**Sample Size for Longitudinal Sequentially Randomized**

**Multiple Assignment Trials with Binary Outcomes**

John J. Dziak, Walter Dempsey, Jamie R. T. Yap, and Inbal Nahum-Shani

Maybe Nick Seewald?

Target journal: Maybe *Psychological Methods*???

Table of Contents

[Introduction 3](#_Toc53837442)

[Model, Notation and Simplifying Assumptions 4](#_Toc53837443)

[Deriving Simplified Formulas for End-of-Study Comparisons 6](#_Toc53837444)

[Overview of General Approach 7](#_Toc53837445)

[Covariance of Model Parameters for a Binary Unrestricted SMART 9](#_Toc53837446)

[Covariance of Model Parameters for a Binary Restricted SMART 10](#_Toc53837447)

[Comparing End-of-Study Outcomes 12](#_Toc53837448)

[Other Estimands 12](#_Toc53837449)

[Method for Simulating Data 15](#_Toc53837450)

[Simple Conditional Method 15](#_Toc53837451)

[Limitations of this Method 16](#_Toc53837452)

[Principal Stratification-Based Method 18](#_Toc53837453)

[Simulation on Binary Unrestricted SMART 19](#_Toc53837454)

[Simulation on Binary Restricted SMART 20](#_Toc53837455)

[Simulation on Area under the Probability Curve 20](#_Toc53837456)

[Discussion 21](#_Toc53837457)

[Table 1: Definitions of Principal Strata 22](#_Toc53837458)

[Table 2: Outcomes to be Simulated in Restricted Case 23](#_Toc53837459)

[Table 3: Outcomes to be Simulated in Unrestricted Case 24](#_Toc53837460)

[Table 4: Data-Generating Parameters for Binary Simulations, Restricted SMART 27](#_Toc53837461)

[Table 5: Simulated Power Results for Some Contrasts of End-of-Study Log Odds 28](#_Toc53837462)

[Table 6: Simulated Power for Some Contrasts of Area under the Probability Curve 29](#_Toc53837463)

[Appendix A: Derivation of Covariance Matrix for Unrestricted Case 30](#_Toc53837464)

[Appendix B: Approximate Covariance Matrix for Restricted Case 33](#_Toc53837465)

[Appendix C: Simplified Sample Size for End-of-Study Contrasts, Unrestricted SMART 35](#_Toc53837466)

[Appendix D: Simplified Sample Size for End-of-Study Contrasts, Restricted SMART 37](#_Toc53837467)

[References 39](#_Toc53837468)

[Acknowledgements 42](#_Toc53837469)

# Introduction

Sequential multiple assignment randomized trials (SMART trials; see Almirall, Nahum-Shani, Sherwood, & Murphy, 2014; Murphy, 2005) are designed to compare efficacy among a set of embedded dynamic adaptive interventions. Often, the efficacy of a candidate adaptive intervention has been operationalized simply by comparing on a single outcome at the end of the study for each participant. However, Lu and colleagues (2016) recommended that more information could be obtained by also considering longitudinal measurements of the outcome during the course of the study. Nahum-Shani and colleagues (2019) and Dziak and colleagues (2019) explained how longitudinal data could be used not only to make potentially more precise comparisons of end-of-study outcomes, but also to estimate other quantities of interest, such as area under the curve (AUC), phase-specific slopes, and delayed effects. They explain that the approach can work well not only for numerical but also for binary outcomes. However, they did not provide sample size or power estimation formulas for planning longitudinal SMART studies to make this comparison.

Kidwell and colleagues (2018) propose sample size formulas for non-longitudinal, end-of-study outcomes in SMART analyses with a logistic model for binary outcomes. Seewald and colleagues (2020) propose sample size formulas for longitudinal SMART analyses for end-of-study outcomes on a numerical variable. Yap and colleagues (2020) propose a general framework for simulating nonnormal outcomes in a SMART, but do not provide specific sample size formulas. However, so far, there is no formula for determining required sample size for longitudinal SMART with binary outcomes. This paper will extend ideas from these papers in order to provide formulas and code to make these sample size calculations.

# Model, Notation and Simplifying Assumptions

We assume that the goal is to compare some quantity between the dynamic treatment regimens (DTRs), also known as adaptive interventions, which are embedded in a planned SMART study. Following past work (Kidwell et al., 2019; Seewald et al., 2020), we consider two simple kinds of SMART design, each in the case of two randomization stages with two available treatments each: the Design I or unrestricted SMART, and the Design II or restricted SMART.

