Biostatistics (2020) **21**, 3, *pp.* 432–448
doi:10.1093/biostatistics/kxy064
Advance Access publication on October 30, 2018

Power analysis in a SMART design: sample size estimation for determining the best embedded dynamic treatment regime

WILLIAM J. ARTMAN*

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, Saunders Research Building, 265 Crittenden Blvd., NY 14642, USA

William_Artman@URMC.Rochester.edu

INBAL NAHUM-SHANI

Institute for Social Research, University of Michigan, 426 Thompson St, Ann Arbor, MI 48106, USA

TIANSHUANG WU

AbbVie Inc., 1 North Waukegan Road, North Chicago, IL 60064, USA

JAMES R. MCKAY

Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, 3535 Market St., Suite 500, Philadelphia, PA 19104, USA

ASHKAN ERTEFAIE

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Saunders Research Building, 265 Crittenden Blvd., Rochester, NY 14642, USA

Summary

Sequential, multiple assignment, randomized trial (SMART) designs have become increasingly popular in the field of precision medicine by providing a means for comparing more than two sequences of treatments tailored to the individual patient, i.e., dynamic treatment regime (DTR). The construction of evidence-based DTRs promises a replacement to *ad hoc* one-size-fits-all decisions pervasive in patient care. However, there are substantial statistical challenges in sizing SMART designs due to the correlation structure between the DTRs embedded in the design (EDTR). Since a primary goal of SMARTs is the construction of an optimal EDTR, investigators are interested in sizing SMARTs based on the ability to screen out EDTRs inferior to the optimal EDTR by a given amount which cannot be done using existing methods. In this article, we fill this gap by developing a rigorous power analysis framework that leverages the multiple comparisons with the best methodology. Our method employs Monte Carlo simulation to compute the number of individuals to enroll in an arbitrary SMART. We evaluate our method through extensive simulation studies. We illustrate our method by retrospectively computing the power in the Extending Treatment Effectiveness of Naltrexone (EXTEND) trial. An R package implementing our methodology is available to download from the Comprehensive R Archive Network.

*To whom correspondence should be addressed.

Keywords: Embedded dynamic treatment regime (EDTR); Monte Carlo; Multiple comparisons with the best; Power; Sample size; Sequential multiple assignment randomized trial (SMART).

1. Introduction

Sequential, multiple assignment, randomized trial (SMART) designs have gained considerable attention in the field of precision medicine by providing an empirically rigorous experimental approach for comparing more than two sequences of treatments tailored to the individual patient, i.e., dynamic treatment regime (DTR) (Lavori *and others*, 2000; Murphy, 2005; Lei *and others*, 2012). A DTR is a treatment algorithm implemented through a sequence of decision rules which dynamically adjusts treatments and dosages to a patient's unique changing need and circumstances (Murphy *and others*, 2001; Murphy, 2003; Robins, 2004; Nahum-Shani *and others*, 2012; Chakraborty and Moodie, 2013; Chakraborty and Murphy, 2014; Laber *and others*, 2014). SMARTs are motivated by scientific questions concerning the construction of an effective DTR. The sequential randomization in a SMART gives rise to several DTRs which are embedded in the SMART by design (EDTR). Many SMARTs are designed to compare more than two EDTRs and identify those showing greatest potential for improving a primary clinical outcome. The construction of evidence-based EDTRs promises an alternative to *ad hoc* one-size-fits-all decisions pervasive in patient care (Chakraborty, 2011).

The advent of SMART designs poses interesting statistical challenges in the planning phase of the trials. In particular, determining an appropriate sample size of individuals to enroll becomes analytically difficult due to the correlation structure between the EDTRs. Previous work includes sizing pilot SMARTs (small scale versions of a SMART) so that each sequence of treatments has a pre-specified number of individuals with some probability by the end of the trial (Almirall *and others*, 2012; Gunlicks-Stoessel *and others*, 2016; Kim *and others*, 2016). The central questions motivating this work are feasibility of the investigators to carry out the trial and acceptability of the treatments by patients. These methods do not provide a means to size SMARTs for comparing EDTRs in terms of a primary clinical outcome.

Alternatively, Crivello *and others* (2007a) proposed a new objective for SMART sample size planning. The question they address is how many individuals need to be enrolled so that the best EDTR has the largest sample estimate with a given probability (Crivello *and others*, 2007b). Such an approach based on estimation alone fails to account for the fact that some EDTRs may be statistically indistinguishable from the true best EDTR for the given data in which case they should not necessarily be excluded as suboptimal. Our approach goes one step further than Crivello's by providing a means to size SMARTs in order to construct narrow confidence intervals which not only tell which is the best EDTR, but also provide the ability to screen out inferior EDTRs. Crivello *and others* (2007a) also discussed sizing SMARTs to attain a specified power for testing hypotheses which compare only two treatments or two EDTRs as opposed to comparing all EDTRs. The work of Crivello *and others* (2007a) focused mainly on a particular common two-stage SMART design whereas our method is applicable to arbitrary SMART designs.

More recently, Ogbagaber *and others* (2016) proposed two methods for sizing a SMART. Their first approach is to choose the sample size in order to achieve a specified power for a global chi-squared test of equality of EDTR outcomes. Their second approach is to choose the sample size in order to detect pairwise differences between EDTR outcomes while adjusting for a specified number of pairwise comparisons using the Bonferroni correction. Their second approach sizes a SMART so that for each pairwise comparison, a difference can be detected with a specified probability $1 - \beta$. Our approach offers an alternative which requires a smaller sample size to achieve the same power.

