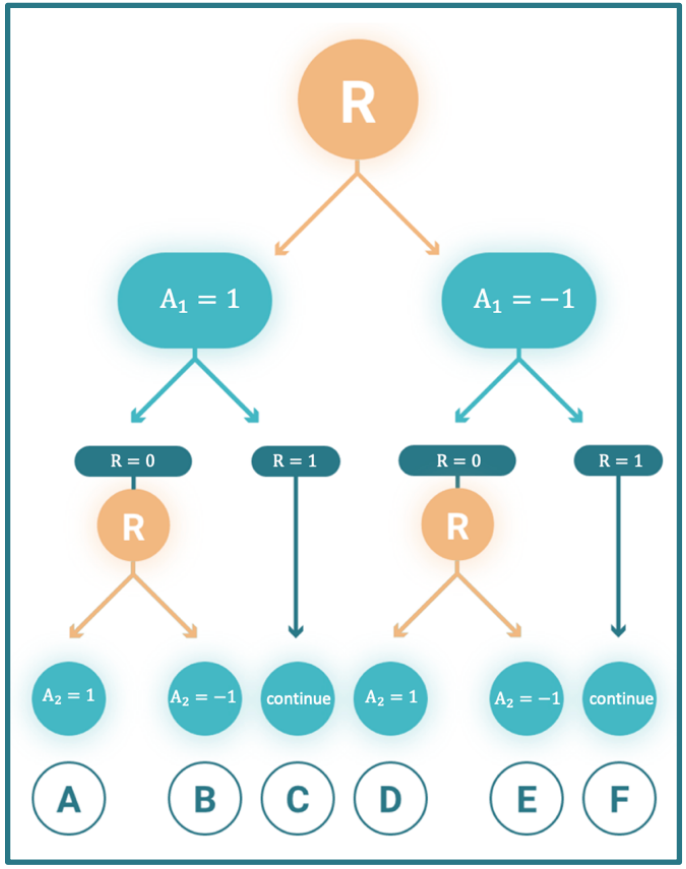
SMART Sample Size Considerations

Prototypical Repeated Measures SMART with a Continuous Primary Outcome

What? The focus of this handout is **prototypical repeated measures (longitudinal) SMARTs with a continuous primary outcome**. This handout provides a table (below) linking 3 common primary research questions in a with the causal effect(s) targeted by it, a description of its associated primary aim hypothesis test and a sample size formula for each question.

Who? This handout is for behavioral intervention scientists who are designing a prototypical **SMART** with a longitudinal (repeated measures) primary outcome and want quick guidance on sample size calculations.



Notation for primary research outcome. This handout concerns a continuous primary research outcome $Y_{it}(a_1, a_2)$ that is measured at times $t = 0, \ldots, T$ (same for all subjects). Specifically, $Y_{it}(a_1, a_2)$ is the primary research outcome that would occur for unit i at time t, had unit i been targeted with the adaptive intervention (a_1, a_2).

Notation for the adaptive interventions. The pair (a_1, a_2) is used to denote one of the four embedded clustered adaptive interventions in a prototypical, repeated measures SMART: $\{(1,1), (-1,1), (1,-1), (-1,-1)\}$.

Notation for the observed data. The observed data in a prototypical SMART includes stage 1 intervention assignment A_{1i}, response status at the end of stage 1 R_i, stage 2 intervention assignment A_{2i}, and the unit level repeated measures outcome Y_{it}.

A prototypical, repeated measures SMART. A prototypical, longitudinal SMART is a two-stage, sequentially randomized trial (see figure to the right), in which

(i) all units are randomized at the beginning of stage 1 to $\mathrm{A}_1 = -1$ vs. $\mathrm{A}_1 = 1$, equal probability ($\Pr(\mathrm{A}_{1\mathrm{i}} = 1) = 1/2$);

(ii) units i that do not respond to $\mathrm{A}_{1\mathrm{i}}$ ($\mathrm{R}_{\mathrm{i}}(\mathrm{A}_{1\mathrm{i}}) = 0$) are re-randomized at the stage 2 decision point to $\mathrm{A}_{2i} = -1$ vs. $\mathrm{A}_{2i} = 1$, with probability ($\Pr(\mathrm{A}_{2\mathrm{i}} = 1 \mid \mathrm{R}_{\mathrm{i}}(\mathrm{A}_{1\mathrm{i}}) = 0) = 1/2$); and

(iii) units i that respond to stage 1 intervention (i.e., $\mathrm{R}_{\mathrm{i}} = 1$) are not re-randomized in stage 2 (e.g., they continue with stage 1 intervention). For responders, A_{2i} is undefined, by design.

