Published in final edited form as:

Stat Methods Med Res. 2017 August ; 26(4): 1572–1589. doi:10.1177/0962280217708654.

Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: Regression estimation and sample size considerations

Timothy NeCamp^{1,2}, Amy Kilbourne^{3,4}, and Daniel Almirall^{1,2}

¹Department of Statistics, University of Michigan, Ann Arbor, MI, USA

²Survey Research Center, Institute for Social Research, Ann Arbor, MI, USA

³Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI, USA

⁴Quality Enhancement Research Initiative, HSR&D, US Department of Veterans Affairs, Washington, DC, USA

Abstract

Cluster-level dynamic treatment regimens can be used to guide sequential treatment decision-making at the cluster level in order to improve outcomes at the individual or patient-level. In a cluster-level dynamic treatment regimen, the treatment is potentially adapted and re-adapted over time based on changes in the cluster that could be impacted by prior intervention, including aggregate measures of the individuals or patients that compose it. Cluster-randomized sequential multiple assignment randomized trials can be used to answer multiple open questions preventing scientists from developing high-quality cluster-level dynamic treatment regimens. In a cluster-randomized sequential multiple assignment randomized trial, sequential randomizations occur at the cluster level and outcomes are observed at the individual level. This manuscript makes two contributions to the design and analysis of cluster-randomized sequential multiple assignment randomized trials. First, a weighted least squares regression approach is proposed for comparing the mean of a patient-level outcome between the cluster-level dynamic treatment regimens embedded in a sequential multiple assignment randomized trial. The regression approach facilitates the use of baseline covariates which is often critical in the analysis of cluster-level trials. Second, sample size calculators are derived for two common cluster-randomized sequential multiple assignment randomized trial designs for use when the primary aim is a between-dynamic treatment regimen comparison of the mean of a continuous patient-level outcome. The methods are motivated by the Adaptive Implementation of Effective Programs Trial which is, to our knowledge, the first-ever cluster-randomized sequential multiple assignment randomized trial in psychiatry.

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Timothy NeCamp, Department of Statistics, University of Michigan, 311 West Hall, 1085 South University, Ann Arbor, MI 48109, USA. tnecamp@umich.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Keywords

Adaptive interventions; adaptive treatment strategies; dynamic treatment regimens; group-randomized; cluster-randomized; Adaptive Implementation of Effective Programs Trial

1 Introduction

Interventions aimed at improving individual-level outcomes often occur at a cluster level.^{1–3} Often, it may be necessary to use a tailored and dynamic approach to intervention in order to address cluster-level heterogeneity.⁴ For example, due to differences in size, geography, or culture, some clusters may require more intensive or longer-duration intervention in order to improve patient-level outcomes.

Cluster-level dynamic treatment regimens (DTRs), also known as adaptive interventions, can be used to guide such sequential intervention decision-making at the cluster level. In a cluster-level DTR, the cluster-level intervention is potentially adapted (or re-adapted) over time based on changes in the cluster that could be impacted by prior intervention, e.g. adapting based on aggregate measures of the individuals that comprise it. A cluster-level DTR may also include intervention components dynamically tailored to the individuals within clusters.

Sequential multiple assignment randomized trials (SMARTs) represent an important data collection tool for informing the construction of DTRs.^{5–9} The focus of most SMARTs to date has been the development of individual-level DTRs to improve individual-level outcomes (e.g. see Methodology Center¹⁰).

There has been much less focus on analytic or design issues related to cluster-randomized SMARTs for developing cluster-level DTRs. In a cluster-randomized SMART, randomizations occur at the cluster level, yet outcomes are at the level of the individuals within the cluster. Using the Adaptive Implementation of Effective Programs Trial¹¹ (ADEPT) as a motivating example, the focus of this paper is on primary aim analysis and sample size considerations in cluster-randomized SMARTs. ADEPT, which is currently in the field, is to our knowledge the first-ever cluster-randomized SMART. The overarching goal of ADEPT is to develop a cluster-level DTR to improve the adoption of an evidence-based practice (EBP) for mood disorders in community-based mental health clinics and thereby improve patient-level mental health outcomes.

This manuscript makes two contributions to the design and analysis of cluster-randomized SMARTs. First, we develop a regression approach for comparing the mean of a continuous patient-level outcome between the cluster-level DTRs embedded in a SMART. The regression approach is an extension of the estimator by Nahum-Sahni et al.¹² and first introduced by Orellana et al.¹³ The regression approach facilitates the use of individual- and cluster-level baseline (pre-randomization) covariates in the analysis of data from a cluster-randomized SMART.

Second, we develop sample size formulae (for the total number of clusters) to be used when the primary aim of the cluster-randomized SMART is a comparison of the mean of a continuous patient-level outcome between two DTRs beginning with different treatments. This is a common primary aim in SMARTs, see Oetting et al.¹⁴ (continuous end of study outcome) and Li et al.¹⁵ (survival outcome).

This regression approach can be used with any cluster-randomized SMART with repeated cluster-level randomizations. Sample size formulae are developed for two common types of two-stage SMART designs: the one used in ADEPT, in which only one group of non-responders are re-randomized, and the most popular type of SMART, in which all non-responders are re-randomized.

Consistent with the proposed regression approach, which facilitates the use of baseline covariates, the sample size formulae allow scientists to incorporate the correlation between a pre-specified baseline cluster-level covariate and patient-level outcomes, which leads to a reduction in the minimum number of clusters necessary.¹⁶ This manuscript extends the work of Ghosh et al.,¹⁷ which develops sample size calculators for a single type of cluster-randomized SMART in a non-regression context, i.e. without covariates.

2 SMARTs with cluster-level randomization

SMARTs are multi-stage randomized trial designs used explicitly for the purpose of building high-quality DTRs.^{5,18} The multiple stages at which randomizations occur correspond to critical intervention decision points. At each decision point, randomization is used to address a question concerning the dosage (duration, frequency or amount), intensity, type, or delivery of treatment.

Here we consider SMARTs for developing cluster-level DTRs where the unit of randomization (and re-randomization) is a cluster and the outcomes are measured at the level of the individual.

2.1 Motivating example: The ADEPT SMART study

A schematic for the ADEPT trial¹¹ is displayed in Figure 1. The overall aim of ADEPT is to develop a cluster-level DTR to improve the adoption of an EBP for mood disorders in community-based mental health clinics across Colorado and Michigan. The patient-level EBP is known as Life Goals,¹⁹ a collaborative care, psychosocial intervention for mood disorders delivered to patients in six individual or group sessions. The primary outcome in ADEPT is a continuous, patient-level measure of mental health quality of life (MH-QOL).

ADEPT includes several interventions: the replicating effectiveness program (REP), REP plus External Facilitation (REP + EF), and REP plus External and Internal Facilitation (REP + EF + IF). REP is a cluster-level intervention focused on standardizing the implementation of the EBP into routine care through toolkit development, provider training, and program assistance. Facilitation is a cluster-level coaching intervention to help support the use of EBPs. EF is by phone and focuses on technical aspects of how to adopt the EBP; IF is in-person and involves working with a clinic manager to further embed the EBP.

ADEPT, which is currently in the field, involves community-based mental health clinics (approximately $N = 60$) that have failed to respond to an initial six months of REP (pre-randomization). During these six months, each clinic $i = 1, \dots, N$ is expected to identify approximately $m_i = 10$ to 25 patients with mood disorders, all of which are followed for patient-level outcomes throughout the study. Clinics that enter the study (i.e. did not respond to REP at month 6) are randomized with equal probability to receive additional REP + EF or REP EF + IF. After another 6 months, (i) REP + EF sites that are still non-responsive are randomized with equal probability to either continue REP + EF or augment with IF (REP + EF + IF) for an additional 12 months, (ii) REP + EF + IF sites that are still non-responsive continue REP + EF + IF, and (iii) facilitation interventions are discontinued for all sites that are responsive. A clinic is identified as “not responding” at months 6 and 12 if $< 50\%$ of the patients identified to be part of Life Goals during months 0–6 have received ≥ 3 Life Goals sessions.

By design, ADEPT has three DTRs embedded within it, which are displayed in Table 1. Each embedded DTR is labeled (a_1, a_2) . For example, DTR $(1, -1)$, offers REP + EF at month 6, then REP + EF is augmented with IF for clinics that remain non-responsive at month 12, whereas EF is discontinued for clinics who are responsive at month 12.

2.2 The prototypical SMART design

In ADEPT, only clinics not responding to REP + EF were re-randomized at the next stage. This type of SMART (but with individual-level randomizations) has been previously employed in autism research, see Kasari et al.²⁰ and Almirall et al.²¹

Many other types of SMART designs are possible (see Methodology Center¹⁰ for a comprehensive list with individual-level randomizations), including SMARTs where all units are subsequently re-randomized to the same set of next-stage intervention options (e.g. Chronis-Tuscano et al.²²) and others where all units are re-randomized, but to different next-stage intervention options depending on response/non-response to first-stage intervention (e.g. Lu et al.²³). Ultimately, the decision to choose a particular type of SMART is driven by scientific considerations.

By far the most common type of SMART is a two-stage design where (i) all units are randomized to two first-stage treatment options, (ii) a subset of units at the end of stage 1 (e.g. non-responders) are re-randomized to second-stage intervention options regardless of choice of first-stage intervention, and (iii) the remaining subset of units (e.g. responders) are not re-randomized. See Figure 2 for a generic example. We call this a “prototypical SMART design” given its popularity. Note that in the case of the prototypical SMART, there are four embedded DTRs, see Table 2.

Published examples of the prototypical SMART (with individual-level randomizations) include Pelham et al.²⁴ in attention-deficit/hyperactivity disorder, Gunlicks-Stoessel et al.²⁵ in adolescent depression, August et al.²⁶ in conduct disorder prevention, Sherwood et al.²⁷ and Naar-King et al.²⁸ in weight loss, and McKay et al.²⁹ in cocaine/alcohol use.