One of the main goals motivating SMARTs is to identify the optimal EDTR. It follows that investigators are interested in sizing SMARTs based on the ability to screen out EDTRs which are inferior to the optimal EDTR by a clinically meaningful amount while including the best EDTR with a specified probability. In this article, we develop a rigorous power analysis framework that leverages the multiple comparisons with

the best (MCB) methodology (Hsu, 1981, 1984, 1996). The main justification for using MCB to adjust for multiple comparisons is that it involves fewer comparisons compared to other methods and thus, it yields greater power for the same sample size with all else being equal (Ertefaie *and others*, 2015).

In Section 2, we give a brief overview of SMARTs, notation, and background on estimation and MCB. In Section 3, we present our power analysis framework. In Section 4, we look at the sensitivity of the power to the covariance matrix of EDTR outcomes. In Section 5, we demonstrate the validity of our method through extensive simulation studies. In Section 6, we apply our method to retrospectively compute the power in the Extending Treatment Effectiveness of Naltrexone (EXTEND) trial. In Section 7, we discuss how to choose the covariance matrix of EDTR outcomes for sample size calculations. In Sections 8 and 9, we give concluding remarks. In the supplementary material available at *Biostatistics* online, we provide additional details about our simulation study, a comparison with the method presented in Ogbagaber *and others* (2016), and additional simulation studies for power analysis when data from a pilot SMART is available. The R package "`smartsizer`" is available to download from the Comprehensive R Archive Network.

2. PRELIMINARIES

2.1. *Sequential multiple assignment randomized trials (SMART)*

In a SMART, individuals proceed through multiple stages of randomization such that some or all individuals may be randomized more than once. Additionally, treatment assignment is often tailored to the individuals' ongoing response status (Nahum-Shani *and others*, 2012). For example, in the Extending Treatment Effectiveness of Naltrexone (EXTEND) trial (see Figure 1 for the study design and Nahum-Shani *and others*, 2017 for more details about this study), individuals were initially randomized to two different criteria of non-response: lenient or stringent. Specifically, all individuals received the same fixed dosage of naltrexone (NTX)—a medication that blocks some of the pleasurable effects resulting from alcohol consumption. After the first 2 weeks, individuals were evaluated weekly to assess response status. Individuals assigned to the lenient criterion were classified as non-responders as soon as they had five or more heavy drinking days during the first 8 weeks of the study, whereas those assigned to the stringent criterion were classified as non-responders as soon as they had two or more heavy drinking days during the first eight weeks. As soon as participants were classified as non-responders, they transitioned to the second stage where they were randomized to two subsequent rescue tactics: switch to combined behavioral intervention (CBI) or add CBI to NTX (NTX + CBI). At week 8, individuals who did not meet their assigned non-response criterion were classified as responders and re-randomized to two subsequent maintenance interventions: add telephone disease management (TDM) to NTX (NTX + TDM) or continue NTX alone. Note that the stage-2 treatment options in the SMART are tailored to the individuals' early response status. This leads to a total of eight EDTRs. For example, one of these EDTRs recommends to start the treatment with NTX and monitor drinking behaviors weekly using the lenient criterion (i.e., 5 or more heavy drinking days) to classify the individual as a non-responder. As soon as the individual is classified as a non-responder, add CBI (NTX + CBI); if at week 8 the individual is classified as a responder, add TDM (NTX + TDM). A primary goal motivating many SMARTs is the determination of optimal EDTR. For example, determining an optimal EDTR in the EXTEND may guide in evaluating a patient's initial response to NTX and in selecting the best subsequent treatment. We develop our power analysis framework with this goal in mind.

One important challenge for power analysis in SMART designs is the correlation of EDTR outcomes. The correlation arises, in part, due to overlapping interventions in distinct EDTRs and because patients' treatment histories may be consistent with more than a single EDTR. For example, patients in distinct EDTRs of the EXTEND trial all receive NTX. Also, patients who are classified as responders in stage 2