Notation used to define causal effects. $Y_{it}(a_1, a_2)$ denotes the primary research outcome that would occur for unit i at time t had unit i been offered the adaptive intervention (a_1, a_2). For every unit i, there are 4 potential values of $Y_{it}(a_1, a_2)$. Note that the observed Y_{it}= $Y_{it}(A_{1i}, A_{2i})$; and note that in a prototypical SMART, if $R_i(A_{1i}) = 1$ then $Y_{it}(A_{1i}, -1) = Y_{it}(A_{1i}, 1)$ for all units that are responding. The unit-level causal effects $\Delta_{it}\ (a_1, a_1', a_2, a_2') = Y_{it}(a_1, a_2) - Y_{it}(a_1', a_2')$ can never be observed, even using a SMART. So the ultimate goal of any prototypical SMART is to generate scientific knowledge about features of Δ_{it} that are useful in constructing a high-quality adaptive intervention. The 3 most common primary research questions in a repeated measures SMART involve marginal mean summaries of Δ_{it}.

Marginal mean modeling. We use a marginal mean model $\mu_t(a_1, a_2; \gamma)$ (with unknown parameters γ) for $E[Y_{it}(a_1, a_2)]$ to express the causal effects that are targeted by each of the primary research questions listed below. There are four means, one for each adaptive intervention (a_1, a_2). For a continuous outcome in a prototypical repeated measures SMART design, we use the following marginal mean model:

$$\mu_t(a_1, a_2; \gamma) = \gamma_0 + I(t \le t^*)(\gamma_1 t + \gamma_2 a_1 t) + I(t > t^*)[\gamma_1 t^* + \gamma_2 t^* a_1 + \gamma_3(t - t^*) + \gamma_4(t - t^*)a_1 + \gamma_5(t - t^*)a_2 + \gamma_6(t - t^*)a_1 a_2]$$

where

- t^* is a knot corresponding to the measurement time immediately before the second-stage randomization (t^* is fixed for all subjects)
- $I(\cdot)$ is the indicator function
- In the sample size table below, we assume t^*=1 and T=2 is fixed for all subjects

Within-unit correlation coefficient ρ. For each (a_1, a_2), let $e_{it}(a_1, a_2; \sigma)$ be a working model for the total error around the marginal mean, that is, for $Y_{it}(a_1, a_2) - E[Y_{it}(a_1, a_2)]$. For the sample size calculators given in this handout, we decompose the total error using a "exchangeable" working covariance

$$e_{it}(a_1, a_2; \sigma) = \eta_i + \epsilon_{it}$$

where $E(\eta_i) = 0$ and $Var(\eta_i) = \sigma^2_{individual}$, $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{it}) = \sigma^2_{residual}$, and $Cov(\eta_i, \epsilon_{it}) = 0$. This working model for the total error leads to a working variance for the total error is given by $Var(e_{ij}) = \sigma^2_{total} = \sigma^2_{ind} + \sigma^2_{residual}$ and the within-unit correlation is given by

$$\rho = \sigma^2_{ind}/\sigma^2_{total} = \frac{\sigma^2_{ind}}{\sigma^2_{ind} + \sigma^2_{residual}}.$$

ρ corresponds to within-person correlation and captures similarities in the multiple measurements take from the same individual.

Minimum significant effect size $\delta = \Delta/\sigma^2_{total}$. δ is defined as the standardized causal effect $\delta = \Delta/\sigma_{total}$ where 3 choices for Δ are given in the table below. In the samples size formulae below, δ will be set to the smallest effect size you would like to detect (i.e., the smallest effect size that is educationally or clinically significant). A smaller δ requires a larger sample size. Typically, values of 0.2, 0.5 and 0.8 are considered small, moderate, and large, respectively.

Additional methodological considerations.

- δ is based on what is the minimum educationally or clinically significant effect size one wants to detect. It is not based solely on pilot data.
- For simplicity, we provide formulae using a working model for the variances σ^2_{total} and ρ that does not differ by the adaptive intervention (a_1, a_2). If variance components are expected to differ by (a_1, a_2), we recommend that you use the largest value. This should lead to more conservative

(greater than needed) estimate of the minimum sample size. If you want sharper bounds, consult Seewald, et al. (2020) or use a simulations experiment.

- $r(a_1) = \Pr(R(a_1) = 1)$ is the probability of response to the first-stage treatment. For simplicity, in table rows #2 and #3, we provide formulae under the working assumption that the response rate is the same across first-stage treatment: $r = r(1) = r(-1)$. If $r(a_1)$ is expected to differ by a_1, we recommend that you (i) use the larger value for question #2, and (ii) use the smaller value for question #3; this should lead to more conservative (greater than needed) estimate of the minimum sample size. If you want sharper bounds, Seewald, et al. (2020)or use a simulations experiment.
- The formulae presented in this handout do not account for expected missing data in the primary research outcome. We recommend that you inflate the total sample size to adjust for the expected rate of missing data.

Additional notation needed to understand the sample size table.

