8

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

ALENA I. OETTING, JANET A. LEVY, ROGER D. WEISS, AND SUSAN A. MURPHY

Introduction

The past two decades have brought new pharmacotherapies as well as behavioral therapies to the field of drug-addiction treatment (Carroll & Onken, 2005; Carroll, 2005; Ling & Smith, 2002; Fiellin, Kleber, Trumble-Hejduk, McLellan, & Kosten, 2004). Despite this progress, the treatment of addiction in clinical practice often 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 specific treatment. Second, as many authors have pointed out (McLellan, 2002; McLellan, Lewis, O'Brien, & Kleber, 2000), addiction is often a chronic condition, with symptoms waxing and waning over time. Third, relapse is common. Therefore, the clinician is faced with, first, finding a sequence of treatments that works initially to stabilize the patient and, next, deciding which types of treatments will prevent relapse in the longer term. To inform this sequential clinical decision making, *adaptive treatment strategies*, that is, treatment strategies shaped by individual patient characteristics or patient responses to prior treatments, have been proposed (Greenhouse, Stangl, Kupfer, & Prien, 1991; Murphy, 2003, 2005; Murphy, Lynch, Oslin, McKay, & Tenhave, 2006; Murphy, Oslin, Rush, & Zhu, 2007; Lavori & Dawson, 2000; Lavori, Dawson, & Rush, 2000; Dawson & Lavori, 2003).

Here is an example of an adaptive treatment strategy for prescription opioid dependence, modeled with modifications after a trial currently in progress within the Clinical Trials Network of the National Institute on Drug Abuse (Weiss, Sharpe, & Ling, 2010).

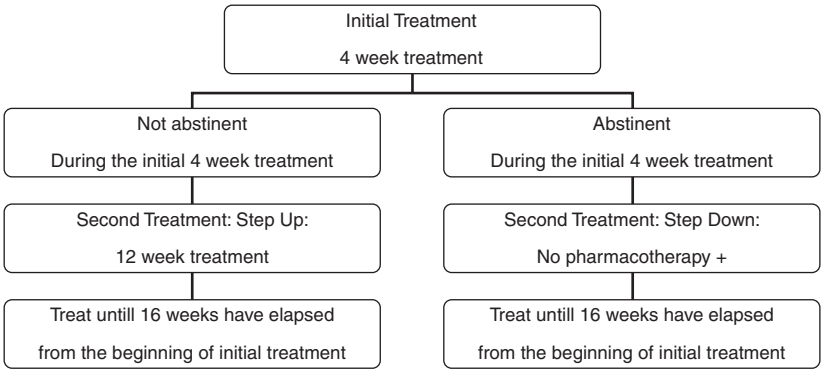

Figure 8.1. An adaptive treatment strategy for prescription opioid dependence.

Example
First, provide all patients with a 4-week course of buprenorphine/naloxone (Bup/Nx) plus medical management (MM) plus individual drug counseling (IDC) (Fiellin, Pantalon, Schottenfeld, Gordon, & O'Connor, 1999), culminating in a taper of the Bup/Nx. If at any time during these 4 weeks the patient meets the criterion for nonresponse,[1] a second, longer treatment with Bup/Nx (12 weeks) is provided, accompanied by MM and cognitive behavior therapy (CBT). However, if the patient remains abstinent[2] from opioid use during those 4 weeks, that is, responds to initial treatment, provide 12 additional weeks of relapse prevention therapy (RPT).

A patient whose treatment is consistent with this strategy experiences one of two *sequences* of two treatments, depicted in Figure 8–1. The two sequences are

1. Four-week Bup/Nx treatment plus MM plus IDC, then if the criterion for nonresponse is met, a subsequent 12-week Bup/Nx treatment plus MM plus CBT.
2. Four-week Bup/Nx treatment plus MM plus IDC, then if abstinence is achieved, a subsequent 12 weeks of RPT.

1. Response to initial treatment is abstinence from opioid use during these first 4 weeks. Nonresponse is defined as any opioid use during these first 4 weeks
2. Abstinence might be operationalized using a criterion based on self-report of opioid use and urine screens.

This strategy might be intended to maximize the number of days the patient remains abstinent (as confirmed by a combination of urine screens and self-report) over the duration of treatment.

Throughout, we use this hypothetical prescription opioid dependence example to make the ideas concrete. In the next section, several research questions useful in guiding the development of an adaptive treatment strategy are discussed. Next, we review the sequential multiple assignment trial (SMART), which is an experimental design developed to answer these questions. We present statistical methodology for analyzing 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 a discussion of suggested areas for future research.

this cite should be moved to

Research Questions to Refine an Adaptive Treatment Strategy

Continuing with the prescription opioid dependence example, we might ask if we could begin with a less intensive behavioral therapy. For example, standard MM (Lavori et al., 2000), which is less burdensome than IDC and focuses primarily on medication adherence, might be sufficiently effective for a large majority of patients; that is, we might ask, In the context of the specified options for further treatment, does the addition of IDC to MM result in a better long-term outcome than the use of MM as the sole accompanying behavioral therapy? Alternatively, if we focus on the behavioral therapy accompanying the second longer 12-week treatment, we might ask, Among subjects who did not respond to one of the initial treatments, which accompanying behavioral therapy is better for the secondary treatment: MM+IDC or MM+CBT?

On the other hand, instead of focusing on a particular treatment component within strategies, we may be interested in comparing entire adaptive treatment strategies. Consider the strategies in Table 8–1. Suppose we are interested in comparing two of these treatment strategies. If the strategies begin with the same initial treatment, then the comparison reduces to a comparison of the two secondary treatments; in our example, a comparison of strategy C with strategy D is obtained by comparing MM+IDC with MM+CBT among nonresponders to MM alone. We also might compare two strategies with different initial treatments. For example, in some settings, CBT may be the preferred behavioral therapy to use with longer treatments; thus, we might ask, if we are going to provide MM+CBT for nonresponders

Table 8.1 Potential Strategies to Consider for the Treatment of Prescription Opioid Dependence

Initial Treatment	*Response to Initial Treatment*	*Secondary Treatment*
Strategy A: Begin with Bup/Nx+MM+IDC; if nonresponse, provide Bup/Nx+MM+CBT; if response, provide RPT		
4-week Bup/Nx treatment + MM+IDC	Not abstinent	12-week Bup/Nx treatment + MM+CBT
4-week Bup/Nx treatment + MM+IDC	Abstinent	RPT
Strategy B: Begin with Bup/Nx+MM+IDC; if nonresponse, provide Bup/Nx+MM+IDC; if response, provide RPT		
4-week Bup/Nx treatment + MM+IDC	Not abstinent	12-week Bup/Nx treatment + MM + IDC
4-week Bup/Nx treatment + MM+IDC	Abstinent	RPT
Strategy C: Begin with Bup/Nx+MM; if nonresponse, provide Bup/Nx+MM+CBT; if response, provide RPT		
4-week Bup/Nx treatment + MM	Not abstinent	12-week Bup/Nx treatment + MM + CBT
4-week Bup/Nx treatment + MM	Abstinent	RPT
Strategy D: Begin with Bup/Nx+MM; if nonresponse, provide Bup/Nx+MM+IDC; if response, provide RPT		
4-week Bup/Nx treatment + MM	Not abstinent	12-week Bup/Nx treatment + MM + IDC
4-week Bup/Nx treatment + MM	Abstinent	RPT

to the initial treatment and RPT to responders to the initial treatment, Which is the best initial behavioral treatment: MM+IDC or MM? This is a comparison of strategies A and C. Alternately, we might wish to identify which of the four strategies results in the best long-term outcome (here, the highest number of days abstinent). Note that the behavioral therapies and pharmacotherapies are illustrative and were selected to enhance the concreteness of this example; of course, other selections are possible.