In an unrestricted (Design I) SMART, participants are assigned randomly either to or for the first stage. In the second stage, all participants are then assigned randomly to either or . Responder status is assessed, but is not used to determine whether a participant will be re-randomized. The meaning of the and options may either differ or be the same between nonresponders and responders; if it is the same, then the unrestricted SMART is essentially just a two-by-two factorial design in which one of the factors is not implemented until the second stage. Eight embedded dynamic treatment regimens, representing possible adaptive interventions which could be considered as candidates for recommendation, can be described using data from this design. Each of the eight can be characterized by a triplet of numbers (), each +1 or -1, representing the treatment to be given at first stage, to responders at second stage, and to nonresponders at second stage.

We assume that binary outcomes are observed at time points where the final time point is the distal outcome. Let denote the potential outcome at time for person .

The investigator’s goal is assumed to be making inference on some contrast , such as the relative risk at the final time point between DTR and DTR . In the power and sample size calculations in this paper, we follow Kidwell and colleagues (2019) and Seewald and colleagues (2020) in making the conservative assumption of no baseline covariates, because it is difficult to specify their distributions and effects in advance.

We can fit a model to compare either relative risk, or log odds, between regimens. In the relative risk (log link function) case, we assume is a linear function of the columns of a design matrix containing indicators of time and treatment, with regression coefficients . In the log odds (logit link function) case, we instead assume is a linear function of the matrix, with regression coefficients . For simplicity we make the conservative assumption of no baseline covariates in this paper, but these would otherwise be appended to the design matrix. If feasible, a model should be chosen so that the estimand of interest is a linear combination of **.**

Let be the vector of responses for participant and let be the vector of potential outcomes for a participant given regimen . We assume that the weighting and replication method of Lu and co-workers (2016) is used (see also Dziak et al., 2019, for more information on this method in the binary case).. Let is the weight associated with participant and DTR . We estimate model coefficients by solving the estimating equations

where and is the working variance structure, where is a diagonal matrix with th entry set to an estimate of,and is a working correlation structure. For a logistic link function, and so where is the diagonal matrix with diagonal . For a log link function, and so where is the diagonal matrix with diagonal . We choose to be an exchangeable correlation matrix here because this potentially allows for a closed form sample size solution (Liu and Liang, 1997; see also Wang, 2014).

The expression above looks superficially different from the one used by Dziak and colleagues (2019), because we have absorbed the compatibility indicator into the weights following Seewald and colleagues (2020); for the specific form of the weights see Appendix A for unrestricted and appendix D for restricted designs.

# Deriving Simplified Formulas for End-of-Study Comparisons

For most of this paper, we assume three equally spaced measurement times for notational convenience, labeled 1, 2, and 3. We assume the first randomization immediately follows time 1, and the second immediately follows time 2. We assume that for each randomization, there are two possibilities, denoted and , and assigned with equal (50%) probability. We also assume conservatively, like Kidwell and co-authors (2018) and Seewald and co-authors (2020), that there are no baseline covariates. Under these assumptions, longitudinal change can be modeled as a piecewise linear trajectory. The ideas considered in this paper all apply to much more general settings, but the simple setting allows them to be presented in the most straightforward way. In particular, suppose that the study being planned will be a SMART with two treatment options at time 1, and two treatment options at time 2. Suppose that the design is either unrestricted (so that every participant is re-randomized at time 2 regardless of time 1 assignment), or a prototypical restricted design in which some means is used to distinguish responders from non-responders to early treatment, and only nonresponders are re-randomized. These are described as Designs I and II in Kidwell and colleagues (2019) and Seewald and colleagues (2020). In the following section, we provide formulas for the variances of some estimands of interest in various situations.

## Overview of General Approach

For many scalar quantities , the central limit theorem can be used to argue that

has approximately a standard normal distribution for large sample sizes *n*, where is the sampling variance of given sample size . Consider a two-sided test of whether and assume without loss of generality that . The null hypothesis is rejected in the correct direction if the test statistic is greater than the critical value , the quartile of the standard normal distribution. For example, for . The probability of rejecting the null hypothesis is then

Thus, counting only power for detecting an effect in the correct direction, power can be approximated by

where is the standard normal probability density function. The needed sample size to get a power of to reject the null hypothesis is therefore found by setting

Recall that the asymptotic sampling variance of a parameter is inversely proportional to the sample size. Thus, the required sample size to test with power and two-sided level can be rewritten as

where is the raw effect size, and is a quantity such that for a given sample size , is its sampling variance.

The practical difficulty here is finding an expression for . This depends on the model, the design, and the estimand. In the case of longitudinal SMART, it is easier if we assume that is a difference between two dynamic treatment regimens on some linear combination of the weighted and replicated estimating equations model coefficients for each regimen. That is, where are the regression coefficients in the longitudinal model, and and are the estimands of interest for two different embedded adaptive interventions. This means that

where for a given sample size . So

where . The precise values of specify the estimand, such as final outcome (see Dziak et al., 2019, and Nahum-Shani et al., 2019, Seewald et al., 2020).