- H_0 is the null hypothesis and H_1 is the alternative hypothesis. All hypothesis tests shown below are two-sided.
- α in the sample size formulae is the two-side type 1 error. Typically, $\alpha = 5\%$ for a primary research question.
- pwr in the sample size formulae is the target power. Typically, $pwr = 80\%$ or higher for a primary aim

Reference.

Seewald NJ, Kidwell KM, Nahum-Shani I, Wu T, McKay JR, Almirall D. Sample size considerations for comparing dynamic treatment regimens in a sequential multiple-assignment randomized trial with a continuous longitudinal outcome. Statistical Methods in Medical Research. 2020;29(7):1891-1912. doi:10.1177/0962280219877520

For a review of correlation structures, including compound-symmetric correlation, see https://online.stat.psu.edu/stat502_fa21/lesson/11/11.3

Primary Research Question (Primary Study Aim)				Sample Size Formula
	Question Type (Comparison)	What is the causal effect targeted by this primary research question?	What is the hypothesis test associated with this primary research question?	
#1	Main effect of first-stage intervention* (A+B+C vs D+E+F)	$\Delta = E[Y(1,A_2)] - E[Y(-1,A_2)] = 2\gamma_2 + 2\gamma_4$ What is the average effect between adaptive interventions that begin with $A_1 = 1$ vs. $A_1 = -1$ on the end-of-study outcome (EOS)?***	$H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$ To test if there is a difference in the mean EOS outcome between adaptive interventions starting with $A_1 = 1$ vs. with $A_1 = -1$.	$N = \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{pwr}\right)^2}{\delta^2} \times (1-\rho^2)$ This calculation is the same as for a 2-arm randomized clinical trial with repeated measures
#2	Main effect of second-stage intervention (tactic), among units that do not respond** (A+D vs B+E)	$\Delta = E[Y(A_1,1)\mid R(A_1)=0] - E[Y(A_1,-1)\mid R(A_1)=0] = 2\gamma_5 \times (\frac{1}{1-r})$ Among units that do not respond to first-stage intervention, what is the average effect between second-stage $A_2 = 1$ vs. $A_2 = -1$ on the end-of-study outcome?***	$H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$ Among units that do not respond to first-stage intervention, to test if there is a difference in the mean EOS outcome between second-stage $A_2 = 1$ vs. $A_2 = -1$.	$N = \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{pwr}\right)^2}{\delta^2} \times (1-\rho^2) \times (\frac{1}{1-r})$ This calculation is the same as for question #1 times an inflation term to account for the non-response rate.
#3	Comparison of two adaptive interventions that begin with different first-stage intervention (e.g., A+C vs. D+F)	$\Delta = E[Y(1,1)] - E[Y(-1,-1)] = 2(\gamma_2 + \gamma_4 + \gamma_5)$ What is the average effect of adaptive intervention (1,1) vs. (-1,-1) on the end-of-study outcome?*** Or, any one of these pair-wise contrasts: $\Delta = E[Y(1,1)] - E[Y(-1,1)] = 2(\beta_1 + \beta_3)$, $\Delta = E[Y(1,-1)] - E[Y(-1,1)] = 2(\beta_1 - \beta_2)$, $\Delta = E[Y(1,-1)] - E[Y(-1,-1)] = 2(\beta_1 - \beta_3)$	$H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$ To test if there is a difference in the mean EOS outcome between adaptive intervention (a_1, a_2) versus adaptive intervention (a_1', a_2') where $a_1 \neq a_1'$.	$N = \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{pwr}\right)^2}{\delta^2} \times (1-\rho^2) \times (2-r)$ This calculation is the same as for question #1 times a design factor to account for the sharing of responders in the other two adaptive interventions.

* This causal effect is a function of the stage 2 randomization probabilities, which are equal to ½ in a prototypical SMART.
** This causal effect is a function of the stage 1 randomization probability, which is equal to ½ in a prototypical SMART.
*** assume t^*=1 and T=2 is fixed for all subjects

Why do we expect the formulae above for the Primary Aim 1 to be conservative?
The sample size formulae in first row of the table looks very similar to the sample size formula for a two-arm randomized controlled trial (RCT). Here we are considering the setting of a longitudinal SMART, so we have more collected more information over time than in an RCT. If we incorporate this extra information into our model (by adjusting for second-stage treatment, for example), we might capture more of the noise in the data, thus increasing the efficiency of our coefficient estimates. Because our efficiency is higher, these sample size formulae tend to be conservative.

What if you want to address your primary using a metric other than comparing end-of-study outcomes? Other metrics include…

1. **Differences in areas under the curve (AUC)**
    a. AUC can be thought of as a comparison of weighted means of the outcomes measured repeatedly during the trial.
2. **Delayed effects**
    a. Delayed effects are defined as a difference in differences: it compares the difference in long term effects vs. short term effects of two AIs.

We do not provide the estimands and hypothesis tests or sample size formulas for using these additional metrics in this handout, but they can be useful alternatives to comparing end-of-study outcomes in some settings.