2.3 Common primary aims in a SMART

This manuscript develops methods for comparing the mean of a continuous individual-level outcome between the DTRs embedded in a cluster-randomized SMART. This comparison can be conceptualized in various ways as a primary aim.^{14,30} (i) To compare first stage intervention options (averaging over the second stage intervention). In ADEPT, this is a comparison of DTR $(-1,.)$ and the DTRs $\{(1,1), (1, -1)\}$ (this was the primary aim in ADEPT, see Kilbourne et al.¹¹). (ii) To compare second stage intervention options (averaging over the first stage intervention). For example, in the prototypical design, this would be a comparison of DTRs $\{(1,1), (-1,1)\}$ and DTRs $\{(1, -1), (-1, -1)\}$ (e.g. see aim 3 in Pelham et al.²⁴). (iii) To compare the mean outcome between two DTRs beginning with the same first-stage treatment. In ADEPT, this is a comparison of DTR $(1,1)$ and $(1,-1)$. (iv) To compare the mean outcome between two DTRs that begin with different first stage treatments. In ADEPT, this is a comparison of $(1,1)$ and $(-1,.)$ or of $(1, -1)$ and $(-1,.)$.

The next section develops a regression estimator that can be used to address all of these primary aims using data from a cluster-randomized SMART. Following that, we derive sample size formulae for aim (iv). Simple extensions of standard sample size formulae may be used for primary aims (i), (ii), and (iii).

3 Methodology

3.1 Marginal mean model

For each SMART participant $j = 1, \dots, m_i$ within each site $i = 1, \dots, N$, we envision a primary end-of-study individual-level outcome Y_{ij} . Let the $p \times 1$ vector $\mathbf{X}_{ij}$ denote a pre-specified set of baseline covariates measured prior to the initial randomization. The baseline covariates, $\mathbf{X}_{ij}$, may be patient-level (e.g. age) or cluster-level (e.g. clinic location).

Denote $E_{a_1, a_2}(Y_{ij}|\mathbf{X}_{ij})$ as the marginal mean of Y_{ij} had the entire population been assigned to the DTR (a_1, a_2) , conditional on baseline covariates, $\mathbf{X}_{ij}$.^{31,32} The mean, $E_{a_1, a_2}(Y_{ij}|\mathbf{X}_{ij})$, averages over the response/non-response measure used the DTR (a_1, a_2) .

Let $\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})$ denote a marginal structural model^{13,33–36} for the mean $E_{a_1, a_2}(Y_{ij}|\mathbf{X}_{ij})$, which is linear in the unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\eta})$. We provide examples below. We denote the causal effects between the DTRs by the $q \times 1$ vector $\boldsymbol{\beta}$, and denote associational effects between $\mathbf{X}_{ij}$ and Y_{ij} by the $p \times 1$ vector $\boldsymbol{\eta}$.

3.1.1 Example 1: ADEPT—An example marginal mean model for the ADEPT study is

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 + \beta_1 a_1 + \beta_2 a_2 I_{a_1=1} + \boldsymbol{\eta}^T \mathbf{X}_{ij} \quad (1)$$

where $I_{a_1=1}$ is an indicator function which equals 1 when $a_1 = 1$.

Here we use a vector $\boldsymbol{\beta}$ with $q = 3$ to capture the causal effects for the three embedded DTRs. The covariates, $\mathbf{X}_{ij}$, could include, for example, the three baseline site-level variables used to stratify the initial randomization: US state (Colorado or Michigan), whether the site

was a primary care or mental health site, and a site-average of individual MH-QOL scores. Using this model to address a primary aim of type (iv) above, the difference between the mean outcome had all clusters received DTR (1,1) and the mean outcome had all clusters received DTR (−1,.)—i.e. $E_{1,1}(Y_{ij}) - E_{-1,1}(Y_{ij})$ —is given by $2\beta_1 + \beta_2$.

3.1.2 Example 2: Prototypical SMART—In the prototypical SMART, we use a vector β with $q = 4$ to capture the causal effects for the 4 embedded DTRs.

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \beta, \eta) = \beta_0 + \beta_1 a_1 + \beta_2 a_2 + \beta_3 a_1 a_2 + \eta^T \mathbf{X}_{ij} \quad (2)$$

Here, the difference between the mean outcome had all clusters received DTR (1,1) and the mean outcome had all clusters received DTR (−1, −1)—i.e. $E_{1,1}(Y_{ij}) - E_{-1,-1}(Y_{ij})$ —is given by $2(\beta_1 + \beta_2)$.

3.2 Estimation

We now present an estimator for the unknown coefficients (β, η) .

3.2.1 Notation—Let $\mathbf{X}_i$ denote the $m_i \times p$ matrix $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \dots, \mathbf{X}_{im_i})^T$ of covariates and let $\mu(\mathbf{X}_i, a_1, a_2; \beta, \eta)$ be the $m_i \times 1$ vector of means

$(\mu(\mathbf{X}_{i1}, a_1, a_2; \beta, \eta), \dots, \mu(\mathbf{X}_{im_i}, a_1, a_2; \beta, \eta))^T$. Let $\mathbf{Y}_i$ be the $m_i \times 1$ vector of responses $(Y_{i1}, Y_{i2}, \dots, Y_{im_i})^T$. Let A_{1i} denote the observed (i.e. randomly assigned) stage 1 treatment. In ADEPT, $A_{1i} = 1$ implies that cluster i received REP + EF as an initial treatment while $A_{1i} = -1$ implies cluster i received REP + EF + IF. Let R_i , a binary variable, denote responder/non-responder status at the end of stage 1. In ADEPT, $R_i = 1$ if cluster i is a responder at the end of the first stage and $R_i = 0$ if cluster i is a non-responder. Let A_{2i} denote the observed (i.e. randomly assigned) stage 2 treatment. Note that, depending on the SMART design, A_{2i} may not be defined for some clusters i depending on the value of (A_{1i}, R_i) . In ADEPT, A_{2i} is defined only for clusters with $A_{1i} = 1$ and $R_i = 0$. In the prototypical SMART, A_{2i} is not defined for clusters with $R_i = 1$. See Tables 1 and 2.

3.2.2 Estimator—Building on Nahum-Shani et al.,¹² Orellana et al.,¹³ and Lu et al.,²³ we obtain estimates of the coefficients (β, η) through solving an estimating equation.

In the estimator, the $m_i \times (q + p)$ matrix $\mathbf{D}(\mathbf{X}_i, a_1, a_2)$ is the derivative of $\mu(\mathbf{X}_i, a_1, a_2; \beta, \eta)$ with respect to (β, η) ; it can be thought of as the “design matrix” for DTR (a_1, a_2) . For example, using the model in equation (1) for ADEPT, the j th row of $\mathbf{D}(\mathbf{X}_i, a_1, a_2)$ is $(1, a_1, a_2 I_{a_1=1}, \mathbf{X}_{ij})$.

The $m_i \times m_i$ matrix $\mathbf{V}(a_1, a_2, \mathbf{X}_i)$ (abbreviated $\mathbf{V}_{i,a_1,a_2}$) is a working model for the covariance of $\mathbf{Y}_i$ conditional on $\mathbf{X}_i$ for DTR (a_1, a_2) , $\text{Cov}_{a_1,a_2}(\mathbf{Y}_i | \mathbf{X}_i)$. In practice, the matrix $\mathbf{V}_{i,a_1,a_2}$ is unknown and must be estimated prior to solving equation (3), see section 3.3.

The function $I(A_{1i}, R_i, A_{2i}, a_1, a_2)$ (abbreviated I_{i,a_1,a_2}) is a cluster-level indicator function which identifies whether (equals 1) or not (equals 0) cluster i was assigned to a sequence of

treatments that is consistent with DTR (a_1, a_2) . For example, in ADEPT, if $A_{1i} = 1$, $R_i = 0$, and $A_{2i} = -1$, then cluster i is consistent only with DTR $(1, -1)$, whereas if $A_{1i} = 1$, $R_i = 1$, then cluster i is consistent with both DTRs $(1, 1)$ and $(1, -1)$.

The weights $W(A_{1i}, A_{2i}, R_i)$, (abbreviated W_i) are the known cluster-level inverse probability weights (IPW),¹³ $W_i = 1 / [f_{A_1}(A_{1i}) f_{A_2|A_1, R}(A_{2i} | A_{1i}, R_i)]$, where $f_{A_1}(a) = Pr(A_1 = a)$ and $f_{A_2|A_1, R}(a | b, c) = Pr(A_2 = a | A_1 = b, R = c)$ are probability mass functions. See Tables 1 and 2 for the known values of W_i in ADEPT and the prototypical SMART.

We obtain estimates $(\hat{\beta}, \hat{\eta})$ by solving for (β, η) in

$$0 = \sum_{i=1}^N U_i(A_{1i}, R_i, A_{2i}, \mathbf{X}_i, \mathbf{Y}_i; \beta, \eta) \triangleq \sum_{i=1}^N \sum_{(a_1, a_2)} I_{i, a_1, a_2} W_i \mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \mathbf{V}_{i, a_1, a_2}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}(\mathbf{X}_i, a_1, a_2; \beta, \eta)) \quad (3)$$

The estimates $(\hat{\beta}, \hat{\eta})$ derived from solving equation (3) are consistent and asymptotically normally distributed assuming the mean model (e.g. equation (1) for ADEPT) is correctly specified. As in the generalized estimating equations literature,^{37,38} there is no requirement that the working model $\mathbf{V}_{i, a_1, a_2}$ be a correct model for $\text{Cov}_{a_1, a_2}(\mathbf{Y}_i | \mathbf{X}_i)$. See supplementary material for a sketch of the derivations.

3.2.3 Intuition for the weights—By design, in the observed data in a SMART, different clusters have different probabilities of being consistent with a specific DTR. For example, clusters assigned to cells A and B are consistent with DTR $(1, 1)$ (see Figure 1 and Table 1). However, clusters assigned to cell A had a 50% chance of being consistent with DTR $(1, 1)$, whereas clusters assigned to cell B had 25% chance of being consistent with DTR $(1, 1)$. Ignoring this known imbalance—i.e. using an unweighted average of observations in cells A and B to estimate the mean outcome had the entire population of clusters been assigned to DTR $(1, 1)$ —would cause the Cell A observations to have an unfairly larger influence on the estimator, leading to bias. The weights are designed to counteract this known imbalance and ensure that all clusters consistent with DTR (a_1, a_2) are represented equally. For example, in ADEPT, clusters in cell A are weighted by $1/0.5 = 2$, whereas clusters in cell B are weighted by $1/0.25 = 4$.