These research questions can be classified into one of four general types, as summarized in Table 8–2. The SMART experimental design discussed in the next section is particularly suited to addressing these types of questions.

A SMART Experimental Design and the Development of Adaptive Treatment Strategies

Traditional experimental trials typically evaluate a single treatment with no manipulation or control of preceding or subsequent treatments. In contrast, the SMART design provides data that can be used both to assess the efficacy of each treatment within a sequence and to compare the effectiveness of strategies as a whole. A further rationale for the SMART design can be found in Murphy et al. (2006, 2007). We focus on SMART designs in which there are two initial treatment options, then two treatment options for initial nonresponders (alternately, initial responders) and one treatment option for initial treatment responders (alternately, initial nonresponders). 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 those employed by the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) (Rush et al., 2003) and the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) (Stroup et al., 2003); additionally, two SMART trials of this type are currently in the field (D. Oslin, personal communication, 2007; W. Pelham, personal communication, 2006).

Data from this experimental design can be used to address questions from each type in Table 8–2. Because SMART specifies sequences of treatments, it allows us to determine the effectiveness of one of the treatment components in the presence of either preceding or subsequent treatments; that is, it addresses questions of both types 1 and 2. Also, the use of randomization supports causal inferences about the relative effectiveness of different treatment strategies, as in questions of types 3 and 4.

Returning to the prescription opioid dependence example, a useful SMART design is provided in Figure 8–2. Consider a question of the first type from Table 8–2. An example is, In the context of the specified options for further treatment, does the addition of IDC to MM result in a better long-term outcome than the use of MM as the sole accompanying behavioral therapy? This question is answered by comparing the pooled outcomes of subgroups 1,2,–3 with those of subgroups 4,5,–6. This is the main effect of the initial behavioral treatment. Note that to estimate the main effect of the initial behavioral treatment, we require outcomes from not only initial nonresponders but also initial responders. Clinically, this makes sense as a particular initial treatment may lead to a good response but this response may not be as durable as other initial treatments. Next, consider a question of the second type, such as, Among those who did not respond to one of the initial treatments, which is the better subsequent behavioral treatment: MM+IDC or MM+CBT? This question is addressed by pooling outcome data from subgroups 1 and 4 and comparing the resulting mean to the

delete hyphens

Table 8.2 Four General Types of Research Questions

Question	*Type of Analysis Required to Answer Question*	*Research Question*
Two questions that concern components of adaptive treatment strategies		
1	Hypothesis test	Initial treatment effect: What is the effect of initial treatment on long-term outcome in the context of the specified secondary treatments? In other words, what is the main effect of initial treatment?
2	Hypothesis test	Secondary treatment effect: Considering only those who did (or did not) respond to one of the initial treatments, what is the best secondary treatment? In other words, what is the main effect of secondary treatment for responders (or nonresponders)?
Two questions that concern whole adaptive treatment strategies		
3	Hypothesis test	Comparing strategy effects: What is the difference in the long-term outcome between two treatment strategies that begin with a different initial treatment?
4	Estimation	Choosing the overall best strategy: Which treatment strategy produces the best long-term outcome?

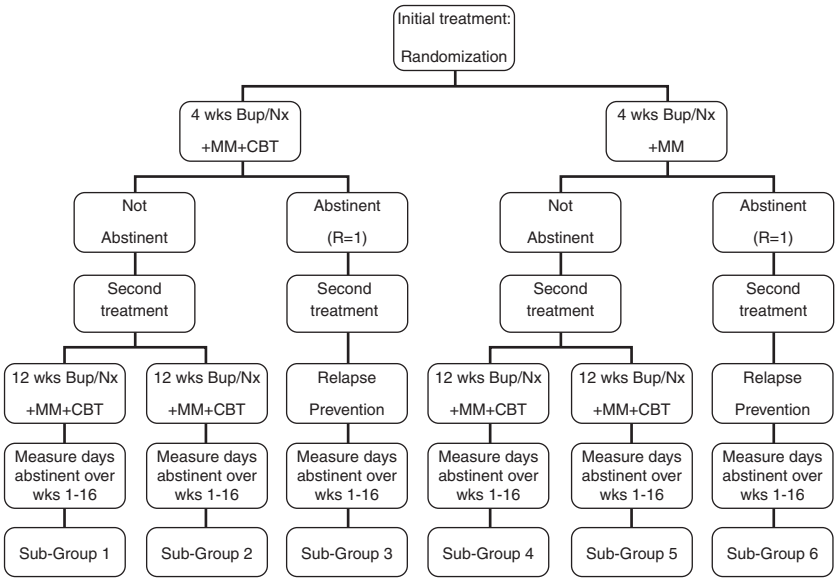

Figure 8.2. SMART study design to develop adaptive treatment strategies for prescription opioid dependence.

pooled outcome data of subgroups 2 and 5. This is the main effect of the secondary behavioral treatment among those not abstinent during the initial 4-week treatment.

An example of the third type question would be to test whether strategies A and C in Table 8–1 result in different outcomes; to form this test, we use appropriately weighted outcomes from subgroups 1 and 3 to form an average outcome for strategy A and appropriately weighted outcomes from subgroups 4 and 6 to form an average outcome for strategy C (an alternate example would concern strategies B and D; see the next section for formulae). Note that to compare strategies, we require outcomes from both initial responders as well as initial nonresponders (e.g., subgroup 3 in addition to subgroup 1 and subgroup 6 in addition to subgroup 4). The fourth type of question concerns the estimation of the best of the strategies. To choose the best strategy overall, we follow a similar ''weighting'' process to form the average outcome for each of the four strategies (A, B, C, D) and then designate as the best strategy the one that is associated with the highest average outcome.