However, a remaining difficulty is that the formula for is a rather complicated expression involving several matrices and vectors, given in theorem I.2 of the supplemental material of Lu et al (2016). Expressing it in terms of interpretable and easily elicited parameters could be very difficult. However, a simpler form can be derived under specific assumptions, such as those considered in this paper, which is simple enough to use for sample size planning in advance of having the full dataset.

## Covariance of Model Parameters for a Binary Unrestricted SMART

In our example with three equally spaced time points, the vector of marginal means at times 1, 2, and 3 under DTR prescribing treatments () is where is the link function (such as log or logit) and where the design matrix can be expressed using three rows (one for each measurement time) and eleven columns (representing the dimensions of time, treatment and their interaction). Specifically, to construct for the unrestricted SMART, let be 0, 1, and 1 for times 1, 2, and 3; and let be 0, 0, and 1 for times 1, 2, and 3, representing amount of time spent undergoing the Stage 1 and Stage 2 treatments. Then for the th row, the entries are 1, , , , , , , , and . (We assume no or interaction). That is,

for the values of and and expressing embedded adaptive intervention Then it is shown in Appendix A (actually will be shown, I hope) that for purposes of forming this linear combination, the approximate per-subject covariance of the regression coefficients can be treated as

where values and are effect-coded (possible values and ), and is thediagonal matrixwithentries[] where . If the proportion of responders differs between levels of , take to be the average of and . The function of is toaccountfor the fact that for nonresponders, is unknown and treated as zero by convention, and for responders, is unknown and treated as zero by convention, so the actual effect codes are sometimes zero instead of or . (I just got this heuristically - is there a way to actually prove this? Or is it slightly too optimistic or just plain wrong?)

## Covariance of Model Parameters for a Binary Restricted SMART

In a restricted (Design II) SMART, responders are not re-randomized. Because of this, there are only four embedded adaptive interventions () representing treatment which could be given at stage 1 to everyone, and treatment which could be given at stage 2 to nonresponders. In this case, the vector of means for times 1, 2, and 3 under embedded adaptive intervention consisting of choices () is where the design matrix can be expressed using three rows (one for each time point) and seven columns instead of the eleven needed previously. Specifically, for the th row, the entries are 1, , , , , , and , where is 0, 1, and 1 for times 1, 2, and 3; and is 0, 0, and 1 for times 1, 2, and 3 (see Dziak and co-authors, 2019). That is,

for the values of and expressing embedded adaptive intervention Let be the proportion of responders, assumed for now to be the same regardless of treatment.

Then Appendix B shows (actually will show, I hope) that for purposes of forming this linear combination,

where is adiagonal matrixwith entries , 1, , 1, 1,1, 1. As before, is related to the idea of a design effect, attenuating the information available on the second randomization. (I just got this heuristically - is there a way to actually prove this? Or is it slightly too optimistic or just plain wrong? It works pretty well, although not perfectly, in a simulation with logit link.) The intuition is that for the first five regression coefficients, corresponding to the intercept, , , , and , all subjects provide information; but for the last coefficients, corresponding to , and , only nonresponders provide information so the effective sample size is only . An equivalent justification is that even though for responders, instead of the specified by the regimen for nonresponders, a value of zero is used for in the longitudinal regression.

# Comparing End-of-Study Outcomes

When comparing end-of-study outcomes on the binary scale (testing a log odds ratio), it is shown in Appendix C and Appendix D that the expressions above leads to a conservative upper bound sample size similar to those of Kidwell and colleagues (2019) and by Seewald and colleagues (2020), specifically:

where for unrestricted SMARTs and for restricted SMARTs, where , and where . This also has a very intuitive explanation because it is a product of four terms. The first two terms (also found in Kidwell et al., 2019) would be the sample size for testing a log odds in a posttest-only randomized controlled trial with binary outcomes. The third term (also found in Seewald et al., 2020) suggests that with higher within-subject correlation, a smaller sample size is needed. The fourth term (found in Kidwell et al, 2019, and Seewald et al, 2020) represents the additional cost of comparing four embedded regimens instead of two observed treatment groups. (This is actually just conjecture, I haven’t proved it or even tested it yet in a simulation – but I think maybe it holds, and maybe the same would hold for log also? But maybe for both logit and log some higher upper bound is needed because of how variance depends on both regimen and time.)

# Other Estimands

Many other estimands could be powered for, other than end of study outcome. It is easier if the estimand is a linear combination of the regression coefficients but this is not necessary. If it is not, an approximate variance can be estimated, using an application of Taylor approximation sometimes known as Cramèr’s delta method (see, e.g., Ferguson, 1996), even if the estimand of interest is a continuous but nonlinear function of . Specifically,

where is the vector of derivatives of the estimand in each *β*. This is useful for estimands such as the area under the curve (AUC) of probabilities in a model with either log or logistic link, which either way is not a linear function of the coefficients.