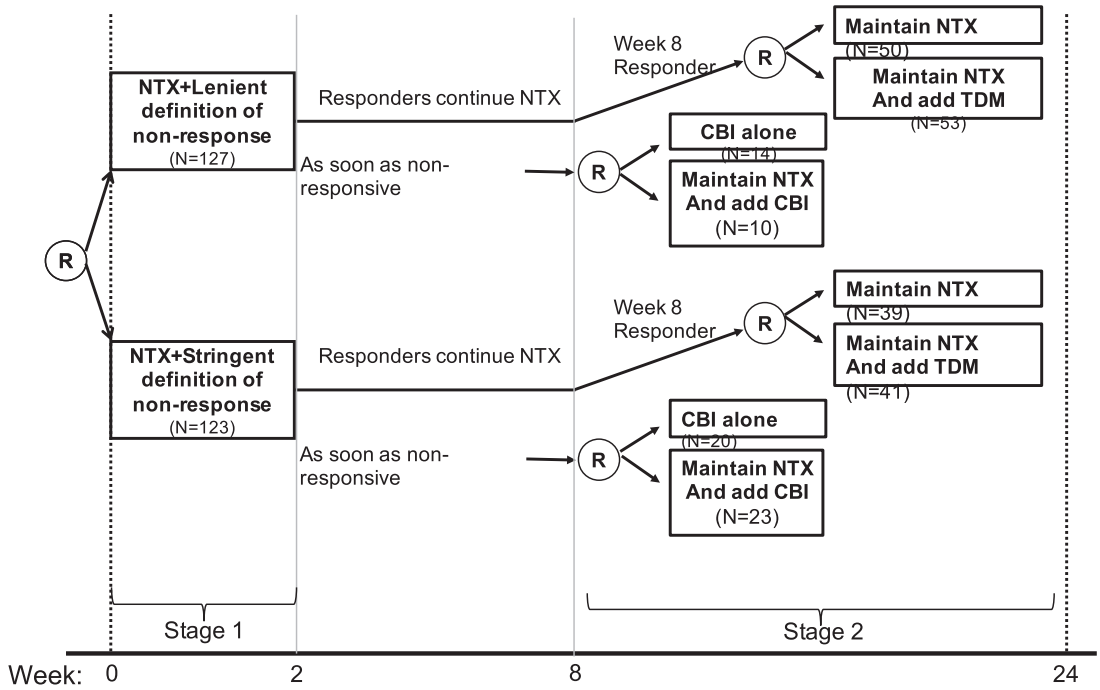

Fig. 1. This diagram shows the structure of the EXTEND trial.

and subsequently randomized to NTX will be consistent with two EDTRs: one where non-responders are offered CBI and one where non-responders are offered NTX + CBI. In Sections 4 and 5, we will discuss the dependence of power on the covariance. We provide guidelines on choosing the covariance matrix in Section 7.

2.2. *Notation*

We focus on notation for two-stage SMART designs, but the methods in this paper are applicable to an arbitrary SMART. We use the same notation as in Ertefaie *and others* (2015). Let O_j and A_j denote the observed covariates and treatment assignment, respectively, at stage j. Let $\bar{O}_j$ and $\bar{A}_j$ denote the covariate and treatment histories up to and including stage j, respectively. Let $\mathcal{T}$ the *treatment trajectory* be the vector of counterfactual treatment assignments for an individual. For example, in a two-stage SMART with stage-2 treatment tailored to response status, $\mathcal{T}$ may be of the form $\mathcal{T} = (A_1, A_2^{\mathrm{R}}, A_2^{\mathrm{NR}})$ where A_2^{R} is the stage-2 treatment assignment had the individual responded and A_2^{NR} is the stage-2 treatment assignment had the individual not responded. The reason these are counterfactual treatment assignments is that for an individual who responds to the stage-1 treatment, A_2^{NR} would be unobserved. Hence, the treatment history $\bar{A}_2$ would be (A_1, A_2) while the treatment trajectory $\mathcal{T}$ would be $(A_1, A_2^{\mathrm{R}}, A_2^{\mathrm{NR}})$ and would include the unobserved counterfactual. Let V be the embedded tailoring variable for the stage-2 treatment. For example, in EXTEND, V is the indicator of response to the stage-1 treatment. Let Y denotes the continuous observed outcome of an individual at the end of the study. Let the kth EDTR be denoted by EDTR_k. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)^t$ be the true mean outcome vector of EDTRs where N is the total number of EDTRs. Let n denote the sample size.

2.3. *Estimation*

We summarize the estimation procedures inverse probability weighting (IPW) and augmented inverse probability weighting (AIPW) introduced in Ertefaie *and others* (2015) for a two-stage SMART, but the method can be extended to arbitrary SMART designs. In order to perform estimation with IPW/AIPW, a marginal structural model (MSM) must be specified. A MSM models the response as a function of the counterfactual random treatment assignments in the treatment trajectory vector $\mathcal{T}$, while ignoring non-treatment covariates. For example, in a two-stage SMART, a MSM is: $m(\mathcal{T};\boldsymbol{\beta}) = \beta_0 + \beta_1 A_1 + \beta_2 A_2^R + \beta_3 A_2^{NR} + \beta_4 A_1 A_2^R + \beta_5 A_1 A_2^{NR}$. Subsequently, the IPW and AIPW estimators $\hat{\boldsymbol{\beta}}_{\text{IPW}}$ and $\hat{\boldsymbol{\beta}}_{\text{AIPW}}$ for $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^t$ may be obtained by solving the following respective estimating equations:

$$\mathbb{P}_n \sum_{k=1}^{N} \dot{m}(\mathcal{T};\beta) w_2(V, \bar{A}_2, k)(Y - m(\mathcal{T};\beta)) = 0 \tag{IPW}$$

$$\begin{aligned} \mathbb{P}_n \sum_{k=1}^{N} \dot{m}(\mathcal{T};\boldsymbol{\beta}) \big[& w_2(V, \bar{A}_2, k)(Y - m(\mathcal{T},\boldsymbol{\beta})) \\ & - \big(w_2(V, \bar{A}_2, k) - w_1(A_1, k)\big) \big(\mathbb{E}[Y \mid \bar{A}_2 = \text{EDTR}_k^V, \bar{O}_2, \gamma] - m(\mathcal{T};\boldsymbol{\beta})\big) \\ & - (w_1(A_1, k) - 1)(\mathbb{E}[\mathbb{E}\{Y \mid \bar{A}_2 = \text{EDTR}_k^V, \bar{O}_2, \gamma\} \mid A_1 = \text{EDTR}_{k,1}, O_1, \gamma] - m(\mathcal{T};\boldsymbol{\beta}))\big] = 0 \end{aligned} \tag{AIPW}$$