3.3 Implementation

Typically, in clustered settings, our working model for $\text{Cov}_{a_1, a_2}(\mathbf{Y}_i | \mathbf{X}_i)$, $\mathbf{V}_{i, a_1, a_2}$, is taken to be exchangeable and independent of $\mathbf{X}_i$, i.e. $\mathbf{V}_{i, a_1, a_2} = \sigma_{a_1, a_2}^{2*} \cdot \text{Exch}_{m_i}(\rho_{a_1, a_2}^*)$. Here σ_{a_1, a_2}^{2*} and ρ_{a_1, a_2}^* are scalars representing the conditional (on $\mathbf{X}_i$) variance and intra-cluster correlation (ICC) of the outcome under DTR (a_1, a_2) , and $\text{Exch}_{m_i}(\rho_{a_1, a_2}^*)$ is an $m_i \times m_i$

exchangeable matrix (i.e. $[\mathbf{Exch}(\rho)]_{ii} = 1$ and $[\mathbf{Exch}(\rho)]_{ij} = \rho$ for $i \neq j$). Given this working model, the estimators $(\hat{\beta}, \hat{\eta})$ can be obtained using the following steps:

Step 1 Solve equation (3) with $\mathbf{V}_{i,a_1,a_2}$ set to the identity matrix to obtain $(\hat{\beta}_0, \hat{\eta}_0)$. For each embedded DTR (a_1, a_2) obtain the residuals

$$\hat{\varepsilon}_{ij,(a_1,a_2)}(\hat{\beta}_0, \hat{\eta}_0) = Y_{ij} - \hat{\mu}(X_{ij}, a_1, a_2; \hat{\beta}_0, \hat{\eta}_0)$$

Step 2 Estimate σ_{a_1,a_2}^{2*} and ρ_{a_1,a_2}^* using

$$\hat{\sigma}_{a_1,a_2}^{2*} = \frac{\sum_{i=1}^N [W_i I_{i,a_1,a_2} \sum_{j=1}^{m_i} \hat{\varepsilon}_{ji,(a_1,a_2)}^2]}{\sum_{i=1}^N W_i I_{i,a_1,a_2} m_i} \quad \text{and} \quad \hat{\rho}_{a_1,a_2}^* = \frac{\sum_{i=1}^N [W_i I_{i,a_1,a_2} \sum_{j=1}^{m_i} \sum_{k \neq j}^{m_i} \hat{\varepsilon}_{ji,(a_1,a_2)} \hat{\varepsilon}_{jk,(a_1,a_2)}]}{\hat{\sigma}_{a_1,a_2}^{2*} \sum_{i=1}^N W_i I_{i,a_1,a_2} m_i (m_i - 1)} \quad (4)$$

Step 3 Solve equation (3) with $\mathbf{V}_{i,a_1,a_2}$ set to $\hat{\mathbf{V}}_{i,a_1,a_2} = \hat{\sigma}_{a_1,a_2}^{2*} \cdot \mathbf{Exch}_{m_i}(\hat{\rho}_{a_1,a_2}^*)$ to obtain $(\hat{\beta}_1, \hat{\eta}_1)$.

Step 4 Repeat Steps 2 and 3 with $\hat{\varepsilon}_{ij,(a_1,a_2)}(\hat{\beta}_1, \hat{\eta}_1)$ to obtain final estimates $(\hat{\beta}, \hat{\eta})$.

In simulations, we do not find appreciable performance gains by iterating Steps 2 and 3 more than twice. Steps 1–4 can be seen as extensions of standard GEE analysis.^{37,38} Also, some analysts may choose to specify a working correlation structure which is equal for all DTRs. In this case, one could take a simple average of the estimates in equation (4) across all regimens (a_1, a_2) . Lastly, it is well known that by replacing the known W_i in each step above with estimated weights, statistical efficiency of the estimators may be improved.^{39–44}

3.4 Standard error estimation

To estimate the variance of $(\hat{\beta}, \hat{\eta})$, we use the plug-in estimator, given by the $(q + p) \times (q + p)$ matrix $1/N \cdot \sum_{\hat{\beta}, \hat{\eta}} = 1/N \cdot \hat{\mathbf{J}}^{-1} \hat{\mathbf{K}} \hat{\mathbf{J}}^{-1}$ where

$$\hat{\mathbf{J}} = \frac{1}{N} \sum_{i=1}^N \sum_{(a_1,a_2)} I_{i,a_1,a_2} W_i \mathbf{D}(\mathbf{X}_i, a_1, a_2)^T \hat{\mathbf{V}}_{i,a_1,a_2}^{-1} \mathbf{D}(\mathbf{X}_i, a_1, a_2)$$

$$\hat{\mathbf{K}} = \frac{1}{N} \sum_{i=1}^N \mathbf{U}_i(A_{1i}, R_i, A_{2i}, \mathbf{X}_i, \mathbf{Y}_i; \hat{\beta}, \hat{\eta}) \mathbf{U}_i^T(A_{1i}, R_i, A_{2i}, \mathbf{X}_i, \mathbf{Y}_i; \hat{\beta}, \hat{\eta})$$

See supplementary materials for an adjustment to the standard errors for the case when weights are estimated.

3.5 Hypothesis testing

For any linear combination of $(\boldsymbol{\beta}, \boldsymbol{\eta})$ say $\mathbf{c}^T(\boldsymbol{\beta}, \boldsymbol{\eta})$ where $\mathbf{c}$ is a $(q + q)$ -dimensional column vector, we use the univariate Wald statistic $Z = \sqrt{N} \mathbf{c}^T(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) / \sqrt{\mathbf{c}^T \hat{\Sigma}_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}} \mathbf{c}}$ to test the null hypothesis $H_0 : \mathbf{c}^T(\boldsymbol{\beta}, \boldsymbol{\eta}) = 0$. For example, in ADEPT, to test the difference in means had the entire population of clusters followed DTR (1,1) versus DTR (-1,.) (i.e. primary aim (iv) above) using the model in equation (1), we set $\mathbf{c} = (0, 2, 1, \mathbf{0}_p)^T$. In large samples, Z has a standard normal distribution under the null hypothesis. Hence, an α level test is, “reject H_0 when $|Z| > z_{\alpha/2}$,” where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal distribution.

4 Sample size formulae

For both ADEPT and the prototypical SMART, we develop sample size formulae for the total number of clusters N for comparing the mean patient-level outcome between two embedded DTRs beginning with different stage 1 treatments. Specifically, for ADEPT, formulae are developed for testing null hypotheses of the form $H_0 : E_{1,b_2}(Y_{ij}) - E_{-1,.}(Y_{ij}) = 0$ for a fixed $b_2 \in (-1, 1)$ against alternate hypotheses of the form

$H_1 : E_{1,b_2}(Y_{ij}) - E_{-1,.}(Y_{ij}) = \delta \sqrt{(\sigma_{1,b_2}^2 + \sigma_{-1,.}^2)/2}$. Here, δ is a standardized effect size⁴⁵ and σ_{b_1,b_2}^2 is the outcome's marginal variance under DTR (b_1, b_2) . For the prototypical SMART, formulae are developed for testing null hypotheses of the form

$H_0 : E_{1,b_2}(Y_{ij}) - E_{-1,c_2}(Y_{ij}) = 0$ for a fixed $(b_2, c_2) \in \{-1, 1\}^2$ against alternate hypotheses

of the form $H_1 : E_{1,b_2}(Y_{ij}) - E_{-1,c_2}(Y_{ij}) = \delta \sqrt{(\sigma_{1,b_2}^2 + \sigma_{-1,c_2}^2)/2}$. The formulae are based on using equation (3) to estimate the coefficients $\boldsymbol{\beta}$ in marginal models of the form (1) or (2) as follows: (i) with or without a pre-specified cluster-level covariate X_i ; (ii) known weights W_i and (iii) an exchangeable working covariance structure for $\mathbf{V}_{i,a_1,a_2}$. In addition, formulae are based on large sample approximations and a constant cluster size $m_i = m$ for all i . Extensions to the unequal cluster size case can be done as in Kerry and Bland⁴⁶ or by conservatively setting m equal to the minimum cluster size. They also rely on the following *working population assumptions*.

1. **Equal exchangeable covariance matrices across regimens:** We assume the true marginal covariance matrices are equal for the two DTRs we are testing (e.g. $\text{Cov}_{1,b_2}(\mathbf{Y}_i) = \text{Cov}_{-1,c_2}(\mathbf{Y}_i) = \sigma^2 \cdot \text{Exch}(\rho)$ in the prototypical SMART).
2. **Conditional covariance inequality:** For a specific DTR, we assume non-responders do not vary from the marginal mean significantly more than responders. This assumption applies to different DTRs based on design, see below. A concern about this assumption should be raised only if the scientist, apriori, believed that, for a specific DTR, non-responders had significantly larger variances than responders or if the response rate was expected to be much larger than 0.5 (which is atypical for SMART designs). See Appendix 1 for details.
3. **Correct marginal mean model:** We assume that $E_{a_1,a_2}(Y_{ij} | X_i) = \mu(X_i, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta})$ for the pre-specified cluster-level X_i where

$\mu(X_i, a_1, a_2; \beta, \eta)$ is of the form (1) or (2). When X_i is not included in equations (1) or (2), this assumption is met trivially.

Each formula is a function of the cluster size m , the effect size δ , the outcome's ICC, ρ , the probability of a cluster responding after receiving a particular initial treatment,

$p_1 = \mathbb{P}(R=1|A_1=1)$ and $p_{-1} = \mathbb{P}(R=1|A_1=-1)$, and the standard normal quantiles $z_{\alpha/2}$ and z_β , where α is the size of our test and $1 - \beta$ is the power. We first provide formulae for estimation without covariates followed by the case when a cluster-level covariate X_i is used.

4.1 ADEPT sample size formula

For ADEPT, working assumption 2 is:

$$E_{1,b_2}[(Y_i - \mu(1, b_2) | (Y_i - \mu(1, b_2))^T | R=0] \leq E_{1,b_2}[(Y_i - \mu(1, b_2) | (Y_i - \mu(1, b_2))^T] = \text{Cov}_{1,b_2}(Y_i |$$

. Also, for ADEPT, working assumption 1 can be relaxed to $\sigma_{1,b_2}^2 \leq \sigma_{-1,\cdot}^2$. Under these assumptions, we obtain the sample size formula

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m-1)|\rho| \cdot \left(1 + \frac{1-p_1}{2}\right)) \quad (5)$$

4.2 Prototypical sample size formula

For prototypical SMART designs, working assumption 2 is: for both DTRs in our test, i.e.