The idea of AUC is more complicated with a binary than with a numerical outcome, because it is necessary to decide whether to compare the curves on the probability, odds, or log odds scales. The probability, the odds, and the log odds are all monotonically related to each other. Thus, when considering a single time point, comparisons between regimens on any of these metrics will agree on sign, although not on value. However, when considering multiple time points, these metrics need not agree even on the sign of differences. Consider a pair of regimens being compared, and suppose the event probabilities were known. For regimen 1, the event probabilities at the three time points are 0.50, 0.99, and 0.80. For regimen 2, the event probabilities at the three time points are 0.50, 0.98, and 0.90. On the probability scale, regimen 1 has an AUC of 3.28 and regimen 2 has 3.36, making regimen 2 appear better. On the log odds scale, however, regimen 1 has an AUC of 4.59 and regimen 2 has 4.33, making regimen 1 appear better. This is not because one answer is incorrect, but of two different estimands, and therefore, two different correct answers. Thus, for a binary outcome, to meaningfully compare area under the curve requires specifying a metric. Dziak and colleagues (2019) argued that the area under the curve for probabilities was more interpretable, because when scaled by the length of the time interval it can be interpreted as the average probability over the continuous time period from the beginning to end of the study. Odds and log odds cannot be averaged in this way.

The AUC on the probability scale for a regimen is a linear combination of the time-specific means for that regimen, although not of the regression coefficients, because the means here are not linear combinations of the regression coefficients. Specifically, the AUC is where is the expected response at time for a participant given regimen , and for evenly spaced measurement times and . It is more convenient to interpret the result if the time interval is rescaled to be 1 unit in length instead of 2, so that the area remains on a 0 to 1 metric and can be interpreted as an averaged time-specific probability averaged over continuous time (see Dziak et al., 2019). Thus, we here use and . As mentioned earlier, the probability curve is not a linear combination of the logistic regression coefficients but an asymptotic sampling variance can still be computed.

Specifically, and , so in order to calculate variance estimates for comparing two regimens and , we need to consider the joint distribution of and . Using the multivariate delta method (a form of Taylor linearization), the asymptotic covariance of the log probabilities (in a log scale model) or log odds ratios (in a logistic model) is

where is estimated by the sandwich covariance estimator from the weighted and replicated generalized estimating equations. Thus the asymptotic covariance of the mean vectors is

where is a diagonal matrix whose entries are the derivatives of the link function, hence for the log link and for the logistic link. Then the asymptotic variance of a linear combination of these vectors is

Note that this is a single number, because, for example, is a matrix when there are 3 time points, and so Cov is a 6 vector times a matrix times a vector, hence a vector. In the special case of comparing rescaled AUC’s with 3 equally spaced waves, and In this case, the asymptotic variance is

(see Dziak et al., 2019). The square root of this variance provides a standard error estimate as usual. However, we do not currently have a simpler formula than this for this situation.

# Method for Simulating Data

In order to test whether the proposed sample size formulas work well, it is necessary to have a way of simulating data from longitudinal SMART designs.

## Simple Conditional Method

Dziak and colleagues (2019) used the following approach in the context of logistic longitudinal SMART. They started by assuming that the conditional means for each possible treatment path at each time conditional on had been elicited, and that the assignment probabilities must be chosen, namely and in a restricted SMART, or and in an unrestricted SMART. They then simulated the data following these steps:

* Randomly assign
* Generate based on . Different initial treatments are allowed to have different response rates.
* Assign . This is done only for responders if the dataset is intended to represent a restricted SMART.
* Choose the appropriate vector of conditional meansto use for each subject based on that subject’s values of , , and . These conditional means must be elicited for each combination of the values of , , and , although equivalently the regression coefficients for the conditional model may be specified instead. The elicitation must be done under the assumption that depends neither on nor , and that depends neither on nor , but may have a (non-causal) effect on all three time periods.
* Generate the responses for each subject from the appropriate mean vector and the specified correlation matrix. This can be done using the bindata package in R (Leisch, Weingessel and Hornik, 1998, 2011), which dichotomizes an underlying multivariate normal distribution with covariance and cutpoints specifically selected to give the desired multivariate binary distribution.

## Limitations of this Method

This method has important limitations. First, it requires the difficult task of eliciting from the distribution conditional on response status , and it does not provide direct control over the correlation matrix conditional on . Also, it is not clear whether it actually accounts for the role of the response rate correctly, because it essentially treats as if it were an observed baseline covariate.