where $\mathbb{P}_n$ denotes the empirical average, $\dot{m}(\mathcal{T},\boldsymbol{\beta}) = \dfrac{\partial m}{\partial \boldsymbol{\beta}}$, $\text{EDTR}_k^v = \big(\text{EDTR}_{k,1}, \text{EDTR}_{k,2}^v\big)$ is the kth EDTR for $V = v$, $w_1(a_1, k) = \dfrac{I_{\text{EDTR}_{k,1}}(a_1)}{p(A_1 = a_1)}$ for $A_1 = a_1$, and $w_2(v, \bar{a}_2, k) = \dfrac{I_{\text{EDTR}_{k,1}}(a_1) I_{\text{EDTR}_{k,2}^v}(a_2)}{p(A_1 = a_1) p(A_2 = a_2 \mid A_1 = a_1, V = v)}$ for $V = v$ and $\bar{A}_2 = \bar{a}_2$.

Then, the EDTR outcome estimators are $\hat{\boldsymbol{\theta}}_{\text{IPW}} = \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\text{IPW}}$ and $\hat{\boldsymbol{\theta}}_{\text{AIPW}} = \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\text{AIPW}}$ where $\boldsymbol{D}$ is a $N \times p$ matrix with kth row of $\boldsymbol{D}$ corresponding to the kth EDTR contrast and p is the number of parameters in the MSM. AIPW is doubly robust in the sense that it will still provide unbiased estimates of the MSM coefficients $\boldsymbol{\beta}$ when either the conditional means or the treatment assignment probabilities are correctly specified. The following theorem from Ertefaie *and others* (2015) is included for the sake of completeness.

THEOREM 2.1 Let $\diamond$ denote IPW or AIPW. Let $\hat{\boldsymbol{\theta}}^{\diamond} = \boldsymbol{D}\hat{\boldsymbol{\beta}}^{\diamond}$. Then, under standard regulatory assumptions, $\sqrt{n}(\hat{\boldsymbol{\theta}}^{\diamond} - \boldsymbol{\theta}) \to N(\mathbf{0}, \boldsymbol{\Sigma}^{\diamond} = \boldsymbol{D}[\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}^{\diamond}\boldsymbol{\Gamma}'^{-1}]\boldsymbol{D}')$ where $\boldsymbol{\Gamma} = -\mathbb{E}[\sum_{k=1}^{N} \dot{m}(\mathcal{T};\boldsymbol{\beta})\dot{m}'(\mathcal{T};\boldsymbol{\beta})]$ and $\boldsymbol{\Lambda}^{\diamond} = \mathbb{E}[\boldsymbol{U}^{\diamond}\boldsymbol{U}'^{\diamond}]$ with

$$\begin{aligned} \boldsymbol{U}^{\text{AIPW}} = \sum_{k=1}^{N} \dot{m}(\mathcal{T};\boldsymbol{\beta}) \big[& w_2(V, A_2, k)(Y - m(\mathcal{T},\boldsymbol{\beta})) \\ & - \big(w_2(V, \bar{A}_2, k) - w_1(A_1, k)\big) \big(\mathbb{E}[Y \mid \bar{A}_2 = \text{EDTR}_k^V, \bar{O}_2, \gamma] - m(\mathcal{T};\boldsymbol{\beta})\big) \\ & - (w_1(A_1, k) - 1)(\mathbb{E}[\mathbb{E}\{Y \mid \bar{A}_2 = \text{EDTR}_k^V, \bar{O}_2, \gamma\} \mid A_1 = \text{EDTR}_{k,1}, O_1, \gamma] - m(\mathcal{T};\boldsymbol{\beta}))\big] \\ \boldsymbol{U}^{\text{IPW}} = \sum_{k=1}^{N} & \dot{m}(\mathcal{T};\boldsymbol{\beta}) w_2(V, \bar{A}_2, k)(Y - m(\mathcal{T};\boldsymbol{\beta})) \end{aligned}$$

The asymptotic variance $\boldsymbol{\Sigma}^{\diamond}$ may be estimated consistently by replacing the expectations with expectations with respect to the empirical measure and $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with its estimate $(\hat{\boldsymbol{\beta}}^{\diamond}, \hat{\boldsymbol{\gamma}})$ and may be denoted as $\hat{\boldsymbol{\Sigma}}^{\diamond} = \boldsymbol{D}[\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\Lambda}}^{\diamond}\hat{\boldsymbol{\Gamma}}^{\prime-1}]\boldsymbol{D}'$.

We will see the sample size needed in a SMART is a function of the asymptotic covariance matrix $\boldsymbol{\Sigma}$ of the EDTR outcomes $\hat{\boldsymbol{\theta}}$. This is because the amount of variation in EDTR outcomes and the correlation between EDTRs determines how easy it is to screen out inferior EDTRs. Identifying the optimal EDTRs and excluding inferior EDTRs may be viewed as the multiple testing problem. In the next section, we discuss how the MCB procedure (Hsu, 1981, 1984, 1996) can be used to address scientific questions concerning the optimal EDTR.

