\mu}_{00}) * (1 - A_{1i}) * (R_i * (1 - A_{2i}) + (1 - R_i))}{0.5 * (R_i * 0.5 + (1 - R_i))} \right)^2$

## 4.2 Sample Size Calculations

In the following, all sample size formulae assume a two-tailed z-test. Let  $\alpha$  be the desired size of the hypothesis test and let  $1-\beta$  be the power of the test, and let  $z_{\alpha/2}$  be the standard normal  $(1-\alpha/2)$  percentile. Approximate normality of the test statistic is assumed throughout.

In order to calculate the sample size, one must also input the desired detectable standardized effect size. We denote the standardized effect size by  $d$  and use the definition for standardized effect size found in Cohen<sup>22</sup>. The standardized effect size between two groups is defined as the difference between the means of the two groups divided by the square root of the pooled variance, which is the square root of the average of the variances of the two groups being compared; in simple notation, we have

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}.$$

All of the sample size formulae below make the working assumption that the variances of the two groups under consideration are equal. Under the equal variance assumption, the formula for the standardized effect size further simplifies to  $\delta = \frac{\Delta\mu}{\sigma}$ . The definition of the variance changes with the analysis under consideration; we will explicitly define the variance assumption as we present each sample size formulae. The standardized effect sizes for the various analyses we are considering are summarized in Table 6.

Table 6. Standardized effect sizes for the four analyses in Table 3

Analysis	Formula for Standardized Effect Size $d$
1	$\delta = \frac{E[Y   A_1 = 1] - E[Y   A_1 = 0]}{\sqrt{\frac{\text{Var}[Y   A_1 = 1] + \text{Var}[Y   A_1 = 0]}{2}}}$
2	$\delta = \frac{E[Y   R = 1, A_2 = 1] - E[Y   R = 1, A_2 = 0]}{\sqrt{\frac{\text{Var}[Y   R = 1, A_2 = 1] + \text{Var}[Y   R = 1, A_2 = 0]}{2}}}$
3	$\delta = \frac{E[Y   A_1 = 1, A_2 = a2] - E[Y   A_1 = 0, A_2 = b2]}{\sqrt{\frac{\text{Var}[Y   A_1 = 1, A_2 = a2] + \text{Var}[Y   A_1 = 0, A_2 = b2]}{2}}}$ <p>where a2 and b2 are the second components in the two prespecified strategies being compared.</p>
4	$\delta = \frac{E[Y   A_1 = a1, A_2 = a2] - E[Y   A_1 = b1, A_2 = b2]}{\sqrt{\frac{\text{Var}[Y   A_1 = a1, A_2 = a2] + \text{Var}[Y   A_1 = b1, A_2 = b2]}{2}}}$ <p>where (a1, a2) = strategy with the highest mean outcome, (b1, b2) = strategy with the next highest mean outcome.</p>

The sample size formulas for Analyses 1 and 2 are standard formulas<sup>23</sup> and assume an equal number in each of the two groups being compared. For Analysis 1, the equal variance assumption means that the variance of outcome Y given the first treatment  $A_1=1$  is equal to the variance of outcome Y given the first treatment  $A_1=0$ ; i.e.  $s^2 = \text{Var}[Y|A_1=1] = \text{Var}[Y|A_1=0]$ . Given desired levels of size, power and standardized effect size, the total sample size required for Analysis 1 is

$$N_1 = 2 * 2 * (z_{\alpha/2} + z_{\beta})^2 * (1/\delta)^2 .$$

Note that the sample size calculation for Analysis 1 does not require the input of a non-response rate.

The formula for the total sample size required for Analysis 2 does depend on a guess for the (intermediate) non-response rate, since this indicates the number of patients who will be randomized a second time. The sample size formulae use the working assumption that the intermediate non-response rates are equal; that is, that the probability of non-response for a patient given initial treatment  $A_1=1$  is the same as the probability of non-response for a patient given initial treatment  $A_1=0$ . This working assumption is only used to size the SMART design and is not used to analyze the data from the trial as can be seen from Table 4. For example, in sizing for the SMART design in Section 3, we would assume that the probability of non-response for a patient given initial treatment IDC is the same as the probability of non-response for a patient given initial treatment MM. In the following sample size formulae, we will denote this identical non-response rate by p. For

Analysis 2, the equal variance assumption means that the variance of outcome Y for non-responders who were given second treatment  $A_2=1$  is equal to the variance of outcome Y for non-responders who were given second treatment  $A_2=0$ ; i.e.  $s^2 = \text{Var}[Y|R=1, A_2=1] = \text{Var}[Y|R=1, A_2=0]$ . The formula for the total required sample size for Analysis 2 is

$$N_2 = 2 * 2 * (z_{\alpha/2} + z_{\beta})^2 * (1/\delta)^2 / p.$$

When calculating the sample sizes to test Analysis 3, two different sample size formulae can be used: one depends on the intermediate non-response rates as an input and one that does not. In both sample size formulae, the equal variance assumption indicates that the variance of outcome Y given treatment strategy ( $A_1=1, A_2=a2$ ) is equal to the variance of outcome Y given treatment strategy ( $A_1=0, A_2=b2$ ); i.e.  $s^2 = \text{Var}[Y|A_1=1, A_2=a2] = \text{Var}[Y|A_1=0, A_2=b2]$ . The formula that uses a guess of the intermediate non-response rate makes two additional working assumptions. First, the response rates are equal for both initial treatments and second, the variability of the outcome Y around the strategy mean ( $A_1=1, A_2=a2$ ), among either responders or non-responders, is less than the variance of the strategy mean and similarly for strategy ( $A_1=0, A_2=b2$ ). This formula is

$$N_{3a} = 2 * (z_{\alpha/2} + z_{\beta})^2 * (2 * (2 * p + 1 * (1 - p))) * (1/\delta)^2.$$

The second formula does not require either of these two additional working assumptions; it specifies the sample size required if the non-response rates are both 1, a “worst case scenario.” This conservative sample size formula for Analysis 3 is

$$N_{3b} = 2 * (z_{\alpha/2} + z_{\beta})^2 * 4 * (1/\delta)^2.$$

We will compare the performance of these two sample size formulae for Analysis 3 in the next section. See the Appendix for a derivation of these formulae.