Furthermore, the method cannot be used in its current form in cases in which is deterministically related to , which may occur by design. Recall that both variables are binary and both are observed at the same time. In such cases, conditional on , the joint distribution of is degenerate because has zero variance. For convenience, suppose that is an event whose probability the researchers are trying to reduce (e.g., substance misuse) rather than increase (e.g., exercise). Then these zero-conditional-variance scenarios may include:

1. by definition. That is, a person is a responder if and only if that person is an abstainer at time 2.
2. and are correlated, but not deterministically related, so that any combination of values is possible. For example, is adequate attendance at therapy sessions, while is substance misuse. Attendees may be more likely to avoid misuse but are not immune to it.
3. is always 0 if , but is not necessarily 1 if . For example, to count as a responder a participant must *both* abstain *and* *also* meet some other criterion such as being gainfully employed.
4. is always 1 if , but is not necessarily 0 if . For example, to count as a responder a participant must *either* abstain *or* meet some other criterion such as adequate session attendance.

Each of these possibilities probably has to be considered separately, and none of them can be addressed using the simulation method in Dziak and colleagues (2020). We therefore follow a more sophisticated simulation approach introduced by Yap and colleagues (2020).

## Principal Stratification-Based Method

Yap and colleagues (2020) used principal stratification to handle in a different way, by dividing participants into hypothetical subgroups (somewhat analogously to Frangakis, Rubin, An, & MacKenzie, 2007) based on their potential outcomes on given each possible value of . Following Yap and colleagues (2020), we assume four subgroups: of people who would always be responders, people who would respond only if , people who would respond only if , and people who would never respond, This is summarized in Table 1.

More specifically, we assume that for a total simulated sample size , and for and ,

For example, if and then 50% of people are always responders, would respond only to , and are never responders. If , then of the sample are always responders and are never responders.

This gives a straightforward way to do the simulation:

1. Randomly assign each simulated participant to a subgroup.
2. Simulate randomization to choose which vector of potential outcomes must be generated for each participant.
3. Simulate the appropriate vector of correlated potential outcomes on for each participant based on the that participant’s subgroup.

In order to do steps 1 and 2, some probabilities must be specified. First, the assignment probabilities must be chosen, namely and in a restricted SMART, or and in an unrestricted SMART. In the demonstration in this paper we assume they all use equal probabilities (50%). Also, the response probabilities must be elicited.

Several more parameters are needed for step 3. They are the expected values of each quantity listed in Table 2 for restricted and Table 3 for unrestricted SMARTs. We make the simplifying assumption that the expected values do not depend on subgroup membership. For example, for each subgroup on which is generated, has the same value. As usual (e.g., Seewald et al., 2020) we make the consistency assumption that and .

Kidwell and colleagues proposed their formulas in a setting where only the final endpoints were being compared. However, it is reasonable to assume that this formula also applies as a conservative approximation even when longitudinal data is used to fit the model, as long as the final endpoint comparison is still the only estimand of interest. We did a simulation study to test this in both the case of a restricted and unrestricted design. The results are presented in the next three sections.

# Simulation on Binary Unrestricted SMART

(I will write this up)

# Simulation on Binary Restricted SMART

(I will write this up)

# Simulation on Area under the Probability Curve

(I will write this up)

# Discussion

This paper demonstrated the approximate agreement of two proposed methods, one using asymptotic formulas and one using simulations, of sample size computation for comparing longitudinal estimands between embedded adaptive interventions in a SMART.

This paper focused on power for pairwise comparisons between AI’s. Rose and co-workers (2019) worked on another kind of sample size planning for SMART studies: how to choose a large enough sample size to find the optimal among a set of candidate interventions, not merely to do pairwise tests. This might be challenging to extend to longitudinal binary data, although it could probably be handled using intensive simulations, simplifying assumptions, and sensitivity analyses.

# Table 1: Definitions of Principal Strata

|  |  |  |
| --- | --- | --- |
| Subgroup | Response status  if | Response status  if |
| 1. Would always respond |  |  |
| 2. Would only respond to |  |  |
| 3. Would only respond to |  |  |
| 4. Would never respond |  |  |

# Table 2: Outcomes to be Simulated in Restricted Case

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Subgroup | | | |
|  | 1 | 2 | 3 | 4 |
|  | Y | Y | Y | Y |
|  | Y | Y | Y | Y |
|  | Y | Y | Y | Y |
|  | Y | Y | N | N |
|  | Y | N | Y | N |
|  | N | N | Y | Y |
|  | N | N | Y | Y |
|  | N | Y | N | Y |
|  | N | Y | N | Y |

# Table 3: Outcomes to be Simulated in Unrestricted Case

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Subgroup | | | |
|  | 1 | 2 | 3 | 4 |
|  | Y | Y | Y | Y |
|  | Y | Y | Y | Y |
|  | Y | Y | Y | Y |
|  | Y | Y | N | N |
|  | Y | Y | N | N |
|  | Y | N | Y | N |
|  | Y | N | Y | N |
|  | N | N | Y | Y |
|  | N | N | Y | Y |
|  | N | Y | N | Y |
|  | N | Y | N | Y |