2.4. *Determining a set of best EDTRs using multiple comparison with the best (MCB)*

The MCB procedure permits identification of a confidence set of EDTRs which cannot be statistically distinguished from the true best EDTR for the given data while adjusting for multiple comparisons. In particular, EDTR_i is considered statistically indistinguishable from the best EDTR for the available data if and only if $\dfrac{\hat{\theta}_i - \hat{\theta}_j}{\sigma_{ij}} \geq -c_{i,1-\alpha}$ for all $j \neq i$ where $\sigma_{ij} = \sqrt{\text{Var}(\hat{\theta}_i - \hat{\theta}_j)}$ and $c_{i,1-\alpha} > 0$ is chosen so that the set of best EDTRs includes the best EDTR with at least a specified probability $1-\alpha$. Then, the set of best can be written as $\hat{\mathcal{B}} := \{\text{EDTR}_i : \hat{\theta}_i \geq \max_{j \neq i}[\hat{\theta}_j - c_{i,1-\alpha}\sigma_{ij}]\}$ where $c_{i,1-\alpha}$ depends on α and the covariance matrix $\boldsymbol{\Sigma}$. The above α represents the type I error rate for excluding the best EDTR from $\hat{\mathcal{B}}$. To control the type I error rate, it suffices to consider the situation in which the true mean outcomes are all equal. Then, a sufficient condition for the type I error rate to be at most α is to choose $c_{i,1-\alpha}$ so that the set of best includes each EDTR with probability at least $1-\alpha$: $\Pr(\text{EDTR}_i \in \hat{\mathcal{B}} \mid \theta_1 = \cdots = \theta_N) = 1-\alpha$ for all $i = 1, ..., N$. It is sufficient for $c_{i,1-\alpha}$ to satisfy:

$$\int \Pr\left(Z_j \leq z + c_{i,1-\alpha}\sigma_{ij}, \text{ for all } j = 1, ..., N\right) \mathrm{d}\phi(z) = 1-\alpha, \tag{2.1}$$

where $\phi(z)$ is the marginal cdf of Z_i and $(Z_1, ..., Z_N)^t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Observe that $c_{i,1-\alpha} > 0$ is a function of $\boldsymbol{\Sigma}$ and $\alpha \leq 0.5$, but not of the sample size n. The integral in (2.1) is analytically intractable, but the $c_{i,1-\alpha}$ may be determined using Monte Carlo methods.

It is important to note that while EDTRs included in the set of best are statistically indistinguishable for the given data, this does not mean that the EDTRs are equivalent in efficacy. This is because SMART designs may not have enough individuals in each EDTR to justify the interpretation of equivalence without an unrealistically large sample size. Our method sizes SMARTs for screening out EDTRs inferior to the best and does not size for testing equivalence.

The MCB procedure has an important advantage over other procedures which adjust for multiple comparisons: MCB provides a set with fewer EDTRs since fewer comparisons yields increased power to exclude inferior EDTRs from the set of best. Specifically, for a SMART design where N is the number of EDTRs, the MCB procedure involves only $N-1$ comparisons whereas, for example, all pairwise multiple comparison procedures entail $\binom{N}{2}$ comparisons.

In the next section, we introduce our Monte Carlo simulation based approach to compute the number of individuals to enroll in a SMART to achieve a specified power to exclude EDTRs inferior by a specified amount from the set of best.

3. METHODS

Let N be the index of the best EDTR, $\Delta_{\min} > 0$ be the minimum detectable difference between the mean best EDTR outcome and the other mean EDTR outcomes, α be the type I error rate, and $\boldsymbol{\Sigma}$ be the asymptotic covariance matrix of $\sqrt{n}\hat{\boldsymbol{\theta}}$ where n is the sample size. Furthermore, let $1-\beta$ denote the desired power to exclude EDTRs with true outcome $\Delta_{\min}$ or more away from that of the true best outcome. Let $\boldsymbol{\Delta}$ be the vector of differences between the mean best EDTR outcome and all other mean EDTR outcomes. So, $\Delta_i = \theta_N - \theta_i$ for all i. We also refer to $\boldsymbol{\Delta}$ as the vector of effect sizes and $\Delta_{\min}$ as the minimum detectable effect size, but this terminology should not be confused with a standardized effect size such as Cohen's d.

We wish to exclude all i from the set of best for which $\Delta_i \geq \Delta_{\min}$. We have that

$$\text{Power} = \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}} \left\{\hat{\theta}_i < \max_{j \neq i}[\hat{\theta}_j - c_{i,1-\alpha}\sigma_{ij}]\right\}\right) \tag{3.2}$$

However, the max operator makes (3.2) analytically and computationally complicated, so we will instead bound the RHS of the following inequality:

$$\Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}} \left\{\hat{\theta}_i < \max_{j \neq i}[\hat{\theta}_j - c_{i,1-\alpha}\sigma_{ij}]\right\}\right) \geq \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}} \left\{\hat{\theta}_i < \hat{\theta}_N - c_{i,1-\alpha}\sigma_{iN}\right\}\right) \tag{3.3}$$