The method for finding the sample size for Analysis 4 relies on an algorithm rather than a formula; we will refer to the resulting sample size as  $N_4$ . Analysis 4 is not a hypothesis test; instead of specifying power to detect a difference in two means, the sample size is based on the desired probability to detect the strategy that results in the highest mean outcome. The standardized effect size in this case involves the difference between the two highest strategy means. The algorithm uses an idea similar to the one used to derive the sample size formula for Analysis 3 that is invariant to the non-response rate. Given a desired level of probability for selecting the correct treatment strategy with the highest mean and a desired treatment strategy effect, the algorithm for Analysis 4 finds the samples sizes that correspond to the range of non-response probabilities and then chooses the largest sample size. Since it is based on a “worst case scenario,” this algorithm will result in a conservative sample size formula. See the Appendix for a derivation of this algorithm.

Example sample sizes are given in Table 7 below. In examining these sample sizes it appears odd at first that as the non-response rate increases, the required sample sizes for Analysis 3 (e.g. a comparison of strategies A versus C) increases. To see why this must be the case, consider two extreme cases, the first in which the response rate is 90% for both initial treatments and the second in which the non-response rate is 90%. In the former case, if  $n$  subjects are assigned to IDC and 90% respond (i.e. 10% do not respond), then the resulting sample size for strategy A is  $(0.9)n + \frac{1}{2}(0.1)n = 0.95n$ . The  $\frac{1}{2}$  occurs due to the randomization of non-responders between CBT and IDC. On the other hand, if only 10% respond (i.e. 90% do not respond), then the resulting sample size for strategy A is  $(0.1)n + \frac{1}{2}(0.9)n = 0.55n$ , which is less than  $0.95n$ . Thus, the higher the expected non-response rate, the larger the initial sample size required for a given power to differentiate between strategies A and C. This apparently nonsensical result only occurs because the number of treatment options (2 options) for non-responders is greater than the number of treatment options for responders (only 1).

Table 7. Example Sample Sizes  
All entries are for *total* sample size

Desired Size <sup>(1)</sup> a	Desired Power <sup>(2)</sup> 1-β	Standardized Effect Size d	Non-response Rate <sup>(3)</sup> p	Analysis 1	Analysis 2	Analysis 3 (sample size varies by p)	Analysis 3 (sample size invariant to p)	Analysis 4
a = .10								
	1-β = .80							
		d = .20						
			p = .5	620	1240	930	1240	358
			p = .9	620	689	1178	1240	358
		d = .50						
			p = .5	99	198	149	198	59
			p = .9	99	110	188	198	59
	1-β = .90							
		d = .20						
			p = .5	864	1728	1297	1729	608
			p = .9	864	960	1,642	1729	608
		d = .50						
			p = .5	138	277	207	277	97
			p = .9	138	154	263	277	97
a = .05								
	1-β = .80							
		d = .20						
			p = .5	784	1568	1176	1568	358
			p = .9	784	871	1490	1568	358
		d = .50						
			p = .5	125	251	188	251	59
			p = .9	125	139	238	251	59
	1-β = .90							
		d = .20						
			p = .5	1056	2112	1584	2112	608
			p = .9	1056	1174	2007	2112	608
		d = .50						
			p = .5	169	338	254	338	97
			p = .9	169	188	321	338	97

<sup>(1)</sup> All entries assume that each statistical test is two tailed; size is not required for Analysis 4 since it is not a hypothesis test.

<sup>(2)</sup> Analysis 4 is not a hypothesis test; we choose the sample size so that the probability that we choose the best treatment, given such a “best” treatment exists (i.e. given that there is a treatment strategy that has a higher mean outcome than the rest) is 1-β.

<sup>(3)</sup> Non-response rates are all equal, i.e.  $p = \Pr\{R=1|A_1=1\} = \Pr\{R=1|A_1=0\}$

## 5. EVALUATION OF SAMPLE SIZE FORMULAE VIA SIMULATION

In this section, the sample size formulae presented in Section 4.2 are evaluated. We examine the robustness of the newly developed methods for calculating sample sizes for Analyses 3 and 4. In addition, a second assessment investigates the power for Analysis 4 to detect the best strategy when the study is sized for one of the other analyses. The second assessment is provided because due to the emphasis on strategies in SMART designs, Analysis 4 is likely to always be of interest.

### 5.1 Simulation Designs

The sample sizes used for the simulations were chosen to give a power level of 0.90 and a type I error of 0.05 when one of Analyses 1-3 is used to size the trial, and a 0.90 probability of choosing the best strategy for Analysis 4 when it is used to size the trial; these sample sizes are shown in Table 8. For Analysis 1-3, power is estimated by the proportion of times out of 1000 simulations that the null hypothesis is correctly rejected; for Analysis 4, the probability of choosing the best strategy is estimated by the proportion of times out 1000 simulations that the correct strategy with the highest mean is chosen. We sized the studies to detect a prespecified standardized effect size of 0.2 or 0.5. We follow Cohen<sup>22</sup> in labeling 0.2 as a “small” effect size and 0.5 as a “medium” effect size. The simulated data reflect the types of scenarios found in substance abuse clinical trials.<sup>24-26</sup> For example, the simulated data exhibits intermediate non-response rates (proportion of simulated subjects with R=1) of 0.5 and 0.9, and the mean outcome for the responders is higher than for non-responders. The simulation designs and the programs used can be found in an online technical report.<sup>27</sup>

For Analysis 3, we need to specify the strategies of interest, and for the purposes of these simulations, we will compare strategies  $(A_1=1, A_2=1)$  and  $(A_1=0, A_2=0)$ . For the simulations to evaluate the robustness of the sample size calculation for Analysis 4, we choose  $(A_1=1, A_2=1)$  to always have the highest mean outcome and generate the data according to two different “patterns”: 1) the strategy means are all different and 2) the mean outcomes of the other three strategies besides  $(A_1=1, A_2=1)$  are all equal. In the second pattern, it is more difficult to detect the “best” strategy because the highest mean must be distinguished from *all* the rest, which are all the “next highest”, instead of just *one* next highest mean.