**Table 4**

Data-Generating Parameters for Binary Simulations, Unrestricted SMART

|  |  |
| --- | --- |
| Parameter | Value |
|  | 0.7 |
|  | 0.6 |
| Specified conditional correlation | 0.3 |
| Approximate marginal correlation = | 0.43 |
|  | 0.366 |
|  | 0.598 |
|  | 0.380 |
|  | 0.444 |
|  | 0.549 |
|  | 0.698 |
|  | 0.514 |
|  | 0.39 |
|  | 0.616 |
|  | 0.362 |
|  | 0.496 |
|  | 0.577 |
|  | 0.395 |

# Table 4: Data-Generating Parameters for Binary Simulations, Restricted SMART

|  |  |
| --- | --- |
| Sample Size | Restricted |
|  | 0.70 |
|  | 0.60 |
| Conditional correlation | 0.30 |
| Approximate marginal correlation = | 0.40 |
|  | 0.225 |
|  | 0.405 |
|  | 0.358 |
|  | 0.176 |
|  | 0.358 |
|  | 0.288 |
|  | 0.563 |
|  | 0.375 |
|  | 0.188 |
|  | 0.071 |
|  | 0.158 |

# Table 5: Simulated Power Results for Some Contrasts of End-of-Study Log Odds

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Working**  **Independence** | | **Working**  **Exchangeable** | |
| **Contrast** | **Effect** | **Predicted** | | **Observed** | **Predicted** | **Observed** |
|  |  | (approx.) | (sharp) |  |  |  |
| **Unrestricted SMART, *n*=223** | | | | | | |
| vs | .643 vs. .416 | 0.663 | 0.693 | 0.696 | 0.744 | 0.729 |
| vs | .643 vs. .375 | 0.797 | 0.837 | 0.824 | 0.877 | 0.852 |
| vs | .643 vs. .545 | 0.182 | 0.184 | 0.200 | 0.202 | 0.232 |
| vs | .643 vs. .504 | 0.314 | 0.320 | 0.332 | 0.353 | 0.369 |
| **Restricted SMART, *n*=230** | | | | | | |
| vs | .370 vs. .142 | 0.911 | 0.945 | 0.946 | 0.963 | 0.955 |
| vs | .370 vs. .176 | 0.796 | 0.831 | 0.820 | 0.868 | 0.842 |
| vs | .314 vs. .142 | 0.745 | 0.785 | 0.801 | 0.827 | 0.827 |
| vs | .314 vs. .176 | 0.542 | 0.565 | 0.569 | 0.610 | 0.606 |

Note: **Only some contrasts are shown, to save space.** **The sample size for each simulation was arbitrarily set at the number suggested by the formula of Kidwell et al (2020) for the second contrast shown in each scenario.**

# Table 6: Simulated Power for Some Contrasts of Area under the Probability Curve

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Contrast** | | | | | | **Size** | **Predicted** | | **Observed** | **Predicted** | **Observed** |
| **Unrestricted SMART, *n*=223** | | | | | | | | | | | |
| vs | | | | | | .643 vs. .416 |  | 0.429 | 0.440 | 0.415 | 0.479 |
| vs | | | | | | .643 vs. .375 |  | 0.521 | 0.532 | 0.505 | 0.572 |
| vs | | | | | | .643 vs. .545 |  | 0.185 | 0.198 | 0.179 | 0.227 |
| vs | | | | | | .643 vs. .504 |  | 0.250 | 0.262 | 0.242 | 0.296 |
| **Restricted SMART, *n*=230** | | | | | | | | | | | |
| vs | | | | | | .370 vs. .142 |  | 0.962 | 0.923 | 0.952 | 0.937 |
| vs | | | | | | .370 vs. .176 |  | 0.932 | 0.879 | 0.918 | 0.898 |
| vs | | | | | | .314 vs. .142 |  | 0.916 | 0.865 | 0.902 | 0.888 |
| vs | | | | | | .314 vs. .176 |  | 0.865 | 0.798 | 0.845 | 0.825 |
|  | | | | | |
|  |  |  |  |  |  |

Note: **Only some contrasts are shown, to save space.** **The sample size for each simulation was arbitrarily set at the number suggested by the formula of Kidwell et al (2020) for the second contrast shown in each scenario.**

# Appendix A: Derivation of Covariance Matrix for Unrestricted Case

The weights are defined to equal the indicator for the compatibility of the participant’s actual assignment with regimen , divided by the probability of that participant’s assignment, as in Lu and colleagues (2016), Dziak and colleagues (2019), Kidwell and colleagues (2019), and Seewald and colleagues (2020). This is closely similar to the use of inverse propensity weights in observational studies. That is,

where is one if the person’s treatment assignments and response variable are compatible with regimen and is zero otherwise. Participant is “compatible” with a particular adaptive intervention () if the participant’s randomly assigned treatments agreed with what the intervention would have recommended, that is, if and then either and , or else and Assuming equal assignment probabilities and an unrestricted SMART,

where is the indicator function (1 for true and 0 for false).