Test Statistics and Sample Size Formulae

In this section, we provide the test statistics and sample size formulae for the four types of research questions summarized in Table 8–2. We assume that subjects are randomized equally to the two treatment options at each step. We use the following notation: A_1 is the indicator for initial treatment, R denotes the response to the initial treatment (response = 1 and nonresponse = 0), A_2 is the treatment indicator for secondary treatment, and Y is the outcome. In our prescription opioid dependence example, the values for these variables are as follows: A_1 is 1 if the initial treatment uses MM+IDC and 0 otherwise, A_2 is 1 if the secondary treatment for nonresponders uses MM+CBT and 0 otherwise, and Y is the number of days the subject remained abstinent over the 16-week study period.

Statistics for Addressing the Different Research Questions

The test statistics for questions 1–3 of Table 8–2 are presented in Table 8–3; the method for addressing question 4 is also given in Table 8–3. The test statistics for questions 1 and 2 are the standard test statistics for a two-group comparison with large samples (Hoel, 1984) and are not unique to the SMART design. The estimator of a strategy mean, used for both questions 3 and 4, as well as the test statistic for question 3 are given in Murphy (2005). In large samples, the three test statistics corresponding to questions 1–3 are

Table 8.3 Test Statistics for Each of the Possible Questions

Type of Question	*Test Statistic*
1[a]	$Z = \frac{(\overline{Y}_{A1=1} - \overline{Y}_{A1=0})}{\sqrt{\frac{S^2_{A1=1}}{N_{A1=1}} + \frac{S^2_{A1=0}}{N_{A1=0}}}}$ where $N_{A1=i}$ denotes the number of subjects who received i as the initial treatment
2[a]	$Z = \frac{(\overline{Y}_{R=0,\ A2=1} - \overline{Y}_{R=0,\ A2=0})}{\sqrt{\frac{S^2_{R=0,\ A2=1}}{N_{R=0,\ A2=1}} + \frac{S^2_{R=0,\ A2=0}}{N_{R=0,\ A2=0}}}}$ where $N_{R=0,\ A2=i}$ denotes the number of nonresponders who received i as the secondary treatment
3[b]	$Z = \frac{\sqrt{N}(\hat{\mu}_{A1=1,\ A2=a2} - \hat{\mu}_{A1=0,\ A2=b2})}{\sqrt{\hat{\tau}^2_{A1=1,\ A2=a2} + \hat{\tau}^2_{A1=0,\ A2=b2}}}$ where N is the total number of subjects and a2 and b2 are the secondary treatments in the two prespecified strategies being compared
4	Choose largest of $\hat{\mu}_{A1=1,\ A2=1}$, $\hat{\mu}_{A1=0,\ A2=1}$, $\hat{\mu}_{A1=1,\ A2=0}$, $\hat{\mu}_{A1=0,\ A2=0}$

[a]The subscripts on Y and S^2 denote groups of subjects. For example $Y_{R=0,A2=1}$ is the average outcome for subjects who do not respond initially (R = 0) and are assigned $A_2 = 1$. $S^2_{R=0,A2=1}$ is the sample variance of the outcome for subjects who do not respond initially (R = 0) and are assigned $A_2 = 1$. Similarly, the subscript on N denotes the group of subjects.
[b]$\hat{\mu}$ is an estimator of the mean outcome and $\hat{\tau}^2$ is the associated variance estimator for a particular strategy. Here, the subscript denotes the strategy. The formulae for $\hat{\mu}$ and $\hat{\tau}^2$ are in Table 8–4.

normally distributed (with mean zero under the null hypothesis of no effect). In Tables 8–3, 8–4, and 8–5, specific values of A_i are denoted by a_i and b_i, where i indicates the initial treatment (i = 1) or secondary treatment (i = 2); these specific values are either 1 or 0.

Sample Size Calculations

In the following, all sample size formulae assume a two-tailed z-test. Let α be the desired size of the hypothesis test, 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.

Table 8.4 Estimators for Strategy Means and for Variance of Estimator of Strategy Means

Strategy Sequence (a_1, a_2)	*Estimator for Strategy Mean:* $\hat{\mu}_{A1=a1,\ A2=a2} = \dfrac{\sum_{i=1}^{N} W_i(a_1, a_2)Y_i}{\sum_{i=1}^{N} W_i(a_1, a_2)_i}$	*N*Estimator for Variance of Estimator of Strategy Mean:* $\hat{\tau}^2_{A1=a1,\ A2=a2} = \frac{1}{N}\sum_{i=1}^{N} W_i(a_1, a_2)^2 \times (Y_i - \hat{\mu}_{A1=a1,\ A2=a2})^2$
(1, 1)	$W_i(1,1) = \frac{A_{1i}}{.5} * \left((1 - R_i) * \frac{A_{2i}}{.5} + R_i\right)$	
(1, 0)	$W_i(1,0) = \frac{A_{1i}}{.5} * \left((1 - R_i) * \frac{(1 - A_{2i})}{.5} + R_i\right)$	
(0, 1)	$W_i(0,1) = \frac{(1 - A_{1i})}{.5} * \left((1 - R_i) * \frac{A_{2i}}{.5} + R_i\right)$	
(0, 0)	$W_i(0,0) = \frac{(1 - A_{1i})}{.5} * \left((1 - R_i) * \frac{(1 - A_{2i})}{.5} + R_i\right)$	

Data for subject *i* are 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 the section Test Statistics and Sample Size Formulae and *N* is the total sample size.

In order to calculate the sample size, one must also input the desired detectable standardized effect size. We denote the standardized effect size by δ and use the definition found in Cohen (1988). The standardized effect sizes for the various research questions we are considering are summarized in Table 8–5.

The sample size formulae for questions 1 and 2 are standard formulae (Jennison & Turnbull, 2000) and assume an equal number in each of the two groups being compared. Given desired levels of size, power, and standardized effect size, the total sample size required for question 1 is

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

The sample size formula for question 2 requires the user to postulate the initial response rate, which is used to provide the number of subjects who will be randomized to secondary treatments. The sample size formula uses the working assumption that the initial response rates are equal; that is, subjects respond to initial treatment at the same rate regardless of the particular initial treatment, $p = Pr[R = 1|A_1 = 1] = Pr[R = 1|A_1 = 0]$. This working assumption is used only to size the SMART and is not used to analyze the

Table 8.5 Standardized Effect Sizes for Addressing the Four Questions in Table 8–2