In order to test the robustness of the sample size formulas, we calculate a sample size given by the relevant formula in Section 4.2, and then simulate data sets of this sample size. The simulated data will not satisfy the working assumptions in one of the following ways:

- the intermediate non-response rates to first level treatments are unequal, i.e.  $\Pr\{R=1|A_1=1\} \neq \Pr\{R=1|A_1=0\}$ ,
- the variances relevant to the analysis are unequal,
- the distribution of the final outcome,  $Y$ , are right skewed (thus for a given sample size, the test statistic is more likely to have a non-normal distribution).

When we challenge the non-response rate equality assumption, we calculate the sample size formula for a particular non-response rate  $p$ , then generate the data with non-response rates  $\Pr\{R=1|A_1=1\} = p-0.05$  and  $\Pr\{R=1|A_1=0\} = p+0.05$ . In challenging the equal variance assumption, we set one of the variances at 81% of the other variance. For example, for Analysis 1, we set  $\text{Var}[Y|A_1=0] = .81*\text{Var}[Y|A_1=1]$ . To challenge the normality assumption for the final outcome, we generate  $Y$  using a gamma distribution with skewness up to 3. See the Appendix for a description of how the final outcomes from a gamma distribution were generated.

We also assess the power of Analysis 4 when it is not used in sizing the trial. For each of the types of analyses in Table 2, we generate a data set that follows the working assumptions for the sample size formula for that analysis (for example, use  $N_2$  to size the study to test the effect of the second treatment on the mean outcome), and then perform Analysis 4 on the data and estimate the probability of choosing the correct strategy with the highest mean outcome.

## 5.2 Robustness of the New Sample Size Formulae

As previously mentioned, since the sample size formulae for Analyses 1 and 2 are standard formulae, we will focus on evaluating the newly developed sample size formulae for Analyses 3 and 4. Table 8a and 8b provide the results of the simulations designed to evaluate the sample size formulas for analyses 3 and 4 respectively.

Table 8a. Investigation of Sample Size Assumption Violations for Analysis 3 comparing strategies (1,1) and (0,0);

The power to reject the null hypothesis for Analysis 3 is shown when sample size is calculated to reject the null hypothesis for Analysis 3 with power of 0.90 and type I error of 0.05 (two-tailed)

Simulation Parameters				Simulation Results (power)			
Effect size	Non-response rate (Default)	Sample size formula	Total sample size	Default working assumptions are correct	Non-equal non-response rates <sup>(1)</sup>	Non-equal variance <sup>(2)</sup>	Non-normal outcome $Y^{(3)}$
0.2	0.5	$N_{3a}$	1584	0.893	0.902	0.900	0.882
0.2	0.9	$N_{3a}$	2007	0.882	0.910	0.916	0.877*
0.5	0.5	$N_{3a}$	254	0.896	0.864*	0.920	0.851*
0.5	0.9	$N_{3a}$	321	0.926*	0.886	0.880	0.898
0.2	0.5	$N_{3b}$	2112	0.950*	0.958*	0.954*	0.974*
0.2	0.9	$N_{3b}$	2112	0.903	0.934*	0.931*	0.898
0.5	0.5	$N_{3b}$	338	0.973*	0.938*	0.971*	0.916
0.5	0.9	$N_{3b}$	338	0.937*	0.890	0.889	0.922*

<sup>(1)</sup>  $\Pr\{R=1|A_1=1, A_2=1\} = p-0.05$  and  $\Pr\{R=1|A_1=0, A_2=0\} = p+0.05$ , where  $p$  is the "default" non-response rate.

<sup>(2)</sup>  $\text{Var}[Y|A_1=0, A_2=0] = .81*\text{Var}[Y|A_1=1, A_2=1]$

<sup>(3)</sup> The final outcome comes from a gamma distribution.

\* The 95% confidence interval for this proportion does not contain 0.90

Considering Table 8a, we see that the Analysis 3 sample size formula  $N_{3a}$  performed extremely well when the expected standardized effect size was 0.20. Resulting power levels were uniformly near 0.90 regardless of either the true intermediate response rates or any of the three violations of the working assumptions. Power levels were less robust when the sample sizes calculated were smaller – i.e. for the 0.50 effect size. For example, when the non-response rates are not equal (by initial treatment) the resulting power is lower than 0.90 in the rows using an assumed non-response rates of 0.5. The more conservative sample size formula,  $N_{3b}$  performed well in all scenarios, regardless of non-response rate or the presence of any of the three violations to underlying assumptions. As the non-response rate approaches 1, the sample sizes are less conservative, but the results for power remain within a 95% confidence interval of 0.90.

Table 8b. Investigation of Sample Size Violations for Analysis 4;  
Probability<sup>(1)</sup> to detect the correct “best” strategy  
when the sample size is calculated to detect the correct maximum strategy mean 90% of  
the time.

Simulation Parameters				Simulation Results (probability)			
Effect size	Non-response rate (Default)	Pattern <sup>(2)</sup>	Sample size <sup>(3)</sup>	Default working assumptions are correct	Non-equal non-response rates <sup>(4)</sup>	Non-equal variance <sup>(5)</sup>	Non-normal outcome Y <sup>(6)</sup>
0.2	0.5	1	608	0.966*	0.984*	0.965*	0.972*
0.2	0.9	1	608	0.962*	0.969*	0.964*	0.962*
0.5	0.5	1	97	0.980*	0.985*	0.966*	0.956*
0.5	0.9	1	97	0.960*	0.919*	0.976*	0.947*
0.2	0.5	2	608	0.964*	0.953*	0.952*	0.944*
0.2	0.9	2	608	0.905	0.929*	0.922*	0.923*
0.5	0.5	2	97	0.922*	0.974*	0.976*	0.948*
0.5	0.9	2	97	0.893	0.917	0.927*	0.885

<sup>(1)</sup> Probability calculated as the percentage of 1000 simulations on which correct strategy mean was selected as the maximum.