Seewald et al (2020) showed that under these assumptions, the asymptotic covariance of the coefficient estimates in the longitudinal model is given by the sandwich formula

where in the case with no baseline covariates the bread is

and the meat is

where means that the cross-product of a vector by itself (i.e., ), and where is the vector of potential outcomes for participant under regimen , and . Thus, for an estimand ,

In order for these quantities to be useful for sample size planning, they must be reexpressed in terms of quantities that can reasonably be elicited or hypothesized in advance.

Under the assumption that the marginal variance matrices are bounded above by some ***V***, ***(we hope that)*** it can be shown (see Seewald and colleagues, 2020) that an approximation(? upper bound?) for this quantity is

**,**

That is, that the covariance of the regression coefficients can be simply approximated by

with the diagonal multipliers as given in the text. (I haven’t been able to prove this, if it is true, but I’m not sure how. It seems approximately true in simulations. Seewald et al (2020) doesn’t use this expression – they take a different approach which I don’t understand, and seem to omit a bunch of algebra I needed. Kidwell et al (2019) had only one time point, and also considered only two regimens at a time. One thing they both did was assume that the two regimens differ at least on a1.)

# Appendix B: Approximate Covariance Matrix for Restricted Case

(Supposed to be like Appendix A, but for restricted)

From Seewald et al (2020) we still have that for an estimand , and link function,

where

and th

Not sure where to go from here. Goal if possible is to prove

**,**

That is, that the covariance of the regression coefficients can be simply approximated by

with the diagonal multipliers as given in the text. Goal is to prove this, if it is true. I worked partway through it in the normal linear case but didn’t get all the way, and this case is more complicated because variance of Y depends on mean which depends on estimand and time, so you can’t assume homoskedasticity as Seewald et al (2020) did to get their formula.

# Appendix C: Simplified Sample Size for End-of-Study Contrasts, Unrestricted SMART

(I’ve actually been assuming that at this level of approximation the sample size doesn’t depend on whether you’re testing relative risk or log odds, as long as you start with the elicited probabilities and then use the right link function when calculating the effect size and fitting the model. The intuition is that they are both ways of testing whether the final probabilities are equal or not at the last time stage, and the probabilities are imagined not to be very close to 0 or 1. It was just an assumption though)

Recall that for the three-wave unrestricted SMART example, the columns of the design matrix are the intercept column, , , , , , , , , and . With a logistic link, the end-of-study log odds for an intervention will be given by with. With a log link, the same formula gives the end-of-study odds.

Suppose that interventions and , differing only on , are being compared. Then . Because a quadratic form is being taken, and and are restricted to equal or , the signs of and do not matter and it is safe to take .

If the interventions additionally differ on , then disregarding sign, will be because the products and will be equal even though the individual terms are not. Likewise, if the interventions differ only on and then |=, and if they differ on all three terms then |.

Assume is the exchangeable correlation matrix

Then in the expression from Appendix A,

we can substitute

Not sure where to go from here. Goal if possible is to prove

Where =2 for unrestricted case – or if this doesn’t hold, some similar expression, or else to conclude it’s not possible.

# Appendix D: Simplified Sample Size for End-of-Study Contrasts, Restricted SMART

(Supposed to be like Appendix C, but for restricted)

From Appendix B,

The expected value at the final measurement point for adaptive intervention () is where.

For example, the time-3 expected value for (+1,+1) is . The difference in final expected values between adaptive intervention () and adaptive intervention () is then , where

For example, the time-3 expected value for (+1,+1) is and the time-3 expected value for (-1,+1) is , so the difference is and so we would use .

Not sure where to go from here.Goal if possible is to prove

Where =2-*r* for restricted case. – or if this doesn’t hold, some similar expression, or else to conclude it’s not possible.

**Baby version of Appendix D, for linear model with independence**

Although I wasn’t able to get a simplified formula in the correlated binary case, it is very much easier to get it one in a homoskedastic independent normal case, and I thought maybe that might be helpful as a template for discussing what further assumptions and approximations are needed.

The expected value at the final measurement point for adaptive intervention () is where. For example, the time-3 expected value for (+1,+1) is . The difference in final expected values between adaptive intervention () and adaptive intervention () is then , where

For example, the time-3 expected value for (+1,+1) is and the time-3 expected value for (-1,+1) is , so the difference is and so we would use . For a homoscedastic linear model under independence, assuming that the proof analogous to Appendix B already, we would have the result

Where

Here,

and so, because the regimen-specific terms balance out over the sum,

The inverse turns out to be

Multiplying by , we get

So

Because of the zero weights, this can be simplified to

Letting So the needed sample size is

Which is equal to expression (10) in Seewald and colleagues (2020) when the correlation is zero.