Theoretically, the bound obtained using (3.3) may be conservative, but it is often beneficial to be conservative when conducting sample size calculations because of unpredictable circumstances such as loss to follow up, patient dropout, and/or highly skewed responses. Since the normal distribution is a location-scale family, the power only depends on the vector of mean differences $\boldsymbol{\Delta}$ and not on $\boldsymbol{\theta}$. Henceforth, we write $\text{Power}_{\alpha,n}(\boldsymbol{\Sigma}, \boldsymbol{\Delta}, \Delta_{\min})$ for the RHS of (3.3). It follows that

$$\begin{aligned}
\text{Power}_{\alpha,n}(\boldsymbol{\Sigma}, \boldsymbol{\Delta}, \Delta_{\min}) &= \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}} \left\{\hat{\theta}_i < \hat{\theta}_N - c_{i,1-\alpha}\sigma_{iN}\right\}\right) \\
&= \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}} \left\{\frac{\sqrt{n}(\hat{\theta}_i - (\hat{\theta}_N - \Delta_i))}{\sigma_{iN}\sqrt{n}} < -c_{i,1-\alpha} + \frac{\Delta_i\sqrt{n}}{\sigma_{iN}\sqrt{n}}\right\}\right) \\
&= \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}} \left\{W_i < -c_{i,1-\alpha} + \frac{\Delta_i\sqrt{n}}{\sigma_{iN}\sqrt{n}}\right\}\right),
\end{aligned} \tag{3.4}$$

where $\boldsymbol{W} = (W_1, ..., W_M)^t \sim N\left(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}\right)$ and

$\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}\left(\frac{\sqrt{n}(\hat{\theta}_i - (\hat{\theta}_N - \Delta_i))}{\sigma_{iN}\sqrt{n}}, \frac{\sqrt{n}(\hat{\theta}_j - (\hat{\theta}_N - \Delta_j))}{\sigma_{jN}\sqrt{n}}\right)$, and M is the number of indices $i : \Delta_i \geq \Delta_{\min}$. Note that $\boldsymbol{W}$, $c_{i,1-\alpha}$, and $\sigma_{iN}\sqrt{n} = \sqrt{\Sigma_{ii} + \Sigma_{NN} - 2\Sigma_{iN}}$ do not depend on the sample size n since $\boldsymbol{\Sigma}$ does not depend on n. If the effect sizes δ_i which are standardized by the standard deviation of the difference are specified instead of Δ_i, then Δ_i may be replaced by $\delta_i\sigma_{iN}\sqrt{n}$. Note that δ_i is not the same as Cohen's d which is standardized by the pooled standard deviation rather than the standard deviation of the difference.

Algorithm 1 SMART power computation

1. Given $\boldsymbol{\Sigma} = \text{Var}(\sqrt{n}\hat{\boldsymbol{\theta}})$, compute $c_{i,1-\alpha}$ for $i = 1, ..., N$.
2. Given $\boldsymbol{\Delta}$ and $\Delta_{\min}$, generate $\boldsymbol{W}^{(k)} = \left(W_1^{(k)}, ..., W_M^{(k)}\right)^t \sim N\left(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}\right)$, for $k = 1, ..., m$,

   where $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}\left(\frac{\sqrt{n}(\hat{\theta}_i - (\hat{\theta}_N - \Delta_i))}{\sigma_{iN}\sqrt{n}}, \frac{\sqrt{n}(\hat{\theta}_j - (\hat{\theta}_N - \Delta_j))}{\sigma_{jN}\sqrt{n}}\right)$, m is the number of Monte Carlo repetitions, and M is the number of indices $i : \Delta_i \geq \Delta_{\min}$.
3. Compute the Monte Carlo probability

   $$\text{Power}_{MC,n,\alpha}(\boldsymbol{\Sigma}, \boldsymbol{\Delta}, \Delta_{\min}) \approx \mathbb{P}_m\left[\mathcal{I}\left(\bigcap_{i:\Delta_i \geq \Delta}\left\{W_i < -c_{i,1-\alpha} + \frac{\Delta_i\sqrt{n}}{\sigma_{iN}\sqrt{n}}\right\}\right)\right]$$

   for some $m \in \mathbb{N}$ where $\mathbb{P}_m$ denotes the empirical average and $\mathcal{I}$ denotes the indicator.

It follows that the power may be computed by simulating normal random variables and substituting the probability in (3.4) with the empirical mean $\mathbb{P}_m$ of the indicator variable as is shown in Algorithm 1.

Recall the main point of this article is to assist investigators in choosing the sample size for a SMART. To this end, we will derive a method for finding the minimum n such that $\text{Power}_{\alpha,n}(\boldsymbol{\Sigma}, \boldsymbol{\Delta}, \Delta_{\min}) \geq 1 - \beta$. We proceed by rewriting the RHS of 3.3:

$$\begin{aligned}\text{Power}_{\alpha,n}(\boldsymbol{\Sigma}, \boldsymbol{\Delta}, \Delta_{\min}) &= \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}}\left\{\frac{\sqrt{n}(\hat{\theta}_i - \hat{\theta}_N + \Delta_i + c_{i,1-\alpha}\sigma_{iN})}{\Delta_i} < \sqrt{n}\right\}\right) \\ &= \Pr\left(\bigcap_{i:\Delta_i \geq \Delta_{\min}}\left\{X_i < c^*_{1-\beta}\right\}\right), \end{aligned} \tag{3.5}$$

where $\boldsymbol{X} = (X_1, ..., X_M)^t \sim N\left(\left(\frac{c_{1,1-\alpha}\sigma_{1N}\sqrt{n}}{\Delta_1}, ..., \frac{c_{M,1-\alpha}\sigma_{MN}\sqrt{n}}{\Delta_M}\right)^t, \tilde{\boldsymbol{\Sigma}}\right)$,