<sup>(2)</sup> 1 refers to the pattern of strategy means such that all are different, but that for (1,1) is always the highest. 2 refers to the pattern of strategy means such that the mean for (1,1) is higher than the other three and the other three are all equal.

<sup>(3)</sup> Calculated to detect the correct maximum strategy mean 90% of the time when the sample size assumptions hold.

<sup>(4)</sup>  $\Pr\{R=1|A_1=1, A_2=1\} = p-0.05$  and  $\Pr\{R=1|A_1=a1, A_2=a2\} = p+0.05$ , where  $p$  is the “default” non-response rate and  $(a1, a2)$  is the strategy with the next highest mean.

<sup>(5)</sup>  $\text{Var}[Y|A_1=1, A_2=1] = .81*\text{Var}[Y|A_1=a1, A_2=a2]$ , where  $(a1, a2)$  is the strategy with the next highest mean.

<sup>(6)</sup> The final outcome comes from a gamma distribution

\* The 95% confidence interval for this proportion does not contain 0.90.

In Table 8b, the conservatism of the sample size calculation  $N_4$  (associated with Analysis 4) is apparent. We can see that  $N_4$  is less conservative for the more difficult scenario where the strategy means besides the highest are all equal, but the probability of correctly identifying the strategy with the highest mean outcome is still about 0.90.

Overall, under different violations of the working assumptions, the sample size formulas for Analysis 3 and Analysis 4 still performed well in terms of power.

As discussed above, we also assess the power for Analysis 4 when the trial was sized using a different analysis. For each of the types of analyses in Table 2, we generate a data set that follows the working assumptions for the sample size formula for that analysis, then evaluate the power of Analysis 4 to detect the optimal strategy. From Tables 9a-9c, we see that in almost all cases, regardless of the starting assumptions used to size the various analyses, we achieve a 0.9 probability or higher of correctly detecting the strategy with the highest mean outcome. The probability falls below 0.9 when the standardized effect size for Analysis 4 falls below 0.1. These results are not surprising as from Table 7 we see that Analysis 4 requires much smaller sample sizes than all the other analyses.

Note that Analysis 4 is more closely linked to Analysis 3 than to Analyses 1 or 2. Analysis 3 is potentially a subset of Analysis 4; this relationship occurs when one of the strategies considered in Analysis 3 is the strategy with the highest mean outcome. The probability of detecting the correct strategy mean as the maximum when sizing for Analysis 3 is generally very good as can be seen from Table 9c. This is due to the fact that the sample sizes required to test the differences between two strategy means (each beginning with a different initial treatment) are much larger than those needed to detect the maximum of four strategy means with a specified degree of confidence. For a z-test of the difference between two strategy means with a two tailed type I error rate of 0.05, power of 0.90, and standardized effect size of 0.20, the sample size requirements range from 1584 to 2112. The sample size required for a 0.90 probability of selecting of the correct strategy mean as a maximum when the standardized effect size between it and the next highest strategy mean is 0.2 is 608. It is therefore not surprising that the selection rates for the correct strategy mean are generally high when powered to detect differences between strategy means each beginning with a different initial strategy.

Table 9a. The probability<sup>(1)</sup> of choosing the correct strategy for Analysis 4 when sample size is calculated to reject the null hypothesis for Analysis 1 (for a two-tailed test with power of 0.90 and type I error of 0.05)

Simulation Parameters			Simulation Results		
Effect size for Analysis 1	Non-response Rate	Sample size	Analysis 1 (power)	Analysis 4 (probability <sup>(1)</sup> )	Effect size for Analysis 4
0.2	0.5	1056	0.880	1.000	0.325
0.2	0.9	1056	0.904	1.000	0.425
0.5	0.5	169	0.934	0.987	0.350
0.5	0.9	169	0.920	0.998	0.630

<sup>(1)</sup> Probability calculated as the percentage of 1000 simulations on which correct strategy mean was selected as the maximum.

Table 9b. The probability<sup>(1)</sup> of choosing the correct strategy for Analysis 4 when sample size is calculated to reject the null hypothesis for Analysis 2 (for a two-tailed test with power of 0.90 and type I error of 0.05)

Simulation Parameters			Simulation Results		
Effect size for Analysis 2	Non-response Rate	Sample size	Analysis 2 (power)	Analysis 4 (probability <sup>(1)</sup> )	Effect size for Analysis 4
0.2	0.5	2112	0.906	0.999	0.133
0.2	0.9	1174	0.895	0.716	0.054
0.5	0.5	338	0.895	0.997	0.372
0.5	0.9	188	0.901	0.978	0.420

<sup>(1)</sup> Probability calculated as the percentage of 1000 simulations on which correct strategy mean was selected as the maximum.

Table 9c. The probability<sup>(1)</sup> of choosing the correct strategy for Analysis 4 when sample size is calculated to reject the null hypothesis for Analysis 3 (for a two-tailed test with power of 0.90 and type I error of 0.05)

Simulation Parameters				Simulation Results		
Effect size for Analysis 3	Non-response rate	Sample size formula	Sample size	Analysis 3 (power)	Analysis 4 (probability <sup>(1)</sup> )	Effect size for Analysis 4
0.2	0.5	$N_{3a}$	1584	0.893	0.939	0.10
0.2	0.9	$N_{3a}$	2007	0.882	0.614	0.02
0.5	0.5	$N_{3a}$	254	0.896	0.976	0.25
0.5	0.9	$N_{3a}$	321	0.926	0.978	0.32
0.2	0.5	$N_{3b}$	2112	0.950	0.953	0.10
0.2	0.9	$N_{3b}$	2112	0.903	0.613	0.02
0.5	0.5	$N_{3b}$	338	0.973	0.989	0.25
0.5	0.9	$N_{3b}$	338	0.937	0.985	0.32

<sup>(1)</sup> Probability calculated as the percentage of 1000 simulations on which correct strategy mean was selected as the maximum.