$(a_1, a_2) = (1, b_2)$ and $(-1, c_2)$,

$$E_{a_1,a_2}[(Y_i - \mu(a_1, a_2) | (Y_i - \mu(a_1, a_2))^T | R=0] \leq E_{a_1,a_2}[(Y_i - \mu(a_1, a_2) | (Y_i - \mu(a_1, a_2))^T] = \text{Cov}_{a_1,a_2}(Y_i |$$

. Under these assumptions, we obtain the sample size formula

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m-1)|\rho| \cdot \left(1 + \frac{(1-p_1) + (1-p_{-1})}{2}\right)) \quad (6)$$

Note this formula is identical to the formula in Ghosh et al.¹⁷ Also, note that we believe working assumptions 1–3 are implicit in their work.

The sample size formulae in formulas (5) and (6) are intuitive. The first two terms in both formulae are identical; these terms compose the formula for the sample size for the difference in means in a 2-arm randomized control trial (RCT) with cluster-level randomization.³ The second term, in particular, is the expression for the variance inflation factor (VIF) arising from cluster-randomized trials. If $\rho = 0$ (i.e. VIF = 1), there is no inflation due to cluster randomization because we have no correlation within clusters. As ρ increases, each new observation within a cluster provides less unique information causing the VIF to increase. This, in turn, leads to an increase in sample size, N .

The third term, which is unique to SMARTs, is used to account for the fact that some clusters are re-randomized depending on response at the end of stage 1; hence, this last term is a function of the rate of response to first stage intervention. To understand this third term, it is useful to consider the following two extremes in the context of the prototypical SMART.

If both response rates (p_1, p_{-1}) are 1, then there is no re-randomization and the design is analogous to a 2-arm cluster-randomized RCT (here, the third term is equal to 1). If, on the other hand, both response rates are 0, then all clusters are randomized twice; here, the third term is equal to 2. Note how the third term is different for ADEPT and the prototypical SMART due to the difference in randomization schemes. Also, the special case where response rates to initial treatments are equal (i.e. $p_1 = p_{-1}$) leads to a clustered version of the sample size formula in Oetting et al.¹⁴

4.3 Sample size formulae with a cluster-level covariate

When including a cluster-level covariate in equations (1) or (2), working assumption 2 is similar for each corresponding design, except that it involves the conditional (on $\mathbf{X}_i$) marginal mean, i.e.

$E_{a_1, a_2}[(Y_i - \mu(\mathbf{X}_i, a_1, a_2) | (Y_i - \mu(\mathbf{X}_i, a_1, a_2) | R=0)] \preceq E_{a_1, a_2}[(Y_i - \mu(\mathbf{X}_i, a_1, a_2) | (Y_i - \mu(\mathbf{X}_i, a_1, a_2) | R=1)]$. Also, our formulae depends on $\text{Cor}(Y, X)$, which is the scalar correlation between the outcome Y_{ij} and the cluster-level covariate X_i under the DTRs in our test. Note that under assumptions 1 and 3, this correlation is constant across these DTRs (i.e.

$\text{Cor}^2(Y, X) \triangleq \text{Cor}_{1, b_2}^2(Y_{ij}, X_i) = \text{Cor}_{-1, c_2}^2(Y_{ij}, X_i)$). We obtain the following sample size formula for ADEPT

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m-1)|\rho^*| \cdot \left(1 + \frac{1-p_1}{2}\right)) \cdot [1 - \text{Cor}^2(Y, X)] \quad (7)$$

For the prototypical SMART, the sample size formula is

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} (1 + (m-1)|\rho^*| \left(1 + \frac{(1-p_1) + (1-p_{-1})}{2}\right)) [1 - \text{Cor}^2(Y, X)]$$

where $\rho^* = (\rho - \text{Cor}^2(Y, X)) / (1 - \text{Cor}^2(Y, X))$.

The use of a covariate leads to two changes in the sample size formulae. First, as expected,¹⁶ depending on the strength of the correlation between X and Y (i.e. $\text{Cor}^2(Y, X)$), the use of a covariate has the potential to reduce the minimum required sample size; this is because the use of covariates may improve the efficiency of our estimate of the coefficients β . Second, there is a reduction in sample size due to the reduction in correlation, ρ^* , which, by definition, is always less than ρ .

4.4 Using the sample size formula for the ADEPT study

To exemplify how the formula can be utilized in practice, we calculate how large of a difference between DTRs (1, -1) and (-1, .) we can detect in ADEPT. This difference would help us understand if it is better to give REP + EF + IF to non-responding clinics initially, or to delay REP + EF + IF until a clinic is non-responsive to REP + EF. In ADEPT, we expect the ICC of patient's MH-QOL to be $\rho = 0.01$ and the probability of responding when initially receiving REP + EF to be $p_1 = 0.2$. Using the true sample size of $N = 60$, a common

cluster size of $m = 10$, and performing an $\alpha = 0.05$ level test ($z_{\alpha/2} = 1.96$), by rearranging formula (5), we conclude that at 80% power ($z_{\beta} = 0.84$), we can detect an effect size of $\delta = 0.282$.

5 Simulations

Simulations were conducted to evaluate the developed formulae and understand their robustness to violations of the working assumptions. Specifically, we evaluate formulae under four scenarios: (1) satisfying all working assumptions, (2) violating working assumption 1, (3) violating working assumption 2, and (4) violating working assumption 3. For each scenario, we compare the nominal power of 0.9 with the estimated power (based on 1000 iterations). Here, we present results for ADEPT; results were similar for the prototypical SMART.

Details concerning the data generative model can be found in Appendix 2. Data were generated to mimic the ADEPT study. We considered different data generative scenarios with varied standardized effect sizes ($\delta = 0.2$ (small), 0.5 (moderate)), cluster sizes ($m = 5, 10, 20$), ICC (ρ or $\rho^* = 0.01$ or 0.1), and, when there is a cluster-level covariate, the correlation between X and Y ($\text{Cor}^2(Y, X) \in [0.04, 0.4]$). We also considered different scenarios constituting violations of the working assumptions (details below). For each scenario, the sample size was selected based on the proposed formulae with nominal power $(1 - \beta) = 0.9$ and Type-I error rate $\alpha = 0.05$; 1000 data sets were generated for each scenario. Each data set was analyzed as in sections 3.3 and 3.4, using the marginal mean model in equation (1).

Table 3 describes simulation results for the sample size formula in formula (5). To violate assumption 1, we made the response variance under DTR($1, b_2$) 1.5 times the response variance under DTR($-1, .$). We could have also violated this assumption by deviating from an exchangeable covariance structure; however, in cluster-randomized trials, it is rare to use another covariance structure.⁴⁷ To violate assumption 2, we made non-responders have significantly larger variance than responders under DTR($1, b_2$).

As expected, when no assumptions are violated (column 5), our estimated power is close to our pre-specified power, 0.9. When assumption 1 is violated (column 6) or assumption 2 is violated (column 7), we see that our power does not reduce dramatically. Hence, we conclude that our sample size formula is robust to violations of working assumptions 1 and 2. Also, the Type-I error rate does not depend on working assumptions 1 or 2. Hence, under each of these scenarios, with the effect size set to 0, the Type-I error rate is close to the nominal rate of 0.05.

Because working assumption 3 will always be true when there are no covariates, we run a second simulation, this time with a cluster-level covariate, to evaluate the robustness of the sample size formula in formula (7) to a violation of this assumption. Specifically, to violate assumption 3, we deviate from the linear marginal mean in equation (1) by generating data with $E_{a_1, a_2}(Y_{ij} | X_i) = \beta_0 + \beta_1 a_1 + \beta_2 a_2 I_{a_1=1} + \eta f_k(X_i)$ where $f_k(X_i) = X_i$ for $X_i \in [-k, k]$, $f_k(X_i) = k$ for $X_i > k$, and $f_k(X_i) = -k$ for $X_i < -k$ (i.e. the linear marginal mean is

misspecified outside of $[-k, k]$). Here η is chosen to maintain the same values of $\text{Cor}(Y, X)$. Setting $k = 2$ indicates a small violation (column 7) and setting $k = 1$ indicates a large violation (column 8). We still, however, analyze the data using the marginal mean model in equation (1). The results are in Table 4.

As expected, when no assumptions are violated (column 6), our estimated power is close to our pre-specified power, 0.9. Note the reduction in sample size caused by the addition of a covariate. Under a small violation (column 7), we see the power is not significantly reduced. Under a large violation (column 8), we see our power is lowest when X and Y are moderately correlated and the sample size is low. This is because when X and Y are weakly correlated, the overall influence of X is small, and hence misspecification of the relationship between X and Y will have little influence on our estimation and power. Also, once again, under these scenarios, with the effect size set to 0, the Type-I error rate is close to the nominal rate of 0.05.

6 Discussion and future work

This manuscript presents a regression estimator and sample size formulae for comparing embedded DTRs using data arising from a cluster-randomized SMART. Methods were motivated by the ADEPT SMART, a study designed to develop a DTR (at the level of community-based mental health clinics) to improve mental health outcomes for patients clustered within those sites.¹¹ Sample size formulae were derived for both ADEPT and for a more common type of SMART.

There are a number of directions for future research in the analysis of cluster-randomized SMARTs. First, relatively straightforward applications of the estimator in equation (3) with different link functions can be used to analyze, for example, binary, count, or zero-inflated outcomes.

Second, in practice, many cluster-randomized SMARTs will collect longitudinal (i.e. repeated measures) research outcomes at the patient-level. A natural next step is to combine the estimator presented here with methods for the analysis of longitudinal SMART outcomes²³ in order to accommodate two levels of clustering: repeated measures within patients within clusters.

Third, future work could also consider the use of variance components models, i.e. mixed effects or random effects models,^{2,48} which are now-standard in the analysis of randomized trials.

Fourth, while this manuscript focuses on the analysis of primary aims in a SMART, in the DTR literature, there is much interest in the development and application of analysis methods designed to generate hypotheses about more individually tailored DTRs.^{49–55} Much of this literature has focused on identifying optimal DTRs at the individual level. Such methods could be extended for the analysis of data arising from cluster-randomized SMARTs to develop optimal cluster-level DTRs.

There are also a number of interesting methodological issues related to the design of cluster-randomized SMARTs (with implications for analysis methods). First, the sample size formulae derived here were limited to cases where our data contain a single cluster-level covariate. Future work may provide extensions to data containing multiple covariates and individual-level covariates.