Research Question	*Formula for Standardized Effect Size* δ
1	$\delta = \dfrac{E[Y \mid A_1 = 1] - E[Y \mid A_1 = 0]}{\sqrt{\frac{Var[Y \mid A_1 = 1] + Var[Y \mid A_1 = 0]}{2}}}$
2	$\delta = \dfrac{E[Y \mid R = 0,\ A_2 = 1] - E[Y \mid R = 0,\ A_2 = 0]}{\sqrt{\frac{Var[Y \mid R = 0,\ A_2 = 0] + Var[Y \mid R = 0,\ A_2 = 0]}{2}}}$
3	$\delta = \dfrac{E[Y \mid A_1 = 1,\ A_2 = a_2] - E[Y \mid A_1 = 0,\ A_2 = b_2]}{\sqrt{\frac{Var[Y \mid A_1 = 1,\ A_2 = a_2] + Var[Y \mid A_1 = 0,\ A_2 = b_2]}{2}}}$ where a_2 and b_2 are the secondary treatment assignments of A_2
4	$\delta = \dfrac{E[Y \mid A_1 = a_1,\ A_2 = a_2] - E[Y \mid A_1 = b_1,\ A_2 = b_2]}{\sqrt{\frac{Var[Y \mid A_1 = a_1,\ A_2 = a_2] + Var[Y \mid A_1 = b_1,\ A_2 = b_2]}{2}}}$ where (a_1, a_2) = strategy with the highest mean outcome, (b_1, b_2) = strategy with the next highest mean outcome; a_i and b_i indicate specific values of A_i, $i = 1,2$

data from it, as can be seen from Table 8–3. The formula for the total required sample size for question 2 is

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

When calculating the sample sizes to test question 3, two different sample size formulae can be used: one that inputs the postulated initial response rate and one that does not. The formula that uses a guess of the initial response rate makes two working assumptions. First, the response rates are equal for both initial treatments (denoted by p), and second, the variability of the outcome Y around the strategy mean ($A_1 = 1$, $A_2 = a_2$), among either initial responders or nonresponders, is less than the variance of the strategy mean and similarly for strategy ($A_1 = 0$, $A_2 = b_2$). This formula is

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

The second formula does not require either of these two working assumptions; it specifies the sample size required if the response rates are both 0, a ''worst-case scenario.'' This conservative sample size formula for addressing question 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 addressing question 3 in the next section. See the Appendix for a derivation of these formulae.

The method for finding the sample size for question 4 relies on an algorithm rather than a formula; we will refer to the resulting sample size as N_4. Since question 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. This algorithm makes the working assumption that $\sigma^2 = \text{Var}[Y|A_1 = a_1,\ A_2 = a_2]$ is the same for all strategies. The algorithm uses an idea similar to the one used to derive the sample size formula for question 3 that is invariant to the 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 question 4 finds the sample sizes that correspond to the range of 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. The online sample size calculator for question 4 can be found at http://methodology.psu.edu/index.php/smart-sample-size-calculation.

change to http://methodology media.psu.edu/smart/samplesize

Example sample sizes are given in Table 8–6. Note that as the response rate decreases, the required sample sizes for question 3 (e.g., comparing two strategies that have different initial treatments) 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 nonresponse rate is 90%. In the former case, if n subjects are assigned to treatment 1 initially and 90% respond (i.e., 10% do not respond), then the resulting sample size for strategy (1, 1) is $0.9 * n + ½ * 0.1 * n = 0.95 * n$. The ½ occurs due to the second randomization of nonresponders between the two secondary treatments. On the other hand, if only 10% respond (i.e., 90% do not respond), then the resulting sample size for strategy (1, 1) is $0.1 * n + ½ * 0.9 * n = 0.55 * n$, which is less than $0.95 * n$. Thus, the lower the expected response rate, the larger the initial sample size required for a given power to differentiate between two strategies. This result occurs because the number of treatment options (two options) for nonresponders is greater than the number of treatment options for responders (only one).

Consider the prescription opioid dependence example. Suppose we are particularly interested in investigating whether MM+CBT or MM+IDC is best for subjects who do not respond to their initial treatment. This is a question of type 2. Thus, in order to ascertain the sample size for the SMART design in Figure 8–2, we use formula N_2. Suppose we decide to

Table 8.6 Example Sample Sizes: All Entries Are for Total Sample Size

Desired Size[(1)] α	*Desired Power*[(2)] $1-\beta$	*Standardized Effect Size* δ	*Initial Response Rate*[(3)] p	*Research Question* 1	2	3 (varies by p)	3 (invariant to p)	4
$\alpha = 0.10$								
	$\beta = 0.20$							
		$\delta = 0.20$						
			$p = 0.5$	620	1,240	930	1240	358
			$p = 0.1$	620	689	1,178	1,240	358
		$\delta = 0.50$						
			$p = 0.5$	99	198	149	198	59
			$p = 0.1$	99	110	188	198	59
	$\beta = 0.10$							
		$\delta = 0.20$						
			$p = 0.5$	864	1,728	1,297	1,729	608
			$p = 0.1$	864	960	1,642	1,729	608
		$\delta = 0.50$						
			$p = 0.5$	138	277	207	277	97
			$p = 0.1$	138	154	263	277	97
$\alpha = 0.05$								
	$\beta = 0.20$							
		$\delta = 0.20$						
			$p = 0.5$	784	1,568	1,176	1,568	358
			$p = 0.1$	784	871	1,490	1,568	358
		$\delta = 0.50$						
			$p = 0.5$	125	251	188	251	59
			$p = 0.1$	125	139	238	251	59
	$\beta = 0.10$							
		$\delta = 0.20$						
			$p = 0.5$	1,056	2,112	1,584	2,112	608
			$p = 0.1$	1,056	1,174	2,007	2,112	608
		$\delta = 0.50$						
			$p = 0.5$	169	338	254	338	97
			$p = 0.1$	169	188	321	338	97

[a]All entries assume that each statistical test is two-tailed; the sample size for question 4 does not vary by α since this is not a hypothesis test.

[b]In question 4, we choose the sample size so that the probability that the treatment strategy with the highest mean has the highest estimated mean is $1-\beta$.

[c]The sample size formulae assume that the response rates to initial treatments are equal: $p = Pr[R=1|A_1=1] = Pr[R=1|A_1=0]$.

size the trial to detect a standardized effect size of 0.2 between the two secondary treatments with the power and size of the (two-tailed) test at 0.80 and 0.05, respectively. After surveying the literature and discussing the issue with colleagues, suppose we decide that the response rate for the two initial treatments will be approximately 0.10 ($p = 0.10$). The number of subjects required for this trial is then $N_2 = 2 * 2 * (z_{\alpha/2} + z_{\beta})^2 * (1/\delta)^2/(1 - p) = 4 * (z_{0.05/2} + z_{0.2})^2 * (1/0.2)^2/0.9 = 871$. Furthermore, as secondary objectives, suppose we are interested in comparing strategy A:—Begin with MM+IDC; if nonresponse, provide MM+CBT; if response, provide RPT—with D—Begin with MM; if nonresponse, provide MM+IDC; if response, provide RPT—(corresponding to a specific example of question 3) and in choosing the best strategy overall (question 4). Using the same input values for the parameters and looking at Table 8–6, we see that the sample size required for question 3 is about twice as much as that required for question 2. Thus, unless we are willing and able to double our sample size, we realize that a comparison of strategies A and D will have low power. However, the sample size for question 4 is only 358 (using desired probability of 0.80), so we will be able to answer the secondary objective of choosing the best strategy with 80% probability.