## 5.2 Summary

Overall, the sample size formulae perform well even when the working assumptions are violated. Additionally, the performance of Analysis 4 is consistently good when sizing for all other analyses; this is most likely due to Analysis 4 requiring smaller sample sizes than the other analyses to achieve good results.

When planning a SMART trial similar to the one considered here, if one is primarily concerned with testing differences between prespecified strategy means, we would recommend using the less conservative formula  $N_{3a}$  if one has confidence in knowledge of the intermediate non-response rates. We recommend this in light of the considerable cost savings that can be accrued by using this approach, in comparison to the more conservative formula  $N_{3b}$ . We comment further on this topic in the discussion section.

## 6. DISCUSSION

In this paper, we demonstrated how a SMART trial can be used to answer research questions about both individual components of an adaptive treatment strategy and the treatment strategies as a whole. We presented statistical methodology to guide the design and analysis of a SMART trial. Two new methods for calculating the sample sizes for a SMART trial were presented. The first is for sizing a study when one is interested in testing the difference in two strategies that have different initial treatments; this formula incorporates knowledge about intermediate response rates. The second new sample size calculation is for sizing a study that has as its goal choosing the strategy that has the highest final outcome. We evaluated both of these methods and found that they performed well in simulations that covered a wide range of plausible scenarios.

The results for the sample size formula for choosing the strategy with the highest mean outcome are particularly promising. Clearly, the sample sizes calculated were more than adequate for the pattern where all of the strategy means varied from each other. Future work might focus on the development of separate sample sizes, depending on the pattern of strategy means expected to occur.

Several comments are in order regarding the violations of assumptions surrounding the values of the intermediate response rates when investigating sample size formula  $N_{3a}$  for Analysis 3. First, we violated the assumption of the homogeneity of response rates across first level treatments such that they differed by 10%; this violation is probably relatively mild. However, based on our experience, intermediate response rates differing by more than 10% in additions clinical trials would be rare. Future research is needed to examine the question regarding the extent to which intermediate response rates can be misspecified when utilizing this modified sample size formula. Clearly, for gross misspecifications, the trialist is probably better off with the more conservative sample size formula. However, the operationalization of “gross misspecification” needs further research.

In the additions, both clinical practice as well as trials are plagued with high percentages of patients dropping out of treatment. SMART designs hold the promise of helping researchers untangle factors affecting patient’s willingness to undergo treatment. In particular, one might be interested in varying subsequent treatments based on a combination of outcomes incorporating both likelihood of withdrawal as well as measures of continued drug use. In this case the subsequent treatments might include behavioral treatments designed to improve the patient’s motivation to adhere. Such a design might have four nominal categories of intermediate responses based on combinations of scores on the instrument predicting withdrawal and measuring ongoing drug use.

In this paper we focused on the simple design in which there are two options for non-responders and one option for responders. Clearly these results hold for the mirror design (one option for non-responders and two options for responders). An important step would be to generalize these results to other designs, such as designs in which there are

equal numbers of options for responders and non-responders or designs in which there are three randomizations. Also, we only consider final outcomes that are continuous. However, in substance abuse, the final outcome variable is often binary; sample size formulae are needed for this setting as well. Alternately, the outcome may be time varying such as the time-varying symptom levels; again, it is important to generalize the above results to this setting.

## APPENDIX

### Sample Size Formulae for Analysis 3

Here, we present the derivation of the sample size formulae for Analysis 3 using results from Murphy.<sup>9</sup>

Suppose we have data from a SMART design modeled after the one presented in Figure 2; that is, there are two options for the initial treatment followed by two treatment options for non-responders and one treatment option for responders. We use the same notation and assumptions listed in Section 4. Suppose that we are interested in comparing two strategies that have different initial treatments, strategies (a1, a2) and (b1, b2). Without loss of generality, let a1=1 and b1=0.

We will make the following working assumptions:

- the intermediate non-response rates for treatments at level 1 are equal and we denote this response rate by  $p$ ; i.e.  $p = \Pr\{R=1|A_1=1\} = \Pr\{R=1|A_1=0\}$ ,
- the marginal variances relevant to the analysis are equal and we denote this variance by  $s^2$ ; in this case,  $s^2 = \text{Var}[Y|A_1=a1, A_2=a2] = \text{Var}[Y|A_1=b1, A_2=b2]$ , and
- the sample sizes will be large enough so that  $\hat{\mu}_{(a1, a2)}$  is approximately normally distributed.

We will denote the mean outcome for strategy  $(A_1, A_2)$  by  $\mu_{(A1, A2)}$ .

The null hypothesis we are interested in testing is

$$H_0: \mu_{(1, a2)} - \mu_{(0, b2)} = 0$$

and the alternative of interest is

$$H_0: \mu_{(1, a2)} - \mu_{(0, b2)} = ds.$$

(Note that  $d$  is the standardized effect size.) As presented in Section 4.1, the test statistic for this hypothesis is

$$Z = \frac{\sqrt{N}(\hat{\mu}_{(1, a2)} - \hat{\mu}_{(0, b2)})}{\sqrt{\hat{\tau}_{(1, a2)}^2 + \hat{\tau}_{(0, b2)}^2}}$$

where  $\hat{\mu}_{(a1, a2)}$  and  $\hat{\tau}_{(a1, a2)}^2$  are as defined in Table 5; in large samples, this test statistic has a standard normal distribution under the null hypothesis<sup>28</sup>. Recall  $N$  is the total sample size for the trial. To find the required sample size  $N$  for a two-sided test with power  $1-\beta$  and size  $\alpha$ , we solve

$$\Pr[Z < -z_{\alpha/2} \text{ or } Z > z_{\alpha/2} \mid \mu_{(1, a2)} - \mu_{(0, b2)} = \delta\sigma] = 1 - \beta$$

for  $N$  where  $z_{\alpha/2}$  is the standard normal  $(1 - \alpha/2)$  percentile. So, we have

$$\Pr[Z < -z_{\alpha/2} \mid \mu_{(1,a2)} - \mu_{(0,b2)} = \delta\sigma] + \Pr[Z > z_{\alpha/2} \mid \mu_{(1,a2)} - \mu_{(0,b2)} = \delta\sigma] = 1 - \beta.$$