Second, in this manuscript, we focus on SMARTs that are useful for developing cluster-level DTRs where the initial and subsequent decisions are all at the cluster-level. However, there is currently much interest by educational scientists in SMARTs aimed at developing DTRs where sequences of intervention decisions are made at *both* the cluster and individual level. For example, we are currently involved in the conduct of a trial where the first stage intervention is at the level of classrooms with children with autism (such classrooms often include one to three children with autism), and the subsequent stages of intervention are at the level of the children themselves.⁵⁶

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. R code for implementing the weighted least squares regression estimator and the sample size formulae for both ADEPT and the prototypical SMART are provided on the first author's website. The authors would like to thank Xi (Lucy) Lu for the helpful comments.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by the following NIH grants: R01MH099898 (Kilbourne & Almirall), P50DA039838 (Almirall), R01HD073975 (Almirall), R01DA039901 (Almirall).

Appendix 1: Derivation of sample size formulae

Here we derive the sample size formulae. We begin by deriving the sample size formula for a prototypical design with no covariates. Then we make simple extensions to derive the formula under an ADEPT design and formulae when a cluster-level covariate is included. All sample size formulae are based on primary aim (iv), that is the marginal mean comparison of two DTRs that begin with a different initial treatment (i.e.

$$E_{1,b_2}(Y_{ij}) - E_{-1,c_2}(Y_{ij}).$$

For simplicity, these derivations assume the marginal mean model is parameterized as follows.

For the prototypical SMART, we postulate a model of the form

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 I\{a_1=1\} + \beta_1 I\{a_1=1, a_2=-1\} + \beta_2 I\{a_1=-1, a_2=1\} + \beta_3 I\{a_1=-1, a_2=-1\} + \boldsymbol{\eta}^T \mathbf{X}_{ij}$$

For ADEPT, we postulate a model of the form

$$\mu(\mathbf{X}_{ij}, a_1, a_2; \boldsymbol{\beta}, \boldsymbol{\eta}) = \beta_0 I\{a_1=1, a_2=1\} + \beta_1 I\{a_1=1, a_2=-1\} + \beta_2 I\{a_1=-1\} + \boldsymbol{\eta}^T \mathbf{X}_{ij}$$

Fitting the re-parameterized models will yield the exact same conclusions as fitting the marginal mean models in equations (1) and (2).

Prototypical design without covariates

For data arising from a prototypical SMART design, we derive the sample size formula for detecting a significant difference between mean outcomes from two treatment regimens, (b_1, b_2) and (c_1, c_2) . Because we are interested in comparing two regimes starting with a different initial treatment, without loss of generality, we let $b_1 = 1$ and $c_1 = -1$.

We are interested in testing the hypothesis

$$H_0: E_{1,b_2}(Y_{ij}) - E_{-1,c_2}(Y_{ij}) = 0$$

Against the alternative

$$H_1: E_{1,b_2}(Y_{ij}) - E_{-1,c_2}(Y_{ij}) = \delta\sigma$$

where $\sigma = \sqrt{(\sigma_{1,b_2}^2 + \sigma_{-1,c_2}^2)/2}$, $\sigma_{a_1,a_2}^2 = \text{var}_{a_1,a_2}(Y_{ij})$, and δ is the standardized effect size.

We make a series of assumptions to derive our sample size formulae. We highlight these assumptions throughout our derivation to illustrate their use in the calculation. Note that our sample size is developed for a fixed cluster size, m (extensions to the unequal cluster size case can be done as in Kerry et al.⁴⁶).

Our test statistic used for the hypothesis test is

$$Z = \frac{\sqrt{N}(\hat{\mu}(1, b_2) - \hat{\mu}(-1, c_2))}{\sqrt{\hat{\tau}^2(1, b_2) + \hat{\tau}^2(-1, c_2) - 2\widehat{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(-1, c_2))}}.$$

Here, $\hat{\tau}^2(a_1, a_2)$ is an estimate of the variance, $\hat{\tau}^2(a_1, a_2) = \text{Var}(\sqrt{N}\hat{\mu}(a_1, a_2))$ which can be calculated using the matrix, $\sum_{\hat{\theta}}$, given the supplementary material.

In large samples and under assumption 3 described in section 4, the distributions of $\hat{\mu}(1, b_2)$ and $\hat{\mu}(-1, c_2)$ can be approximated by a normal distribution, $\hat{\tau}^2(1, b_2) \approx \hat{\tau}^2(1, b_2)$, $\hat{\tau}^2(-1, c_2) \approx \hat{\tau}^2(-1, c_2)$, and $\widehat{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(-1, c_2)) \approx \text{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(-1, c_2)) = 0$ (the covariance is 0 due to the independence of estimators of marginal means with different initial treatments). Thus, Z approximately has a standard normal distribution under the null hypothesis.

Note that these calculations are exactly the same as those highlighted in section 3.5. Specifically, under the original parameterization, letting $\mathbf{c} = (0, 2, b_2 - c_2, b_2 + c_2, \mathbf{0}_p)^T$ then $\mathbf{c}^T(\hat{\beta}, \hat{\eta}) = \hat{\mu}(1, b_2) - \hat{\mu}(-1, c_2)$ and $\mathbf{c}^T \sum_{\hat{\beta}, \hat{\eta}} \mathbf{c} = \hat{\tau}^2(1, b_2) + \hat{\tau}^2(-1, c_2) - 2\widehat{Cov}(\sqrt{N}\hat{\mu}(1, b_2), \sqrt{N}\hat{\mu}(-1, c_2))$.

Under the alternative, our test statistic is normal with approximate mean

$\sqrt{N}\delta\sigma / \sqrt{\tau^2(1, b_2) + \tau^2(-1, c_2)}$ and variance 1. Doing standard power calculations¹⁴ for a hypothesis test of size α , in order to obtain desired power of $1 - \beta$, we need to find N that satisfies

$$z_\beta \approx -z_{\alpha/2} + \frac{\delta\sigma\sqrt{N}}{\sqrt{\tau^2(1, b_2) + \tau^2(-1, c_2)}} N = \frac{(z_\beta + z_{\alpha/2})^2 (\tau^2(1, b_2) + \tau^2(-1, c_2))}{\delta^2 \sigma^2} \quad (8)$$

Everything in this formula can be explicitly found except $\tau^2(1, b_2)$ and $\tau^2(-1, c_2)$. Hence, we now aim to derive upper bounds for these variables in order to write our sample size formula in terms of either known or easily elicited quantities.

Note that under the parameterization, for any DTR (a_1, a_2) , $\tau^2(a_1, a_2)$

$$\tau^2(a_1, a_2) = \text{Var}[\sqrt{N}\hat{\mu}(a_1, a_2)] = [\sum_{\hat{\beta}}]_{(a_1, a_2)} = [\mathbf{J}^{-1} E[\mathbf{U}_i \mathbf{U}_i^T] \mathbf{J}^{-1}]_{(a_1, a_2)}, \text{ with } \mathbf{U}_i \text{ and } \mathbf{J}$$

defined in the s. Here, to simplify notation, in the 4×4 matrices, $M = \sum_{\hat{\beta}} \mathbf{J}$, or $E[\mathbf{U}_i \mathbf{U}_i^T]$, we define $[M]_{(a_1, a_2)}$ as the diagonal element corresponding to DTR (a_1, a_2) (e.g. the (3,3)

element for DTR $(-1, 1)$). We define the $m \times m$ matrix $\sum_{a_1, a_2} := \text{Cov}_{a_1, a_2}(\mathbf{Y}_i)$. Also, the vector $\mathbf{1}_m$ is defined as the $m \times 1$ vector of 1's. Lastly, as defined in section 4, p_{a_1} is the probability of responding given the cluster had received initial treatment a_1 .

After simplification, we find that $\mathbf{J}$ is a diagonal matrix with diagonal elements

$$[\mathbf{J}]_{(a_1, a_2)} = E[W_i I_{i, a_1, a_2} \mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} \mathbf{1}_m] = \mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} \mathbf{1}_m$$

For $E[\mathbf{U}_i \mathbf{U}_i^T]_{(a_1, a_2)}$, we perform the following simplification

$$\begin{aligned} E[\mathbf{U}_i \mathbf{U}_i^T]_{(a_1, a_2)} &= E[W_i^2 I_{i, a_1, a_2} \mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} (\mathbf{Y}_i \\ &\quad - \boldsymbol{\mu}(a_1, a_2) \| (\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2) \|^T \mathbf{V}_{i, a_1, a_2}^{-1} \mathbf{1}_m)] \\ &= \mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} E_{a_1, a_2} [W_i (\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2) \| (\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2) \|^T \mathbf{V}_{i, a_1, a_2}^{-1} \mathbf{1}_m \\ &= \mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} [2E_{a_1, a_2} [(\mathbf{Y}_i \\ &\quad - \boldsymbol{\mu}(a_1, a_2) \| (\mathbf{Y}_i - \boldsymbol{\mu}(a_1, a_2) \|^T R = 1] p_{a_1} \end{aligned}$$

$$\begin{aligned}
& +4E_{a_1,a_2}[(Y_i \\
& -\mu(a_1,a_2)\|(Y_i-\mu(a_1,a_2))^T R \\
& =0](1 \\
& -p_{a_1})\|\mathbf{V}_{i,a_1,a_2}^{-1}\mathbf{1}_m=2\cdot\mathbf{1}_m^T\mathbf{V}_{i,a_1,a_2}^{-1}\sum_{a_1,a_2}\mathbf{V}_{i,a_1,a_2}^{-1}\mathbf{1}_m+2(1-p_{a_1})\cdot\mathbf{1}_m^T\mathbf{V}_{i,a_1,a_2}^{-1}E_{a_1,a_2}[(Y_i \\
& -\mu(a_1,a_2)\|(Y_i-\mu(a_1,a_2))^T R \\
& =0]\mathbf{V}_{i,a_1,a_2}^{-1}\mathbf{1}_m
\end{aligned}$$

To go from line 1 to 2, we assume Robin's consistency assumption holds, i.e. that the cluster's observed outcomes equal the cluster's potential outcomes under the observed DTR.
⁵⁷ Under this assumption, we are able to switch from E , which is an expected value over observed data, to E_{a_1,a_2} which is the expected value had the entire population received DTR (a_1, a_2) .^{31,32}

For further simplification, we now make assumption 2. This assumption is equivalent to assuming, for a specific DTR (a_1, a_2) (we drop the a_1, a_2 from the subscripts for convenience)

$$|(\sigma_R^2\rho_R - \sigma_{NR}^2\rho_{NR})p_{a_1} + (\mu_R - \mu_{NR})^2(p_{a_1}|(1 - 2p_{a_1})| \leq (\sigma_R^2 - \sigma_{NR}^2)p_{a_1} + (\mu_R - \mu_{NR})^2(p_{a_1}|(1 - 2p_{a_1})|$$

, where $\mu_R, \sigma_R^2, \rho_R$ are the mean, variance, and ICC of responders had the whole population received DTR (a_1, a_2) , (i.e. $\mu_R = E_{a_1,a_2}(Y_{ij}|R_i=1)$, $\sigma_R^2 = \text{Var}_{a_1,a_2}(Y_{ij}|R_i=1)$, $\rho_R = \text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}|R_i=1)/\text{Var}_{a_1,a_2}(Y_{ij}|R_i=1)$, similarly defined for NR and non-responders (i.e. conditional on $R=0$). Again, p_{a_1} is the probability of response, given initial treatment a_1 .