$\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}\left(\frac{\sqrt{n}\left(\hat{\theta}_i - \hat{\theta}_N\right)}{\Delta_i}, \frac{\sqrt{n}\left(\hat{\theta}_j - \hat{\theta}_N\right)}{\Delta_j}\right)$, M is the number of indices $i : \Delta_i \geq \Delta_{\min}$, and $c^*_{1-\beta}$ is the $1 - \beta$ equicoordinate quantile for the probability in (3.5). It follows from (3.5) that $n = (c^*_{1-\beta})^2$. Here, we write the quantile $c^*_{1-\beta}$ with an asterisk to distinguish it from the quantile $c_{i,1-\alpha}$ which controls the type I error rate α. The constant $c^*_{1-\beta}$ may be computed using Monte Carlo simulation to find the inverse of equation (3.5) after first computing the $c_{i,1-\alpha}$'s as is shown in Algorithm 2. The above procedure works because the $c_{i,1-\alpha}$'s do not change with n, so the distribution of $\boldsymbol{X}$ is constant as a function of n. Our approach for computing n is an extension of the sample size computation method in the appendix of Hsu (1996) to the SMART setting when $\boldsymbol{\Sigma}$ is known. Algorithms 1 and 2 are implemented in an R package "`smartsizer`" available at the Comprehensive R Archive Network.

In the next section, we will explore the sensitivity of the power to the covariance matrix.

4. Sensitivity of power to $\boldsymbol{\Sigma}$

We now examine how sensitive the power is to the choice of $\boldsymbol{\Sigma}$. We will address the case in which $\boldsymbol{\Sigma}$ is unknown in Section 5. For simplicity, we consider the most conservative case in which the effect sizes are

Algorithm 2 SMART sample size computation

1. Given $\boldsymbol{\Sigma} = \text{Var}(\sqrt{n}\hat{\boldsymbol{\theta}})$, compute $c_{i,1-\alpha}$ for $i = 1, ..., N$.
2. Given $\boldsymbol{\Delta}$ and $\Delta_{\min}$, generate $\boldsymbol{X}^{(k)} = \left(X_1^{(k)}, ..., X_M^{(k)}\right)^t \sim N\left(\begin{pmatrix} c_{1,1-\alpha}\sigma_{1N}\sqrt{n}/\Delta_1 \\ c_{2,1-\alpha}\sigma_{2N}\sqrt{n}/\Delta_2 \\ \vdots \\ c_{M,1-\alpha}\sigma_{MN}\sqrt{n}/\Delta_M \end{pmatrix}, \tilde{\boldsymbol{\Sigma}}\right)$, for $k = 1, ..., m$, where $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}\left(\frac{\sqrt{n}\left(\hat{\theta}_i - \hat{\theta}_N\right)}{\Delta_i}, \frac{\sqrt{n}\left(\hat{\theta}_j - \hat{\theta}_N\right)}{\Delta_j}\right)$, m is the number of Monte Carlo repetitions, and M is the number of indices $i : \Delta_i \geq \Delta_{\min}$.
3. Find the $1 - \beta$ equicoordinate quantile $c^*_{1-\beta}$ of the simulated $\boldsymbol{X}^{(k)}$ for each $k = 1, ..., m$.
4. The sample size is $n \approx \left(c^*_{1-\beta}\right)^2$.

all equal: $\Delta_i = \Delta$ for all i. In Figure 2, we evaluated the power over a grid of values for $\boldsymbol{\Sigma}$ using Equation 3.4 and Algorithm 1. These plots suggest the trend that higher correlations and lower variances tend to yield higher power which means that in order to obtain conservative power estimates, larger variances, and smaller correlations should be used. Furthermore, the correlation between best and non-best EDTRs appears to have a greater influence on power than the correlation between two inferior EDTRs as we see in the left-hand plot of Figure 2. We discuss this further in Section 7.

It is analytically difficult to prove monotonicity for a general $\boldsymbol{\Sigma}$ structure. However, it can be proven the power is a monotone non-decreasing function of the correlation and a monotone non-increasing function of the variance for an exchangeable covariance matrix. We conjecture this property is true in general for n sufficiently large. To confirm that a conservative estimate of power is obtained, one may compute the power for different values of the correlation and variance and confirm the monotone trend when using a non-exchangeable covariance matrix.

THEOREM 4.1 Let $\boldsymbol{\Sigma}$ be exchangeable: $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_N + \rho\sigma^2\left(\mathbf{1}_N\mathbf{1}_N^t - \boldsymbol{I}_N\right)$ where $\boldsymbol{I}_N = \text{diag}(1, ..., 1)$ and $\mathbf{1}_N = (1, ..., 1)^t$. The power is an increasing function of ρ and a decreasing function of σ^2.