Without loss of generality, assume that  $\delta\sigma > 0$  so that

$$\Pr[Z < -z_{\alpha/2} \mid \mu_{(1,a2)} - \mu_{(0,b2)} = \delta\sigma] = 0$$

and

$$\Pr[Z > z_{\alpha/2} \mid \mu_{(1,a2)} - \mu_{(0,b2)} = \delta\sigma] = 1 - \beta.$$

Now,  $E[\hat{\mu}_{(1,a2)} - \hat{\mu}_{(0,b2)}] = \mu_{(1,a2)} - \mu_{(0,b2)}$ , and so we have

$$\Pr\left[\frac{\sqrt{N}(\hat{\mu}_{(1,a2)} - \hat{\mu}_{(0,b2)} - \delta\sigma)}{\sqrt{\hat{\tau}_{(1,a2)}^2 + \hat{\tau}_{(0,b2)}^2}} > z_{\alpha/2} - \frac{\delta\sigma\sqrt{N}}{\sqrt{(\hat{\tau}_{(1,a2)}^2 + \hat{\tau}_{(0,b2)}^2)}}\right] = 1 - \beta.$$

Note the distribution of

$$\frac{\sqrt{N}(\hat{\mu}_{(1,a2)} - \hat{\mu}_{(0,b2)} - \delta\sigma)}{\sqrt{\hat{\tau}_{(1,a2)}^2 + \hat{\tau}_{(0,b2)}^2}}$$

follows a standard normal distribution in large samples.<sup>28</sup> Define

$\tau_{(a1,a2)}^2 = \text{Var}[\sqrt{N}\hat{\mu}_{(a1,a2)}]$ . Also

$$\frac{\sqrt{\tau_{(1,a2)}^2 + \tau_{(0,b2)}^2}}{\sqrt{\hat{\tau}_{(1,a2)}^2 + \hat{\tau}_{(0,b2)}^2}}$$

is close to 1 in large samples<sup>9</sup> thus we have

$$z_{\beta} \approx -z_{\alpha/2} + \frac{\delta\sigma\sqrt{N}}{\sqrt{\tau_{(1,a2)}^2 + \tau_{(0,b2)}^2}} \quad (1)$$

Now, using Equation (10) in Murphy<sup>9</sup> for  $k = 2$  treatment levels,

$$\tau_{(a1,a2)}^2 = E_{a1,a2}\left[\frac{(Y - \mu_{(a1,a2)})^2}{\Pr(a1)\Pr(a2 \mid R, a1)}\right]$$

$$\begin{aligned}
&= E_{a1,a2} \left[ \frac{(Y - \mu_{(a1,a2)})^2}{\Pr(a1) \Pr(a2 | 1, a1)} \middle| R = 1 \right] \Pr_{a1}[R = 1] \\
&+ E_{a1,a2} \left[ \frac{(Y - \mu_{(a1,a2)})^2}{\Pr(a1) \Pr(a2 | 0, a1)} \middle| R = 0 \right] \Pr_{a1}[R = 0]
\end{aligned}$$

for all values of  $a1, a2$ ; the subscripts on  $E$  and  $\Pr$  (namely  $E_{a1,a2}$  and  $\Pr_{a1}$ ) indicate expectations and probabilities calculated as if all subjects were assigned  $a1$  as the initial treatment and then, if non-response, assigned treatment  $a2$ . If we are willing to make the assumption (\*) that

$$E_{a1,a2}[(Y - \mu_{(a1,a2)})^2 | R] \leq E_{a1,a2}[(Y - \mu_{(a1,a2)})^2]$$

for both  $R=1$  and  $R=0$ , (that is, the variability of the outcome around the strategy mean among either responders or non-responders is no more than the variance of the strategy mean) then

$$\begin{aligned}
\tau_{(a1,a2)}^2 &\leq E_{a1,a2}[(Y - \mu_{(a1,a2)})^2] \frac{\Pr_{a1}[R = 1]}{\Pr(a1) \Pr(a2 | 1, a1)} \\
&+ E_{a1,a2}[(Y - \mu_{(a1,a2)})^2] \frac{\Pr_{a1}[R = 0]}{\Pr(a1) \Pr(a2 | 0, a1)} \\
&\leq \sigma^2 \left( \frac{\Pr_{a1}[R = 1]}{\Pr(a1) \Pr(a2 | 1, a1)} + \frac{\Pr_{a1}[R = 0]}{\Pr(a1) \Pr(a2 | 0, a1)} \right)
\end{aligned}$$

So, we have that

$$\tau_{(a1,a2)}^2 \leq \sigma^2 \left( \frac{\Pr_{a1}[R = 1]}{\Pr(a1) \Pr(a2 | 1, a1)} + \frac{\Pr_{a1}[R = 0]}{\Pr(a1) \Pr(a2 | 0, a1)} \right). \quad (2)$$