Suppose that we conduct the trial with 871 subjects. The hypothetical data set[3] and SAS code for calculating the following values can be found at http://www.stat.lsa.umich.edu/~samurphy/papers/APPAPaper/. For question 2, the value of the z-statistic is

$$\frac{(\overline{Y}_{R=0,\ A2=1} - \overline{Y}_{R=0,\ A2=0})}{\sqrt{\frac{S^2_{R=0,\ A2=1}}{N_{R=0,\ A2=1}} + \frac{S^2_{R=0,\ A2=0}}{N_{R=0,\ A2=0}}}} = \frac{(5.8619 - 4.3135)}{\sqrt{\frac{109.3975}{391} + \frac{98.5540}{396}}} = 2.1296,$$

which has a two-sided p value of 0.0332. Using the formulae in Table 8–4, we get the following estimates for the strategy means:

$$[\hat{\mu}_{(1,1)},\quad \hat{\mu}_{(1,0)},\quad \hat{\mu}_{(0,1)},\quad \hat{\mu}_{(0,0)}] = [7.1246 \quad 4.9994 \quad 6.3285 \quad 5.6364].$$

3. We generated this hypothetical data so that the true underlying effect size for question 2 is 0.2, the true effect size for question 3 is 0.2, and the strategy with the highest mean in truth is (1, 1), with an effect size of 0.1. Furthermore, the true response rates for the initial treatments are 0.05 for $A_1 = 0$ and 0.15 for $A_1 = 1$. When we considered 1,000 similar data sets, we found that the analysis for question 2 led to significant results 78% of the time and the analysis for question 3 led to significant results 54% of the time. The latter result and the fact that we did not detect an effect for question 3 in the analysis is unsurprising, considering that we have half the sample size required to detect an effect size of 0.2. Furthermore, across the 1,000 similar simulated data sets the best strategy (1, 1) was detected 86% of the time.

The corresponding estimates for the variances of the estimates of the strategy means are

$$[\,\tau^2_{(1,1)},\quad \tau^2_{(1,0)},\quad \tau^2_{(0,1)},\quad \tau^2_{(0,0)}\,] = [396.4555\quad 352.8471\quad 456.5727\quad 441.0138].$$

Using these estimates, we calculate the value of the corresponding z-statistic for question 3:

$$\frac{\sqrt{\mathrm{N}}(\hat{\mu}_{\mathrm{A1=1,\ A2=1}} - \hat{\mu}_{\mathrm{A1=0,\ A2=0}})}{\sqrt{\tau^2_{\mathrm{A1=1,\ A2=1}} + \tau^2_{\mathrm{A1=0,\ A2=0}}}} = \frac{\sqrt{871}(7.1246 - 5.6364)}{\sqrt{396.4555 + 441.0138}} = 1.5178,$$

which has a two-sided p value of 0.1291, which leads us not to reject the null hypothesis that the two strategies are equal. For question 4, we choose (1, 1) as the best strategy, which corresponds to the strategy:

1. First, supplement the initial 4-week Bup/Nx treatment with MM+IDC.
2. For those who respond, provide RPT. For those who do not respond, continue the Bup/Nx treatment for 12 weeks but switch the accompanying behavioral treatment to MM+CBT.

Evaluation of Sample Size Formulae Via Simulation

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

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 questions 1–3 is used to size the trial and a 0.90 probability of choosing the best strategy for question 4 when it is used to size the trial; these sample sizes are shown in Table 8–6. For questions 1–3, power is estimated by the proportion of times out of 1,000 simulations that the null hypothesis is correctly rejected; for question 4, the probability of choosing the best strategy is estimated by the proportion of times out of 1,000 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 (1988) 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 (Gandhi et al., 2003; Fiellin et al., 2006; Ling et al., 2005). For example, the simulated data exhibit initial response rates (i.e., the proportion of simulated subjects with $R = 1$) of 0.5 and 0.1, and the mean outcome for the responders is higher than for nonresponders.

For question 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 ($A_1 = 0$, $A_2 = 0$); these are strategies A and D, respectively, from Table 8–1. For the simulations to evaluate the robustness of the sample size calculation for question 4, we choose strategy A 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 strategy A 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 formulae, we calculate a sample size given by the relevant formula in Sample Size Calculations and then simulate data sets of this sample size. However, the simulated data will not satisfy the working assumptions in one of the following ways:

- the intermediate response rates to initial treatments are unequal, that is, $Pr[R = 1|A_1 = 1] \neq Pr[R = 1|A_1 = 0]$
- the variances relevant to the question are unequal (for question 4 only)
- the distribution of the final outcome, Y, is right-skewed (thus, for a given sample size, the test statistic is more likely to have a nonnormal distribution).

We also assess the power of question 4 when it is not used in sizing the trial. For each of the types of research questions in Table 8–2, we generate a data set that follows the working assumptions for the sample size formula for that question (e.g., use N_2 to size the study to test the effect of the second treatment on the mean outcome) and then perform question 4 on the data and estimate the probability of choosing the correct strategy with the highest mean outcome.

The descriptions of the simulation designs for each of questions 1–4 as well as the parameters for all of the different generative models can be found at http://www.stat.lsa.umich.edu/~samurphy/papers/APPAPaper/.

Robustness of the New Sample Size Formulae

As previously mentioned, since the sample size formulae for questions 1 and 2 are standard, we will focus on evaluating the newly developed sample size formulae for questions 3 and 4. Table 8–7a and b provides the results of the simulations designed to evaluate the sample size formulae for questions 3 and 4, respectively.

Considering Table 8–7a, we see that the question 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 initial response rates or any of the three violations of the working assumptions. Power levels were less robust when the sample sizes were smaller (i.e., for the 0.50 effect size). For example, when the initial response rates are not equal, the resulting power is lower than 0.90 in the rows using an assumed response rate of 0.5. The more conservative sample size formula, N_{3b}, performed well in all scenarios, regardless of response rate or the presence of any of the three violations to underlying assumptions. As the response rate approaches 0, the sample sizes are less conservative but the results for power remain within a 95% confidence interval of 0.90.

In Table 8–7b, the conservatism of the sample size calculation N_4 (associated with question 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.