For DTR (a_1, a_2) , this condition is satisfied if the probability of response is less than or equal to 0.5 (which is typical for prototypical SMART designs), the non-responders of that regimen have a variance which is less than or equal to the variance of responders of the regimen, and both responders and non-responders have similar within cluster covariances.

Under assumption 2 and using our previous simplification, we can bound our expression for $E[\mathbf{U}_i\mathbf{U}_i^T]_{(a_1,a_2)}$ by

$$E[\mathbf{U}_i\mathbf{U}_i^T]_{(a_1,a_2)} \leq 2(1 + (1 - p_{a_1})\|\cdot\mathbf{1}_m^T\mathbf{V}_{i,a_1,a_2}^{-1}\sum_{a_1,a_2}\mathbf{V}_{i,a_1,a_2}^{-1}\mathbf{1}_m$$

We next utilize the fact that our working covariance matrix, $\mathbf{V}_{i,a_1,a_2}$, is exchangeable. With some linear algebra, this assumption allows us to perform the following simplification

$$\frac{2(2 - p_{a_1})\|\mathbf{1}_m^T\mathbf{V}_{i,a_1,a_2}^{-1}\sum_{a_1,a_2}\mathbf{V}_{i,a_1,a_2}^{-1}\mathbf{1}_m}{(\mathbf{1}_m^T\mathbf{V}_{i,a_1,a_2}^{-1}\mathbf{1}_m)^2} = \frac{2(2 - p_{a_1})\|\mathbf{1}_m^T\sum_{a_1,a_2}\mathbf{1}_m}{m^2}$$

Next, using assumption 1, we exploit the exchangeable population covariance structure (i.e.

$\text{Cov}_{a_1, a_2}(\mathbf{Y}_i | \sigma_{a_1, a_2}^2 \cdot \text{Exch}(\rho_{a_1, a_2})$, where $\rho_{a_1, a_2} = \text{Cor}_{a_1, a_2}(Y_{i1}, Y_{i2})$). Putting everything together, we obtain

$$\begin{aligned} \tau^2(a_1, a_2) &= \text{Var}[\sqrt{N}\hat{\mu}(a_1, a_2)] \stackrel{3}{=} [\mathbf{J}^{-1}E[\mathbf{U}_i\mathbf{U}_i^T]\mathbf{J}^{-1}]_{(a_1, a_2)} \stackrel{2}{\leq} \frac{2(2 - p_{a_1}|\mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} \sum_{a_1, a_2} \mathbf{V}_{i, a_1, a_2}^{-1} \mathbf{1}_m|}{(\mathbf{1}_m^T \mathbf{V}_{i, a_1, a_2}^{-1} \mathbf{1}_m|^2} \\ &= \frac{2(2 - p_{a_1}|\mathbf{1}_m^T \sum_{a_1, a_2} \mathbf{1}_m|}{m^2} \stackrel{1}{=} \frac{2(2 - p_{a_1}|\sigma_{a_1, a_2}^2[1 + (m - 1)|\rho_{a_1, a_2}]}{m} \end{aligned} \quad (9)$$

where the numbering above the equalities and inequalities illustrates which assumption is being used.

We utilize the across regimen covariance equality of assumption 1 in order to simplify things further. Note that if one had good estimates of σ_{a_1, a_2}^2 and ρ_{a_1, a_2} , then you could easily obtain N by plugging in the values into equation (9) and then using these estimates in equation (8).

Using this equality, we combine equation (9) with equation (8) and simplify to obtain

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m - 1)|\rho|) \cdot \left(1 + \frac{(1 - p_1) + (1 - p_{-1})}{2}\right)$$

ADEPT design without covariates

All the calculations done above are nearly identical for the ADEPT case. The only major difference arises from the lack of re-randomization of clusters receiving initial treatment $a_1 = -1$. This in fact makes the calculations simpler for $\tau^2(-1, \cdot)$. In particular, we assume assumptions 1 and 3; however, assumption 2 only needs to be assumed for DTR $(1, b_2)$. Under these assumptions we obtain

$$\tau^2(1, b_2) \leq \frac{2(2 - p_1|\sigma_{1, b_2}^2[1 + (m - 1)|\rho_{1, b_2}]}{m}, \tau^2(-1, \cdot) = \frac{2\sigma_{-1}^2[1 + (m - 1)|\rho_{-1, \cdot}]}{m} \quad (10)$$

After utilizing the across regimen population covariance equality of assumption 1, we combine equation (10) with equation (8) and simplify to obtain

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} \cdot (1 + (m - 1)|\rho|) \cdot \left(1 + \frac{1 - p_1}{2}\right)$$

For this sample size formula, we actually only need to assume $\sigma_{1,b_2}^2 \leq \sigma_{-1,}^2$ (as opposed to the equality assumed in assumption 1) to ensure our power is larger than $1 - \beta$.

With a cluster-level covariate

In this section, we write the sample size formulae when adding a single cluster-level covariate to the model (i.e. the $m \times 1$ vector $\mathbf{X}_j \triangleq (X_{j1}, X_{j2}, \dots, X_{jm})^T = (X_{j1}, X_{j2}, \dots, X_{jm})^T$).

First, for DTR (a_1, a_2) , we define $\sigma_{a_1,a_2}^{2*} \triangleq E_{a_1,a_2}(\text{Var}_{a_1,a_2}(Y_{ij}|X_i|))$, and

$\rho_{a_1,a_2}^* \triangleq E_{a_1,a_2}(\text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}|X_i|)/\sigma_{a_1,a_2}^{2*})$, where the expectations are taken over X_i . Note in the homoscedastic case, the expectation is unnecessary since the conditional variances and covariances are constant for all X_i .

The key to extending our formulae to the covariate case is observing that the numerator of our sample size formulae will now be in terms of the average *conditional* (on X) variances and ICCs, σ_{a_1,a_2}^{2*} and ρ_{a_1,a_2}^* while the denominator remains in terms of the overall variance, σ_{a_1,a_2}^2 . Since $\sigma_{a_1,a_2}^2 = \sigma_{a_1,a_2}^{2*} + \eta^2 \text{Var}(X_i)$ and

$\text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}) = E_{a_1,a_2}[\text{Cov}_{a_1,a_2}(Y_{i1}, Y_{i2}|X_i|) + \eta^2 \text{Var}(X_i)]$, then σ_{a_1,a_2}^{2*} and ρ_{a_1,a_2}^* must be less than or equal to σ_{a_1,a_2}^2 and ρ_{a_1,a_2} . Thus, the numerator of our formulae is reduced while the denominator remains the same. With some algebra, this reduction is shown to be $1 - \text{Cor}^2(Y_{j1}, X_j)$. Note that this reduction can be shown to be the same reduction arising from including a cluster-level covariate in clustered RCTs, see Hedges et al.⁵⁸

For simplification, we do all calculations assuming our covariate has mean 0 (this eliminates covariance between our marginal mean estimates). When our covariate does not have mean 0, one can show that mean centering our covariate does not change the value of our test statistic, and hence does not change our power. Thus, our sample size formulae remain valid when the covariate does not have mean 0.

Prototypical design

Using assumptions 1–3, the relationships above, and doing similar algebra as in the non-covariate case (except now everything is in terms of σ_{a_1,a_2}^{2*} and ρ_{a_1,a_2}^*), we obtain for both DTRs of interest

$$\tau^2(a_1, a_2) \leq \frac{2(2 - p_1|\sigma_{a_1,a_2}^{2*}|[1 + (m - 1)|\rho_{a_1,a_2}^*|])}{m} \quad (11)$$

Making assumption 1 (on marginal population variances and correlations) will lead to equality of expected conditional variances and correlations due to the simple relationship between the conditional and marginal variances and covariances highlighted above. Hence, we define $\rho^* \triangleq \rho_{1,b_2}^* = \rho_{-1,c_2}^*$ and $\sigma^{2*} \triangleq \sigma_{1,b_2}^{2*} = \sigma_{-1,c_2}^{2*}$.

We also take advantage of the fact that

$\text{Cor}_{1,b_2}^2(Y_{ij}, X_i) = \eta^2 \text{Var}(X_i | \sigma_{1,b_2}^2) = \eta^2 \text{Var}(X_i | \sigma_{-1,c_2}^2)$, i.e. is also equal across both regimes.

We define $\text{Cor}^2(Y, X) = \text{Cor}_{1,b_2}^2(Y_{ij}, X_i) = \text{Cor}_{-1,c_2}^2(Y_{ij}, X_i)$.

Ultimately, we obtain

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} (1 + (m-1)|\rho^*| \left(1 + \frac{(1-p_1) + (1-p_{-1})}{2} \right) [1 - \text{Cor}^2(Y, X)])$$

ADEPT design

Similar to the prototypical design, under assumptions 1–3, for DTR (1, b_2), we obtain the same bound as in equation (11). For DTR (–1, .), without utilizing assumption 2, we obtain

$$\tau^2(-1, .) = \frac{2\sigma_{-1,.}^{2*} [1 + (m-1)|\rho_{-1,.}^*|]}{m}$$

And, utilizing equality of population covariance across regimens, we end with

$$N = \frac{4(z_\beta + z_{\alpha/2})^2}{m\delta^2} (1 + (m-1)|\rho^*| \left(1 + \frac{1-p_1}{2} \right) [1 - \text{Cor}^2(Y, X)])$$

Also, with some algebra, ρ^* can be expressed as $\rho^* = (\rho - \text{Cor}^2(Y, X)) / (1 - \text{Cor}^2(Y, X))$, allowing our two covariate sample size formulae to be a function purely of ρ and $\text{Cor}^2(Y, X)$.