Since (\*\*) patients are randomized equally to the two initial treatment options and since there are two treatment options for non-responders ( $R=1$ ) and one option for responders ( $R=0$ ), then for non-response rate  $p$ ,

$$\tau_{(a1,a2)}^2 \leq \sigma^2 * 2(2 * p + 1 * (1 - p)).$$

Rearranging Equation (1) gives us

$$2 * (2 * (2 * p + 1 * (1 - p)) * (1 / \delta)^2 (z_\beta + z_{\alpha/2})^2 = N_{3a}.$$

which is the sample size formula given in Section 4.2 that depends on the non-response rate  $p$ . If we assume that everyone is a non-responder, i.e.  $p=1$ , then going through the above arguments once again, we see that we do not need either of the two working assumptions (\* or \*\*) and we obtain the conservative sample size formula,  $N_{3b}$ :

$$2 * 4 * (1 / \delta)^2 (z_\beta + z_{\alpha/2})^2 = N_{3b}.$$

#### Sample Size Calculation for Analysis 4

We now present the algorithm for calculating the sample size for Analysis 4. As in the previous section, suppose we have data from a SMART design modeled after the one presented in Figure 2; we use the same notation and assumptions listed in Section 4. Suppose that we are interested in identifying the strategy that has the highest mean outcome. We will denote the mean outcome for strategy  $(A_1, A_2)$  by  $\mu_{(A_1, A_2)}$ .

We make the following assumptions:

- the marginal variances of the final outcome given the strategy are all equal and we denote this variance by  $s^2$ . This means that,  $s^2 = \text{Var}[Y|A_1=a_1, A_2=a_2]$  for all  $(a_1, a_2)$  in  $\{(1,1), (1,0), (0,1), (0,0)\}$
- The sample sizes will be large enough so that  $\hat{\mu}_{(a_1, a_2)}$  is approximately normally distributed.
- the correlation between the final outcome  $Y$  given treatment strategy  $(1, 1)$  and  $Y$  given treatment strategy  $(1, 0)$  is the same as the correlation between  $Y$  given treatment strategy  $(0, 1)$  and  $Y$  given treatment strategy  $(0, 0)$ ; we denote this identical correlation by  $\rho$ .

The correlation of the treatment strategies is directly related to the intermediate non-response rates. The final outcome under two different treatment strategies will be correlated to the extent that they share responders. For example, if the non-response rate for treatment  $A_1=1$  is 1, then everyone is a non-responder and the means calculated for  $Y$  given strategy  $(1, 1)$  and for  $Y$  given strategy  $(1, 0)$  will not share any responders to treatment  $A_1=1$ ; so, the correlation between the two strategies will be 0. On the other hand, if the non-response rate for treatment  $A_1=1$  is 0, then everyone is a responder to  $A_1=1$  and therefore, the mean outcome for strategy  $(1, 1)$  and strategy  $(1, 0)$  will be directly related, i.e. completely correlated. Two treatments strategies that each begin with a different initial treatment are not correlated since the strategies do not overlap, i.e. they do not share any subjects.

For the algorithm, the user must specify the following quantities:

- the value of  $s^2$ ,
- the desired standardized effect size,  $d$ , and
- the desired probability that the strategy estimated to have the largest mean outcome does in fact have the largest mean,  $p$ .

We assume that three of the strategies have the same mean and the one remaining strategy produces the largest mean, this is an extreme scenario in which it is most difficult to detect the presence of an effect. Without loss of generality, we choose strategy  $(1, 1)$  to have the largest mean.

Consider the following algorithm as a function of  $N$ :

1. For every value of  $\rho$  in  $\{0, 0.01, 0.02, \dots, 0.99, 1\}$  perform the following simulation:

- Generate  $K=20,000$  samples from a multivariate normal with

$$\blacksquare \text{ mean } M = \begin{bmatrix} \mu_{(1,1)} \\ \mu_{(1,0)} \\ \mu_{(0,1)} \\ \mu_{(0,0)} \end{bmatrix} = \begin{bmatrix} \delta * \sigma \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and}$$

$$\blacksquare \text{ covariance matrix } \Sigma = \frac{1}{N} \begin{bmatrix} 4\sigma^2 & \rho * 4\sigma^2 & 0 & 0 \\ \rho * 4\sigma^2 & 4\sigma^2 & 0 & 0 \\ 0 & 0 & 4\sigma^2 & \rho * 4\sigma^2 \\ 0 & 0 & \rho * 4\sigma^2 & 4\sigma^2 \end{bmatrix}$$

This gives us 20,000 samples,  $\{V_1, \dots, V_k, \dots, V_{20000}\}$ , where each  $V_k$  is a vector of four entries of outcomes, one from each treatment strategy. For example,  $V_k^t = [\hat{\mu}_{(1,1),k} \quad \hat{\mu}_{(1,0),k} \quad \hat{\mu}_{(0,1),k} \quad \hat{\mu}_{(0,0),k}]$ .

- Count how many times out of  $\{V_1, \dots, V_{20000}\}$  that  $\hat{\mu}_{(1,1),k}$  is highest; divide this count by 20,000, and call this value  $C_\rho(N)$ .  $C_\rho(N)$  is the estimate for the probability of correctly identifying the strategy with the highest mean.
2. At the end of step 1, we will have a value of  $C_\rho(N)$  for each  $\rho$  in  $\{0, 0.01, 0.02, \dots, 0.99, 1\}$ . Let  $\pi_N^* = \min_{\rho} C_\rho(N)$ ; the value of  $\pi_N^*$  is the lowest probability of detecting the best strategy mean.

Next, we perform a search over the space of possible values of  $N$  to find the value for which  $\pi_N^* = \pi$ .  $N_4$  is the value of  $N$  for which  $\pi_N^* = \pi$ .