Table 8.7a Investigation of Sample Size Assumption Violations for Question 3, Comparing Strategies A and D

Simulation Parameters				*Simulation Results (Power)*		
Effect Size	*Initial Response Rate (Default)*	*Sample Size Formula*	*Total Sample Size*	*Default Working Assumptions Are Correct*	*Unequal Initial Response Rates*	*Non-Normal Outcome Y*
0.2	0.5	N_{3a}	1,584	0.893	0.902	0.882
0.2	0.1	N_{3a}	2,007	0.882	0.910	0.877[a]
0.5	0.5	N_{3a}	254	0.896	0.864[a]	0.851[a]
0.5	0.1	N_{3a}	321	0.926[a]	0.886	0.898
0.2	0.5	N_{3b}	2,112	0.950[a]	0.958[a]	0.974[a]
0.2	0.1	N_{3b}	2,112	0.903	0.934[a]	0.898
0.5	0.5	N_{3b}	338	0.973[a]	0.938[a]	0.916
0.5	0.1	N_{3b}	338	0.937[a]	0.890	0.922[a]

The power to reject the null hypothesis for question 3 is shown when sample size is calculated to reject the null hypothesis for question 3 with power of 0.90 and type I error of 0.05 (two-tailed).
[a]The 95% confidence interval for this proportion does not contain 0.90.

Table 8.7b Investigation of Sample Size Violations for Question 4: Probability[a] 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	*Initial Response Rate (Default)*	*Pattern*[b]	*Sample Size*[c]	*Default Working Assumptions Are Correct*	*Unequal Initial Response Rates*	*Unequal Variance*	*Non-Normal Outcome Y*
0.2	0.5	1	608	0.966[d]	0.984[d]	0.965[d]	0.972[d]
0.2	0.1	1	608	0.962[d]	0.969[d]	0.964[d]	0.962[d]
0.5	0.5	1	97	0.980[d]	0.985[d]	0.966[d]	0.956[d]
0.5	0.1	1	97	0.960[d]	0.919[d]	0.976[d]	0.947[d]
0.2	0.5	2	608	0.964[d]	0.953[d]	0.952[d]	0.944[d]
0.2	0.1	2	608	0.905	0.929[d]	0.922[d]	0.923[d]
0.5	0.5	2	97	0.922[d]	0.974[d]	0.976[d]	0.948[d]
0.5	0.1	2	97	0.893	0.917	0.927[d]	0.885

[a]Probability calculated as the percentage of 1,000 simulations on which correct strategy mean was selected as the maximum.

[b]1 refers to the pattern of strategy means such that all are different but that the mean for ($A_1 = 1$, $A_2 = 1$), that is, strategy A, is always the highest. 2 refers to the pattern of strategy means such that the mean for strategy A is higher than the other three and the other three are all equal.

[c]Calculated to detect the correct maximum strategy mean 90% of the time when the sample size assumptions hold.

[d]The 95% confidence interval for this proportion does not contain 0.90.

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

As discussed, we also assess the power for question 4 when the trial was sized for a different research question. For each of the types of research questions in Table 8–2, we generate a data set that follows the working assumptions for the sample size formula for that question, then evaluate the power of question 4 to detect the optimal strategy. From Table 8–8a–c, we see that in almost all cases, regardless of the starting assumptions used to size the various research questions, 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 question 4 falls below 0.1. These results are not surprising as from Table 8–6 we see that question 4 requires much smaller sample sizes than all the other research questions.

Note that question 4 is more closely linked to question 3 than to question 1 or 2. Question 3 is potentially a subset of question 4; this relationship occurs when one of the strategies considered in question 3 is the strategy with the highest mean outcome. The probability of detecting the correct

Table 8.8a The Probability[a] of Choosing the Correct Strategy for Question 4 When Sample Size Is Calculated to Reject the Null Hypothesis for Question 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 Question 1	*Initial Response Rate*	*Sample Size*	*Question 1 (Power)*	*Question 4 (Probability[a])*	*Effect Size for Question 4*
0.2	0.5	1,056	0.880	1.000	0.325
0.2	0.1	1,056	0.904	1.000	0.425
0.5	0.5	169	0.934	0.987	0.350
0.5	0.1	169	0.920	0.998	0.630

[a]Probability calculated as the percentage of 1,000 simulations on which correct strategy mean was selected as the maximum.

Table 8.8b The Probability[a] of Choosing the Correct Strategy for Question 4 When Sample Size Is Calculated to Reject the Null Hypothesis for Question 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 Question 2	*Initial Response Rate*	*Sample Size*	*Question 2 (Power)*	*Question 4 (Probability[a])*	*Effect Size for Question 4*
0.2	0.5	2,112	0.906	0.999	0.133
0.2	0.1	1,174	0.895	0.716	0.054
0.5	0.5	338	0.895	0.997	0.372
0.5	0.1	188	0.901	0.978	0.420

[a]Probability calculated as the percentage of 1,000 simulations on which correct strategy mean was selected as the maximum.

strategy mean as the maximum when sizing for question 3 is generally very good, as can be seen from Table 8–8c. 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 1,584–2,112. The sample size required for a 0.90 probability of selecting 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

Table 8.8c The Probability[a] of Choosing the Correct Strategy for Question 4 When Sample Size Is Calculated to Reject the Null Hypothesis for Question 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 Question 3	*Initial Response Rate*	*Sample Size Formula*	*Sample Size*	*Question 3 (Power)*	*Question 4 (Probability[a])*	*Effect Size for Question 4*
0.2	0.5	N_{3a}	1,584	0.893	0.939	0.10
0.2	0.1	N_{3a}	2,007	0.882	0.614	0.02
0.5	0.5	N_{3a}	254	0.896	0.976	0.25
0.5	0.1	N_{3a}	321	0.926	0.978	0.32
0.2	0.5	N_{3b}	2,112	0.950	0.953	0.10
0.2	0.1	N_{3b}	2,112	0.903	0.613	0.02
0.5	0.5	N_{3b}	338	0.973	0.989	0.25
0.5	0.1	N_{3b}	338	0.937	0.985	0.32

[a]Probability calculated as the percentage of 1,000 simulations on which correct strategy mean was selected as the maximum.

powered to detect differences between strategy means each beginning with a different initial treatment.

Summary

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

When planning a SMART 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 initial 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.

Discussion

In this chapter, we demonstrated how a SMART 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. Two new methods for calculating the sample sizes for a SMART 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 initial 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.

Several comments are in order regarding the violations of assumptions surrounding the values of the initial response rates when investigating sample size formula N_{3a} for question 3. First, we examined violations of the assumption of the homogeneity of response rates across initial treatments such that they differed by 10% (initial response rates differing by more than 10% in addictions clinical trials are rare) and found that the sample size formula performed well. Future research is needed to examine the question regarding the extent to which initial 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 addictions and in many other areas of mental health, both clinical practice as well as trials are plagued with subject nonadherence to treatment. In these cases sophisticated causal inferential methods are often utilized when trials are ''broken'' in this manner. An alternative to the post hoc, statistical approach to dealing with nonadherence is to consider a proactive experimental design such as SMART. The SMART design provides the means for considering nonadherence as one dimension of nonresponse to treatment. That is, nonadherence is an indication that the treatment must be altered in some way (e.g., by adding a component that is designed to improve motivation to adhere, by switching the treatment). In particular, one might be interested in varying secondary treatments based on both adherence measures and measures of continued drug use.