I also made some attempt to do this for the exchangeable case with positive correlation. If we assume exchangeable errors, we have

so

where and . Then after algebra,

so

Because the off-diagonal blocks are all zeros, we can invert the upper and lower diagonal matrices separately. Only the lower one is of interest for our contrast because the multipliers for the first three rows and columns are zeroes.

Therefore, for ,

*(This is unfortunately as far as I got so far)*

# References

Almirall, D., Nahum-Shani, I., Sherwood, N. E., & Murphy, S. A. (2014). Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine, 4(3)*, 260-274. DOI: 10.1007/s13142-014-0265-0.

Dziak, J. J., Yap, J. R. T., Almirall, D., McKay, J. R., Lynch, K. G., & Nahum-Shani, I. (2019) A Data Analysis Method for Using Longitudinal Binary Outcome Data from a SMART to Compare Adaptive Interventions. *Multivariate Behavioral Research*. DOI: 10.1080/00273171.2018.1558042

Ferguson, T. S. (1996). *A course in large sample theory*. New York: Chapman & Hall.

Frangakis, C. E., Rubin, D. B., An, M. W., & MacKenzie, E. (2007). Principal stratification designs to estimate input data missing due to death. *Biometrics, 63,* 641-649.

Kidwell, D. M., Seewald, N. J., Tran, Q., Kasari, C., & Almirall, D. (2018). Design and analysis considerations for comparing dynamic treatment regimens with binary outcomes from sequential multiple assignment randomized trials. *Journal of Applied Statistics, 45*: 1628-1651. DOI: 10.1080/02664763.2017.1386773.

Leisch, F., Weingessel, A. and Hornik, K. (1998). On the generation of correlated artificial binary data. Working Paper Series, SFB “Adaptive Information Systems and Modelling in Economics and Management Science”, Vienna University of Economics.

Leisch, F., Weingessel, A. and Hornik, K. (2015). Package 'bindata': Generation of Artificial Binary Data (version 0.9-19). Obtained at https://cran.r-project.org/web/packages/bindata/bindata.pdf.

Lu, X., Nahum-Shani, I., Kasari, C., Lynch, K., Oslin, D., Pelham, W., . . . Almirall, D. (2016). Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in SMART studies. *Statistics in Medicine, 35,* 1595-1615. DOI: 10.1002/sim.6819.

Lumley, T., Kronmal, R., & Ma, S. (2006). Relative risk regression in medical research: models, contrasts, Estimators, and algorithms (UW Biostatistics Working Paper Series, Working Paper 293). University of Washington. Accessed at <http://biostats.bepress.com/uwbiostat/paper293/>

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine, 24,* 1455-1481. DOI: 10.1002/sim.2022.

Liu, G., & Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. Biometrics, 53: 937-947.

Nahum-Shani, I., Dziak, J. J., & Collins, L. M. (2018). Multilevel factorial designs with experiment-induced clustering. *Psychological Methods, 23*(3), 458–479.

Nahum-Shani, I., Almirall, D., Yap, J. R. T., McKay, J. R., Lynch, K. G., Freiheit, E. A., & Dziak, J. J. (2020). SMART longitudinal analysis: A tutorial for using repeated outcome measures from SMART studies to compare adaptive interventions. Psychological Methods, 25(1), 1–29.

R Core Team (2019). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Accessed at <http://www.R-project.org/>.

Rose, E. J., Laber, E. B., Davidian, M., Tsiatis, A. A., Zhao, Y.-Q. & Kosorok, M. R. (2019). Sample size calculations for SMARTs. arXiv preprint, https://arxiv.org/abs/1906.06646

Seewald, N. J., Kidwell, K.M., Nahum-Shani, I., Wu, T., McKay, J.R., Almirall, D. (2020). Sample Size Considerations for the Analysis of Continuous Repeated-Measures Outcomes in Sequential Multiple-Assignment Randomized Trials. *Statistical Methods in Medical Research*, 2020.

Wang, M. (2014). Generalized estimating equations in longitudinal data analysis: A review and recent developments. *Advances in Statistics*, 2014: 303-728.

Yap, J. R. T., Dziak, J. J., Maiti, R., Lynch, K., McKay, J., Chakraborty, B., & Nahum-Shani, I. (2020). Planning SMARTs: A sample size estimation strategy for comparing dynamic treatment regimens with longitudinal excess zero count outcomes. To be submitted.

# Acknowledgements

This research was supported by awards R01 DA039901 and P50 DA039838 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions as mentioned above. If we don’t list Nick as a co-author, we should acknowledge him here; if we do list him as a co-author then of course we won’t list him here but we should list Danny though.