### Generating the Final Outcome From a Gamma Distribution

Instead of generating the final outcome given a particular history  $(A1, R, A2)$  by a normal distribution with mean  $E[Y|A1, R, A2]$  and variance  $\text{Var}[Y|A1, R, A2]$ , we generate from a Gamma distribution with the same mean and variance. That is, instead of generating

$$Y|A1, R, A2 \sim N(E[Y|A1, R, A2], \text{Var}[Y|A1, R, A2])$$

we generate  $Y|A1, R, A2 \sim \text{Gamma}(a, b)$  where

$$a = \frac{E[Y | A1, R, A2]^2}{\text{Var}[Y | A1, R, A2]} \text{ and } b = \frac{\text{Var}[Y | A1, R, A2]}{E[Y | A1, R, A2]}.$$

The mean of this gamma distribution is  $ab$  and the variance is  $ab^2$ . The skewness of this distribution is calculated by  $2 * \sqrt{1/a}$ , in other words:

$$\text{skewness} = 2 * \sqrt{\frac{\text{Var}[Y | A1, R, A2]}{E[Y | A1, R, A2]^2}}.$$

The values for  $E[Y|A_1, R, A_2]$  and  $\text{Var}[Y|A_1, R, A_2]$  are prespecified for the simulation.

## REFERENCES

1. Carroll KM, Onken LS. Behavioral therapies for drug abuse. *American Journal of Psychiatry*. 2005; **162(8)**: 1452-1460.
2. Carroll KM. Recent advances in psychotherapy of addictive disorders. *Current Psychiatry Reports*. 2005; **7**: 329-336.
3. Ling W, Smith D. Buprenorphine: blending practice and research. *Journal of Substance Abuse Treatment*. 2002; **23**: 87-92.
4. Fiellin DA, Kleber H, Trumble-Hejduk JG, McLellan AT, Kosten TR. Consensus statement on office based treatment of opioid dependence using buprenorphine. *Journal of Substance Abuse Treatment*. 2004; **27**: 153-159.
5. McLellan AT. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. *Addiction*. 2002; **97**: 249-252.
6. McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness. Implications for treatment, insurance, and outcomes evaluation. *Journal of the American Medical Association*. 2000; **284(13)**: 1689-1695.
7. Greenhouse J, Stangl D, Kupfer D, Prien R. Methodological Issues in Maintenance Therapy Clinical Trials. *Archives of General Psychiatry*. 1991; **48(3)**: 313-318.
8. Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society*. 2003; **65**: 331-366.
9. Murphy SA. An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine*. 2005; **24**: 1455-1481.
10. Murphy SA, Lynch KG, Oslin DA, McKay JR, Tenhave T. Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence*. 2006. doi:10.1016/j.drugalcdep.2006.09.008
11. Murphy SA, Oslin DW, Rush AJ, Zhu J. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology*. 2007; **32**: 257-262.
12. Lavori PW, Dawson R. A design for testing clinical strategies: Biased adaptive within-subject randomization. *Journal of the Royal Statistical Association*. 2000; **163**: 29-38.
13. Lavori PW, Dawson R, Rush AJ. Flexible treatment strategies in chronic disease: clinical and research implications. *Biological Psychiatry*. 2000; **48**: 605-614.
14. Dawson R, Lavori PW. Comparison of designs for adaptive treatment strategies: baseline vs. adaptive randomization. *Journal of Statistical Planning and Inference*. 2003; **117**: 365-385.
15. Weiss R, Sharpe JP, Ling W. A Two-Phase Randomized Controlled Clinical Trial of Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling for Opioid Analgesic Dependence. National Institute on Drug Abuse Clinical Trials Network. 2006.
16. Pantalon MV, Fiellin DA, Schottenfeld RS, Gordon L, O'Connor PG. Manual for Enhanced Medical Management of Opioid Dependence with Buprenorphine. Yale University School of Medicine Primary Care and Substance Abuse Center, West Haven VA/CT Healthcare System. New Haven CT: Unpublished manuscript; 1999.
17. Fiellin D, Pantalon M, Schottenfeld R, Gordon L, O'Connor P. Manual for Standard Medical Management of Opioid Dependence with Buprenorphine. Yale University

- School of Medicine, Primary Care Center and Substance Abuse Center, West Haven VA/CT Healthcare System. New Haven CT: Unpublished manuscript; 1999.
18. Lavori PW, Rush JA, Wisniewski SR, Alpert J, Fava M, Kupfer DJ, et al. Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biological Psychiatry*. 2001; **50**: 792-801.
  19. Rush AJ, Crismon ML, Kashner TM, Toprac MG, Carmody TJ, Trivedi MH, et al. Texas medication algorithm project, phase 3 (TMAP-3): rationale and study design. *J. Clin. Psychiatry*. 2003; **64**(4): 357-369.
  20. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, et al. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophrenia Bulletin*. 2003; **29**(1): 15-31.
  21. Hoel P. Introduction to Mathematical Statistics. 5th ed. New York: John Wiley and Sons; 1984.
  22. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.; 1988.
  23. Jennison C, Turnbull B. Group Sequential Methods with Applications to Clinical Trials. Boca Raton, FL City: Chapman & Hall; 2000.
  24. Gandhi DH, Jaffe JH, McNary S, Kavanagh GJ, Hayes M, Currens M. Short-term outcomes after brief ambulatory opioid detoxification with buprenorphine in young heroin users. *Addiction*. 2003; **98**: 453-462.
  25. Fiellin DA, Pantalon MV, Chawarski MC, Moore BA, Sullivan LE, O'Connor PG, et al. Counseling plus buprenorphine-naloxone maintenance therapy for opioid dependence. *The New England Journal of Medicine*. 2006; **355**(4): 365-374.
  26. Ling W, Amass L, Shoptow S, Annon JJ, Hillhouse M, Babcock D, et al. A multi-center randomized trial of buprenorphine-naloxone versus clonidine for opioid detoxification: findings from the National Institute on Drug Abuse Clinical Trials Network. *Addiction*. 2005; **100**: 1090-1100.
  27. Scott A, Levy J, Murphy SA. Evaluation of Sample Size Formulae for Developing Adaptive Treatment Strategies Using a SMART Design. University Park, PA: The Pennsylvania State University, The Methodology Center; 2007 May.
  28. Murphy SA, Van Der Laan MJ, Robins JM; Conduct Problems Prevention Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*. 2001; **96**(456): 1410-1423.