Appendix 2: Data-generative models used in simulation experiments

Below we describe how we generated data for our simulations.

Without covariates

For Table 3, we generate data ($A_1, R, A_2, \mathbf{Y}$), using the values in Table 5 and, for each of the N clusters, doing the following:

1. Generate A_1 to be 1 or –1 with equal probability.
2. Generate R to be 1 with probability p_{A_1} and 0 otherwise.
3. Generate A_2 to be 1 or –1 with equal probability, for clusters with $A_1 = 1, R = 0$.
4. Generate the $m \times 1$ vector $\mathbf{Y} = \mu_{A_1, R, A_2} \mathbf{1}_m + \varepsilon$, where $\varepsilon \sim \text{MVN}(\mathbf{0}_m, \sum_{A_1, R, A_2} \sigma_{A_1, R, A_2}^2 \cdot \text{Exch}_m(\rho_{A_1, R, A_2})$. Here μ_{A_1, R, A_2} , σ_{A_1, R, A_2}^2 , ρ_{A_1, R, A_2} are the cell means, variances, and ICCs since they correspond to each cell in Figure 1.

Under the specified means, variances, and ICCs in Table 5, one can easily obtain the desired marginal (over R) means, variances, and ICCs under a specific DTR using the laws of total expectation and variation. For example, to obtain the marginal mean under DTR (1,1), one would calculate $\mu_{1,1}, p_1 + \mu_{1,0,1}(1 - p_1)$. To calculate the variance under DTR (1,1), one would calculate $\sigma_{1,1}^2, p_1 + \sigma_{1,0,1}^2(1 - p_1) + p_1(1 - p_1)(\mu_{1,1} - \mu_{1,0,1})^2$. To calculate the covariance under DTR (1,1), one would calculate

$$\sigma_{1,1}, \rho_{1,1}, p_1 + \sigma_{1,0,1}^2 \rho_{1,0,1}(1 - p_1) + p_1(1 - p_1)(\mu_{1,1} - \mu_{1,0,1})^2.$$

When no assumptions were violated (row 1 of Table 5), the cell means and variances were first chosen to give marginal means and variances which are both similar to results expected in ADEPT and produce effect sizes matching Table 3. After obtaining the correct effect size, the cell ICCs were then chosen also to match values specified in Table 3. To violate assumptions (row 2 and 3 of Table 5), the cell means, variances, and ICCs from row 1 were altered to create the correct violations.

With a cluster-level covariate

To generate data for Table 4, we use a continuous cluster-level covariate. We generate data (X, A_1, R, A_2, Y) , using the values in Table 6 and, for each of the N clusters, doing the following:

1. Generate A_1 to be 1 or -1 with equal probability.
2. Generate R to be 1 with probability p_{A_1} and 0 otherwise.
3. Generate A_2 to be 1 or -1 with equal probability, for clusters with $A_1 = 1, R = 0$.
4. Generate a single cluster-level covariate X from Normal $(0,1)$.
5.
 - a. Generate $m \times 1$ vector $Y = (\mu_{A_1, R, A_2} + \eta X | \mathbf{1}_m + \varepsilon$ for Column 6
 - b. Generate $m \times 1$ vector $Y = (\mu_{A_1, R, A_2} + \eta f_k(X) | \mathbf{1}_m + \varepsilon$ for Columns 7,8.

where $\varepsilon \sim \text{MVN}(\mathbf{0}_m, \sum_{A_1, R, A_2} | \text{with } \sum_{A_1, R, A_2} = \sigma_{A_1, R, A_2}^{2*} \cdot \text{Exch}_m(\rho_{A_1, R, A_2}^*)$. Here μ_{A_1, R, A_2} , $\sigma_{A_1, R, A_2}^{2*}$, ρ_{A_1, R, A_2}^* are the cell means, conditional cell variances, and conditional cell ICCs since they correspond to each cell in Figure 1. Also, f_k is the same piecewise function defined in section 5 (i.e. which is non-linear outside of $[-k, k]$).

Under the specified conditional means, variances, and ICCs in Table 6, one can again obtain the desired conditional (on X only) and marginal means, variances, and ICCs under a specific DTR using the laws of total expectation and variation. For example, to obtain the conditional variance under DTR (1,1), one would calculate

$$\sigma_{1,1}^{2*} = \sigma_{1,1}^{2*}, p_1 + \sigma_{1,0,1}^{2*}(1 - p_1) + p_1(1 - p_1)(\mu_{1,1} - \mu_{1,0,1})^2. \text{ For data generated as in step 5(a), to obtain the marginal variance under DTR (1,1), one would calculate } \sigma_{1,1}^2 = \sigma_{1,1}^{2*} + \eta^2 \text{Var}(X).$$

For data generated as in step 5(b), we instead calculate $\sigma_{1,1}^2 = \sigma_{1,1}^{2*} + \eta^2 \text{Var}(f_k(X))$. Both $\text{Var}(X)$ and $\text{Var}(f_k(X))$ can be found using the known distribution of X .

The cell means, variances, and ICCs were chosen for the same reason as in the non-covariate case. The parameter, η , was chosen to match the $\text{Cor}(Y, X)$ values in Table 4.

Table 5

Pre-specified simulation values for Table 3.

Simulation	p_{-1}	p_{-1}	$\mu_{1,1,}$	$\mu_{1,0,1}$	$\mu_{1,0,-1}$	$\mu_{-1,1,}$	$\mu_{-1,0,}$	$\sigma_{1,1,}^1$	
Table 3, Row 1, Col 5	0.2	0.3	34.71	32.71	28	32.7	31	63.36	
Table 3, Row 1, Col 6	0.2	0.3	34.71	32.71	28	32.14	31.44	63.36	
Table 3, Row 1, Col 7	0.2	0.3	33.36	33.05	28	32.7	31	1	
Simulation	$\sigma_{1,0,1}^2$	$\sigma_{1,0,-1}^2$	$\sigma_{-1,1,}^2$	$\sigma_{-1,0,}^2$	$\rho_{1,1,}$	$\rho_{1,0,1}$	$\rho_{1,0,-1}$	$\rho_{-1,1,}$	$\rho_{-1,0,}$
Table 3, Row 1, Col 5	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006
Table 3, Row 1, Col 6	63.36	60	43	43	0.0	0.0	0.0	0.0076	0.0076
Table 3, Row 1, Col 7	79.73	60	63.39	63.39	0.9	0.007	0.0	0.0006	0.0006

Table 6

Pre-specified simulation values for Table 4.

Simulation	k	p_1	p_{-1}	η	$\mu_{1,1,}$	$\mu_{1,0,1}$	$\mu_{1,0,-1}$	$\mu_{-1,1,}$	$\mu_{-1,0,}$	$\sigma_{1,1,}^{2*}$
Table 4, Row 1, Col 6		0.2	0.3	4.47	34.94	32.94	28	32.7	31	63.36
Table 4, Row 1, Col 7	2	0.2	0.3	4.69	34.95	32.95	28	32.7	31	63.36
Table 4, Row 1, Col 8	1	0.2	0.3	6.66	34.98	32.98	28	32.7	31	63.36
Simulation		$\sigma_{1,0,1}^{2*}$	$\sigma_{1,0,-1}^{2*}$	$\sigma_{-1,1,}^{2*}$	$\sigma_{-1,0,}^{2*}$	$\rho_{1,1,}^*$	$\rho_{1,0,1}^*$	$\rho_{1,0,-1}^*$	$\rho_{-1,1,}^*$	$\rho_{-1,0,}^*$
Table 4, Row 1, Col 6	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006	
Table 4, Row 1, Col 7	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006	
Table 4, Row 1, Col 8	63.36	60	63.39	63.39	0.0	0.0	0.0	0.0006	0.0006	

References

1. Murray, DM. Design and analysis of group-randomized trials. Vol. 29. USA: Oxford University Press; 1998.
2. Raudenbush, SW., Bryk, AS. Hierarchical linear models: applications and data analysis methods. Vol. 1. Thousand Oaks, California: Sage; 2002.

3. Donner, A., Klar, N. Design and analysis of cluster randomization trials in health research. Vol. 1. New York, New York: Wiley; 2010.
4. Kilbourne AM, Abraham KM, Goodrich DE, et al. Cluster randomized adaptive implementation trial comparing a standard versus enhanced implementation intervention to improve uptake of an effective re-engagement program for patients with serious mental illness. *Implement Sci.* 2013; 8:1–14. [PubMed: 23279972]
5. Murphy S. An experimental design for the development of adaptive treatment strategies. *Stat Med.* 2005; 24:1455–1481. [PubMed: 15586395]
6. Lei H, Nahum-Shani I, Lynch K, et al. A SMART design for building individualized treatment sequences. *Ann Rev Clin Psychol.* 2012; 8:21–48. [PubMed: 22224838]
7. Chakraborty, B., Moodie, E. Statistical methods for dynamic treatment regimes. New York, New York: Springer; 2013.
8. Lavori PW, Dawson R. Introduction to dynamic treatment strategies and sequential multiple assignment randomization. *Clin Trials.* 2014; 11:393–399. [PubMed: 24784487]
9. Kosorok, MR., Moodie, EE. Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine. Philadelphia, Pennsylvania: SIAM; 2015.
10. Methodology Center. Example SMART studies. 2016. <https://methodology.psu.edu/ra/adap-inter/projects> accessed 20 April 2017
11. Kilbourne AM, Almirall D, Eisenberg D, et al. Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. *Implement Sci.* 2014; 9:132. [PubMed: 25267385]
12. Nahum-Shani I, Qian M, Almirall D, et al. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychol Meth.* 2012; 17:457–477.
13. Orellana L, Rotnitzky A, Robins J. Dynamic regime marginal structural mean models for estimating optimal dynamic treatment regimes, part I: main content. *Int J Biostat.* 2010; 6 Article 8.
14. Oetting, A., Levy, J., Weiss, R., et al. Statistical methodology for a SMART design in the development of adaptive treatment strategies. In: Shrout, P., Keyes, K., Ornstein, K., editors. *Causality and psychopathology: finding the determinants of disorders and their cures.* Arlington, VA: Oxford University Press; 2010. p. 179–205.
15. Li Z, Murphy SA. Sample size formulae for two-stage randomized trials with survival outcomes. *Biometrika.* 2011; 98:503–518. [PubMed: 22363091]
16. Spybrook, J., Bloom, H., Congdon, R., et al. Optimal design plus empirical evidence. 2011. <http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od> accessed 20 April 2017
17. Ghosh, P., Cheung, Y., Chakraborty, B. Sample size calculations for clustered SMART designs. In: Kosorok, MR., Moodie, EE., editors. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*, chapter 5. Alexandria, Virginia: SIAM; 2015. p. 55–70.
18. Lavori P, Dawson D. A design for testing clinical strategies: biased individually tailored within-subject randomization. *J Royal Stat Soc Ser A.* 2000; 163:29–38.
19. Kilbourne AM, Goodrich DE, Lai Z, et al. Life goals collaborative care for patients with bipolar disorder and cardiovascular disease risk. *Psychiat Serv.* 2012; 63:1234–1238.
20. Kasari C, Kaiser A, Goods K, et al. Communication interventions for minimally verbal children with autism: sequential multiple assignment randomized trial. *J Am Acad Child Adolesc Psychiatr.* 2014; 53:635–646.
21. Almirall D, DiStefano C, Chang YC, et al. Longitudinal effects of adaptive interventions with a speech-generating device in minimally verbal children with ASD. *J Clin Child Adolesc Psychol.* 2016; 45:442–456.
22. Chronis-Tuscano A, Wang CH, Strickland J, et al. Personalized treatment of mothers with ADHD and their young at-risk children: a SMART pilot. *J Clin Child Adolesc Psychol.* 2016; 45:510–521.
23. Lu X, Lynch KG, Oslin DW, et al. Comparing treatment policies with assistance from the structural nested mean model. *Biometrics.* 2016; 72:10–19. [PubMed: 26363892]