In this chapter we focused on the simple design in which there are two options for nonresponders and one option for responders. Clearly, these results hold for the mirror design (one option for nonresponders 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 nonresponders or designs in which there are three randomizations. 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 time-varying symptom levels; again, it is important to generalize the results to this setting.

Appendix

Sample Size Formulae for Question 3

Here, we present the derivation of the sample size formulae N_{3a} and N_{3b} for question 3 using results from Murphy (2005).

Suppose we have data from a SMART design modeled after the one presented in Figure 8–2; that is, there are two options for the initial treatment, followed by two treatment options for nonresponders and one treatment option for responders. We use the same notation and assumptions listed in Test Statistics and Sample Size Formulae. Suppose that we are interested in comparing two strategies that have different initial treatments, strategies (a_1, a_2) and (b_1, b_2). Without loss of generality, let $a_1 = 1$ and $b_1 = 0$.

To derive the formulae N_{3a} and N_{3b}, we will make the following working assumption: The sample sizes will be large enough so that $\hat{\mu}_{(a1,\ a2)}$ is approximately normally distributed.

We use three additional assumptions for formula N_{3a}. The first is that the response rates for the initial treatments are equal and the second two assumptions are indicated by * and **.

The marginal variances relevant to the research question are $\sigma_0^2 = \mathrm{Var}[Y|A_1 = a_1,\ A_2 = a_2]$ and $\sigma_1^2 = \mathrm{Var}[Y|A_1 = b_1,\ A_2 = b_2]$. Denote the mean outcome for strategy (A_1, A_2) by $\mu_{(A1,A2)}$. The null hypothesis we are interested in testing is

$$\mathrm{H0} : \mu_{(1,a2)} - \mu_{(1,b2)} = 0$$

and the alternative of interest is

$$\mathrm{H1} : \mu_{(1,a2)} - \mu_{(1,b2)} = \delta\sigma$$

where $\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_0^2}{2}}$. (Note that δ is the standardized effect size.)

As presented in Statistics for Addressing the Different Research Questions, the test statistic for this hypothesis is

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

where $\hat{\mu}_{(a1,a2)}$ and $\hat{\tau}^2_{(a1,a2)}$ are as defined in Table 8–5; in large samples, this test statistic has a standard normal distribution under the null hypothesis

(Murphy, Van Der Laan, Robins, & Conduct Problems Prevention Group, 2001). Recall that 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 α, we solve

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

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

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

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

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

and

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

Define $\tau^2_{(a1,a2)} = \mathrm{Var}[\sqrt{N}\hat{\mu}_{(a1,a2)}]$. Note that

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

is close to 1 in large samples (Murphy, 2005). Now, $\mathrm{E}[\hat{\mu}_{(1,\ a2)} - \hat{\mu}_{(0,\ b2)}] = \mu_{(1,\ a2)} - \mu_{(0,\ b2)}$, so we have

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

Note the distribution of

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

follows a standard normal distribution in large samples (Murphy et al., 2001). Thus, we have

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

Now, using equation 10 in Murphy (2005) for $k=2$ steps1 (initial and secondary) of treatment,

$$\tau^2_{(a1,\ a2)} = E_{a1,a2}\left[\frac{(Y-\mu_{(a1,\ a2)})^2}{Pr(a_1)\ Pr(a_2 \mid R,\ a_1)}\right]$$

$$= E_{a1,a2}\left[\frac{(Y-\mu_{(a1,\ a2)})^2}{Pr(a_1)\ Pr(a_2 \mid 1,\ a_1)}\Big|\ R=1\right]Pr_{a1}[R=1]$$

$$+ E_{a1,a2}\left[\frac{(Y-\mu_{(a1,\ a2)})^2}{Pr(a_1)\ Pr(a_2 \mid 0,\ a_1)}\Big|\ R=0\right]Pr_{a1}[R=0]$$

for all values of a_1, a_2; the subscripts on E and Pr (namely, $E_{a1,a2}$ and Pr_{a1}) indicate expectations and probabilities calculated as if all subjects were assigned a_1 as the initial treatment and then, if nonresponse, assigned treatment a_2. If we are willing to make the assumption (*) that

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

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

$$\tau^2_{(a1,\ a2)} \le E_{a1,a2}[(Y-\mu_{(a1,\ a2)})^2]\frac{Pr_{a1}[R=1]}{Pr(a_1)\ Pr(a_2 \mid 1,\ a_1)}$$

$$+ E_{a1,a2}[(Y-\mu_{(a1,\ a2)})^2]\frac{Pr_{a1}[R=0]}{Pr(a_1)\ Pr(a_2 \mid 0,\ a_1)}.$$

Thus, we have

$$\tau^2_{(a1,\ a2)} \le \sigma^2_{(a1,\ a2)}\left(\frac{Pr_{a1}[R=1]}{Pr(a_1)\ Pr(a_2 \mid 1,\ a_1)} + \frac{Pr_{a1}[R=0]}{Pr(a1)\ Pr(a_2 \mid 0,\ a_1)}\right) \qquad (2)$$

where $\sigma^2_{(a1,\ a2)}$is the marginal variance of the strategy in question.

Since (**) nonresponding subjects ($R=0$) are randomized equally to the two initial treatment options and since there is one treatment option for responders ($R=1$), for a common initial response rate $p=Pr[R=1|A_1=1]=Pr[R=1|A_1=0]$,

$$\tau^2_{(a1,\ a2)} \le \sigma^2_{(a1,\ a2)} * 2 * (2 * (1-p) + 1 * p))$$

Rearranging equation 1 gives us

$$N \approx \left(\frac{\sqrt{\tau^2_{(1,\ a2)} + \ \tau^2_{(0,\ b2)}}}{\delta\sigma}(z_\beta \ + z_{\alpha/2}) \right)^2$$

$$\leq \left(\frac{\sqrt{(\sigma_1^2 + \ \sigma_0^2)(2 * (2 * (1 - \mathrm{p}) + \mathrm{p}))}}{\delta\sqrt{\frac{\sigma_1^2 + \sigma_0^2}{2}}}(z_\beta \ + z_{\alpha/2}) \right)^2$$

Simplifying, we have the formula

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

which is the sample size formula given in Sample Size Calculations that depends on the response rate *p*.

Going through the arguments once again, we see that we do not need either of the two working assumptions (*) or (**) to obtain the conservative sample size formula, $N_{3\mathrm{b}}$:

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

Sample Size Calculation for Question 4

We now present the algorithm for calculating the sample size for question 4. As in the previous section, suppose we have data from a SMART design modeled after the one presented in Figure 8–2; we use the same notation and assumptions listed in Test Statistics and Sample Size Formulae. 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 $\mu_{(\mathrm{A1,A2})}$.