24. Pelham WE, Fabiano GA, Waxmonsky JG, et al. Treatment sequencing for childhood adhd: a multiple-randomization study of adaptive medication and behavioral interventions. *J Clin Child Adoles Psychol.* 2016; 45:396–415.
25. Gunlicks-Stoessel M, Mufson L, Westervelt A, et al. A pilot SMART for developing an adaptive treatment strategy for adolescent depression. *J Clin Child Adoles Psychol.* 2016; 45:480–494.
26. August GJ, Piehler TF, Bloomquist ML. Being “smart” about adolescent conduct problems prevention: executing a SMART pilot study in a juvenile diversion agency. *J Clin Child Adoles Psychol.* 2016; 45:495–509.
27. Sherwood NE, Butryn ML, Forman EM, et al. The bestfit trial: a smart approach to developing individualized weight loss treatments. *Contem Clin Trials.* 2016; 47:209–216.
28. Naar-King S, Ellis DA, Idalski Carcone A, et al. Sequential multiple assignment randomized trial (SMART) to construct weight loss interventions for African American adolescents. *J Clin Child Adoles Psychol.* 2016; 45:428–441.
29. McKay JR, Drapkin ML, Van Horn DH, et al. Effect of patient choice in an adaptive sequential randomization trial of treatment for alcohol and cocaine dependence. *J Consult Clin Psychol.* 2015; 83:1021. [PubMed: 26214544]
30. Almirall D, Nahum-Shani I, Sherwood N, et al. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Transl Behav Med.* 2014; 4:260–274. [PubMed: 25264466]
31. Neyman J, Iwaskiewicz K, Kolodziejczyk S. Statistical problems in agricultural experimentation. *Suppl J Royal Stat Soc.* 1935; 2:107–180.
32. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat.* 1978; 6:34–58.
33. Robins JM. Association, causation, and marginal structural models. *Synthese.* 1999; 121:151–179.
34. Hernán M, Brumback B, Robins J. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology.* 2000; 11:561–570. [PubMed: 10955409]
35. Robins JM, Hernan M, Brumback B. Marginal structural models and causal inference. *Epidemiology.* 2000; 11:550–560. [PubMed: 10955408]
36. Murphy SA, van der Laan MJ, Robins JM, et al. Marginal mean models for dynamic regimes. *J Am Stat Assoc.* 2001; 96:1410–1423. [PubMed: 20019887]
37. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; 73:13–22.
38. Liang KY, Zeger SL. Regression analysis for correlated data. *Ann Rev Pub Health.* 1993; 14:43–68. [PubMed: 8323597]
39. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc.* 1995; 90:106–121.
40. Hernan M, Brumback B, Robins J. Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Stat Med.* 2002; 21:1689–1709. [PubMed: 12111906]
41. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica.* 2003; 71:1161–1189.
42. Bembom O, van der Laan M. Statistical methods for analyzing sequentially randomized trials. *J Natl Cancer Inst.* 2007; 99:1577–1582. [PubMed: 17971533]
43. Brumback BA. A note on using the estimated versus the known propensity score to estimate the average treatment effect. *Stat Prob Lett.* 2009; 79:537–542.
44. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Stat Med.* 2014; 33:721–737. [PubMed: 24114884]
45. Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
46. Kerry SM, Bland MJ. Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. *Stat Med.* 2001; 20:377–390. [PubMed: 11180308]

47. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev.* 2009; 77:378–394.
48. Hedeker, D., Gibbons, RD. *Longitudinal data analysis*. Vol. 451. Hoboken, New Jersey: John Wiley & Sons; 2006.
49. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. *Ann Stat.* 2011; 39:1180. [PubMed: 21666835]
50. Linn KA, Laber EB, Stefanski LA. Interactive q-learning for quantiles. *Journal of the American Statistical Association.* 2017; 112:638–649. [PubMed: 28890584]
51. Moodie EE, Dean N, Sun YR. Q-learning: flexible learning about useful utilities. *Stat Biosci.* 2014; 6:223–243.
52. Laber EB, Zhao Y. Tree-based methods for individualized treatment regimes. *Biometrika.* 2015; 102:501–514. [PubMed: 26893526]
53. Zhang Y, Laber EB, Tsiatis A, et al. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics.* 2015; 71:895–904. [PubMed: 26193819]
54. Zhao YQ, Zeng D, Laber EB, et al. New statistical learning methods for estimating optimal dynamic treatment regimes. *J Am Stat Assoc.* 2015; 110:583–598. [PubMed: 26236062]
55. Zhou, X., Mayer-Hamblett, N., Khan, U., et al. Residual weighted learning for estimating individualized treatment rules. *J Am Stat Assoc.* 2015. <http://dx.doi.org/10.1080/01621459.2015.1093947>. (accessed 20 April 2017)
56. Kasari, C., Gulsrud, A., Almirall, D. Getting SMART about social and academic engagement of elementary aged students with autism spectrum disorder. 2016. p. 1758 <http://ies.ed.gov/funding/grantsearch/details.asp?ID> accessed 20 April 2017
57. Robins, J. Causal inference from complex longitudinal data. In: Maia, B., editor. *Latent=variable modeling and applications to causality Lecture notes in statistics*. New York: Springer-Verlag; 1997.
58. Hedges, L., Rhoads, C. *Statistical power analysis in education research*. National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education; 2009. NCSER 2010-3006
59. Lu X, Nahum-Shani I, Kasari C, et al. Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in SMART studies. *Stat Med.* 2016; 35:1595–1615. [PubMed: 26638988]

Figure 1. Schematic of ADEPT. The encircled R signifies randomization; cluster-level randomizations occurred at baseline and after six months of REP + EF or REP + EF + IF following identification of clinic responder status.

Figure 2.
Schematic of a prototypical SMART design.

Table 1

The three DTRs embedded in ADEPT (Figure 1).

DTR Label (a_1, a_2)	Second-stage treatment	Status at end of second-stage	Third-stage treatment	A_1	R	A_2	Cell in figure	Known IPW
(1, 1)	REP + EF	Resp	REP	1	1	1	A	2
(1, -1)	REP + EF	Non Resp	REP + EF	1	0	1	B	4
		Resp	REP	1	1	1	A	2
(-1, .)	REP + EF + IF	Non Resp	REP + EF + IF	1	0	-1	C	4
		Resp	REP	-1	1	1	D	2
		Non Resp	REP + EF + IF	-1	0	0	E	2

IPW: inverse probability weights.

Table 2

The four DTRs embedded in a prototypical SMART (Figure 2).

DTR Label (a_1, a_2)	First-stage treatment	Status at end of first-stage	Second-stage treatment	A_1	R	A_2	Cell in figure	Known IPW
(1, 1)	T1 ₁	Resp	T2 ₁	1	1	1	A	2
		Non resp	T2 ₂	1	0	1	B	4
(1, -1)	T1 ₁	Resp	T2 ₁	1	1	1	A	2
		Non resp	T2 ₃	1	0	-1	C	4
(-1, 1)	T1 ₂	Resp	T2 ₄	-1	1	1	D	2
		Non resp	T2 ₅	-1	0	1	E	4
(-1, -1)	T1 ₂	Resp	T2 ₄	-1	1	1	D	2
		Non resp	T2 ₆	-1	0	-1	F	4

IPW: inverse probability weights.

Table 3

Power analysis of formula (5).

ICC, ρ	Effect size, δ	Cluster size, m	Sample size, N	Assumptions are correct	Violating assumption 1	Violating assumption 2
0.01	0.2	5	306	0.894	0.891	0.886
		20	88	0.917	0.890	0.876 ^a
	0.5	5	49	0.909	0.898	0.880 ^a
10		26	0.906	0.878 ^a	0.893	
0.1	0.2	5	412	0.910	0.901	0.870 ^a
		20	213	0.922 ^a	0.902	0.891
	0.5	5	66	0.909	0.888	0.898
20		34	0.915	0.913	0.889	

^aThe proportion is significantly different from 0.9 at the 5% level.

ICC: intra-cluster correlation.

Table 4

Power analysis of formula (7).

ICC, ρ^*	Effect size, δ	Cluster size, m	Cor ² (Y, X)	Sample size, N	Assumptions are correct	Small violation of assumption 3	Large violation of assumption 3
0.01	0.2	5	0.238	233	0.909	0.904	0.859 ^a
		20	0.238	65	0.903	0.879 ^a	0.777 ^a
	0.5	5	0.043	47	0.891	0.901	0.902
		10	0.066	24	0.893	0.903	0.897
0.1	0.2	5	0.243	305	0.918	0.900	0.890
		20	0.243	159	0.915	0.922 ^a	0.864 ^a
	0.5	5	0.043	63	0.898	0.916	0.919 ^a
		20	0.043	32	0.908	0.920 ^a	0.899

^aThe proportion is significantly different from 0.9 at the 5% level.

ICC: intra-cluster correlation.