We make the following assumptions:

- The marginal variances of the final outcome given the strategy are all equal, and we denote this variance by σ^2. This means that $\sigma^2 = \mathrm{Var}[Y|A_1 = a_1,\ A_2 = a_2]$ for all (a_1, a_2) in {(1,1), (1,0), (0,1), (0,0)}.
- The sample sizes will be large enough so that $\hat{\mu}_{(\mathrm{a1,\ a2})}$is approximately normally distributed.
- The correlation between the estimated mean outcome for strategy (1, 1) and the estimated mean outcome for strategy (1, 0) is the same as the correlation between the estimated mean outcome for strategy (0, 1) and the estimated mean outcome resulting for strategy (0, 0); we denote this identical correlation by ρ.

The correlation of the treatment strategies is directly related to the initial response rates. The final outcome under two different treatment strategies will be correlated to the extent that they share responders. For example, if the response rate for treatment $A_1 = 1$ is 0, then everyone is a nonresponder and the means calculated for Y given strategy (1, 1) and for Y given strategy (1, 0) will not share any responders to treatment $A_1 = 1$; thus, the correlation between the two strategies will be 0. On the other hand, if the response rate for treatment $A_1 = 1$ is 1, then everyone is a responder to $A_1 = 1$ and, therefore, the mean outcomes for strategy (1, 1) and strategy (1, 0) will be directly related (i.e., completely correlated). Two treatment 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 desired standardized effect size, δ
- the desired probability that the strategy estimated to have the largest mean outcome does in fact have the largest mean, π

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 ρ in {0, 0.01, 0.02, . . . , 0.99, 1} perform the following simulation:
   - Generate $K = 20{,}000$ samples of $[\hat{\mu}_{(1,1)} \quad \hat{\mu}_{(1,0)} \quad \hat{\mu}_{(0,1)} \quad \hat{\mu}_{(0,0)}]^{\mathrm{T}}$ from a multivariate normal with

$$\text{mean M} = \begin{bmatrix} \mu_{(1,1)} \\ \mu_{(1,0)} \\ \mu_{(0,1)} \\ \mu_{(0,0)} \end{bmatrix} = \begin{bmatrix} \delta/2 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and}$$

$$\text{covariance matrix } \Sigma = \frac{1}{N} \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}$$

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

- Count how many times out of $V_1, \ldots, V_{20000}$ 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 ρ in $\{0,\ 0.01,\ 0.02,\ \ldots,\ 0.99,\ 1\}$. Let $\pi_N^* = \min_\rho C_\rho(N)$; the value of π_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 $\pi_N^* = \pi$.

The online calculator for the sample size for question 4 can be found at http://methodology.psu.edu/index.php/smart-sample-size-calculation.

change to http://methodologymedia.psu.edu/smart/samplesize

References

Carroll, K. M. (2005). Recent advances in psychotherapy of addictive disorders. *Cu Psychiatry Reports, 7*, 329–336.

Carroll, K. M., & Onken, L. S. (2005). Behavioral therapies for drug abuse. *American Journal of Psychiatry, 162*(8), 1452–1460.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dawson, R., & Lavori, P. W. (2003). Comparison of designs for adaptive treatment strategies: Baseline vs. adaptive randomization. *Journal of Statistical Planning and Inference, 117*, 365–385.

Fiellin, D. A., Kleber, H., Trumble-Hejduk, J. G., McLellan, A. T., & Kosten, T. R. (2004). Consensus statement on office based treatment of opioid dependence using buprenorphine. *Journal of Substance Abuse Treatment, 27*, 153–159.

Fiellin, D., Pantalon, M., Schottenfeld, R., Gordon, L., & O'Connor, P. (1999). *Manual for standard medical management of opioid dependence with buprenorphine*. New Haven, CT: Yale University School of Medicine, Primary Care Center and Substance Abuse Center, West Haven VA/CT Healthcare System.

Fiellin, D. A., Pantalon, M. V., Chawarski, M. C., Moore, B. A., Sullivan, L. E., O'Connor, P. G., et al. (2006). Counseling plus buprenorphine-naloxone maintenance therapy for opioid dependence. *New England Journal of Medicine, 355*(4), 365–374.

Gandhi, D. H., Jaffe, J. H., McNary, S., Kavanagh, G. J., Hayes, M., & Currens, M. (2003). Short-term outcomes after brief ambulatory opioid detoxification with buprenorphine in young heroin users. *Addiction, 98*, 453–462.

Greenhouse, J., Stangl, D., Kupfer, D., & Prien, R. (1991). Methodological issues in maintenance therapy clinical trials. *Archives of General Psychiatry, 48*(3), 313–318.

Hoel, P. (1984). *Introduction to mathematical statistics* (5th ed.). New York: John Wiley & Sons.

Jennison, C., & Turnbull, B. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall.

Lavori, P.W., & Dawson, R. (2000). A design for testing clinical strategies: Biased adaptive within-subject randomization. *Journal of the Royal Statistical Association, 163*, 29–38.

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

Ling, W., Amass, L., Shoptow, S., Annon, J. J., Hillhouse, M., Babcock, D., et al. (2005). 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, 100*, 1090–1100.

Ling, W., & Smith, D. (2002). Buprenorphine: Blending practice and research. *Journal of Substance Abuse Treatment, 23*, 87–92.

McLellan, A. T. (2002). Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. *Addiction, 97*, 249–252.

McLellan, A. T., Lewis, D. C., O'Brien, C. P., & Kleber, H. D. (2000). Drug dependence, a chronic medical illness. Implications for treatment, insurance, and outcomes evaluation. *Journal of the American Medical Association, 284*(13), 1689–1695.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, 65*, 331–366.

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

Murphy, S. A., Lynch, K. G., Oslin, D.A., McKay, J. R., & Tenhave, T. (2006). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence*. doi:10.1016/j.drugalcdep.2006.09.008.

Murphy, S. A., Oslin, D. W., Rush, A. J., & Zhu, J. (2007). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology, 32*, 257–262.

Murphy, S. A., Van Der Laan, M. J., Robins, J. M., & Conduct Problems Prevention Group (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association, 96*(456), 1410–1423.

Rush, A. J., Crismon, M. L., Kashner, T. M., Toprac, M. G., Carmody, T. J., Trivedi, M. H., et al. (2003). Texas medication algorithm project, phase 3 (TMAP-3): Rationale and study design. *J. Clin. Psychiatry, 64*(4), 357–369.

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D, Canive, J. M., et al. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin, 29*(1), 15–31.

Weiss, R., Sharpe, J. P., & Ling, W. A. (2010). 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. Retrieved June 14, 2020 from http://www.clinicaltrials.gov/ct/show/NCT00316277?order=1