5. SIMULATION STUDY

We have explored how the power changes in terms of a known covariance matrix. In this section, we present simulation studies for two different SMART designs in which we evaluate the assumption of a known covariance matrix. In practice, the true covariance matrix is estimated consistently by some $\hat{\boldsymbol{\Sigma}}$ (see Section 2.3 for more details). The designs and generating models are based on those discussed in Ertefaie *and others* (2015). For each SMART, we simulated 1000 datasets across a grid of sample sizes n. We computed the sets of best EDTRs using the estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$ obtained using the AIPW estimation method after correctly specifying an appropriate MSM and conditional means (see Appendix A and the Tables and Figures of the supplementary material available at *Biostatistics* online for more details).

5.1. *SMART simulation design 1*

In SMART simulation design 1, the stage-2 randomization is tailored based on response to the stage-1 treatment assignment. Individuals are considered responders if and only if $O_{21} > 0$ where O_{21} is an

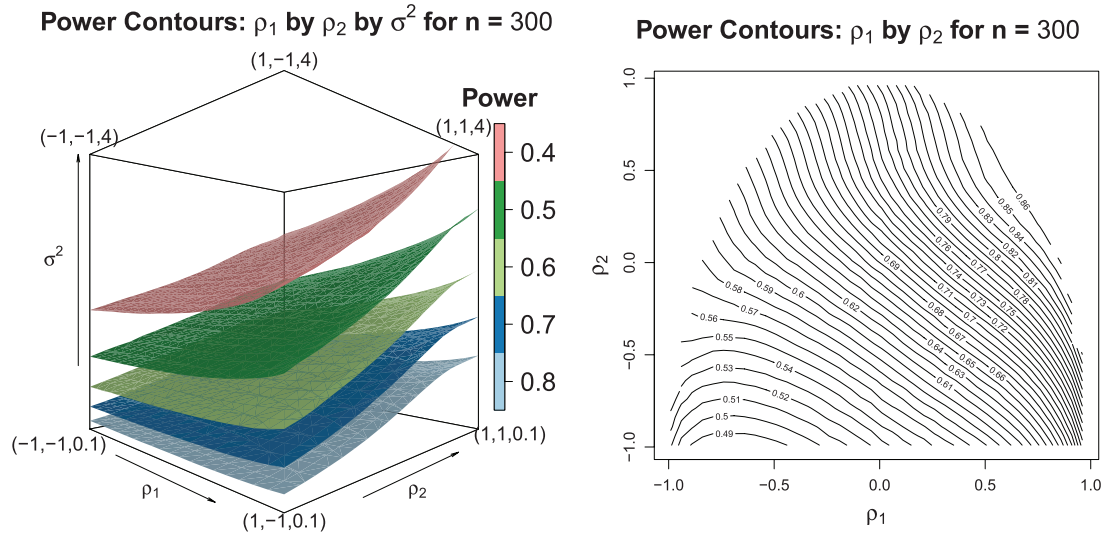

Fig. 2. The left-hand plot shows the 3D contours of the power (denoted by shade/color) as a function of ρ_1, ρ_2, σ^2 where $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_2\sigma \\ 0 & 0 & \rho_2\sigma & \sigma^2 \end{pmatrix}$ and the fourth EDTR is best. $\mathbf{\Delta} = (0.25, 0.25, 0.25, 0)$ and $\Delta_{\min} = 0.25$. The right-hand plot shows power contours over the correlations where $\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \rho_1 \\ 0 & 0 & 1 & \rho_2 \\ 0 & \rho_1 & \rho_2 & 1 \end{pmatrix}$. Note that the power appears monotone with respect to ρ_1 and ρ_2. The finger-shaped boundary is due to the feasible region of values for ρ_1 and ρ_2 such that $\mathbf{\Sigma}$ is positive definite. The sequence of contour curves in the left-hand plot in ascending order from $\sigma^2 = 0.1$ to $\sigma^2 = 4$ corresponds to the order of the power key from 0.8 to 0.4.

intermediate outcome. Non-responders to the stage-1 treatment are subsequently re-randomized to one of the two intervention options while responders continue on the initial treatment assignment. We generated 1000 data sets for each sample size $n = 100, 150, ..., 500$ according to the following scheme:

1. (a) Generate $O_{11}, O_{12} \sim N(0, 1)$ (baseline covariates)
   (b) Generate $A_1 \in \{-1, +1\}$ from a Bernoulli distribution with probability 0.5 (stage-1 treatment option indicator)
2. (a) Generate $O_{21} \mid O_{11}, A_1 \sim N(0.5O_{11} + 0.5I(A_1 = -1), 1)$ and $O_{22} \mid O_{12}, A_1 \sim N(0.5O_{12} + 0.5I(A_1 = +1), 1)$ (intermediate outcomes)
   (b) Generate $A_2^{NR} \in \{-1, +1\}$ from a Bernoulli distribution with probability 0.5 (stage-2 treatment option indicator for non-responders)
3. $Y \mid O_{11}, O_{12}, O_{21}, O_{22}, A_1, A_2^{NR} \sim$ Normal with unit variance and mean equal to

$$1 + O_{11} - O_{12} + O_{22} + O_{21} + A_1(\delta + O_{11}/2) + I(O_{21} < 0)A_2^{NR}\delta/2$$

where $\delta = 0.25$

The true $\boldsymbol{\theta}$ is $(1.802, 1.300, 1.699, 1.197)$. Note the first EDTR is the best. The vector of effect sizes $\mathbf{\Delta}$ is $(0, 0.502, 0.103, 0.605)$ and the minimum detectable effect size $\Delta_{\min}$ was set to 0.5. We computed the

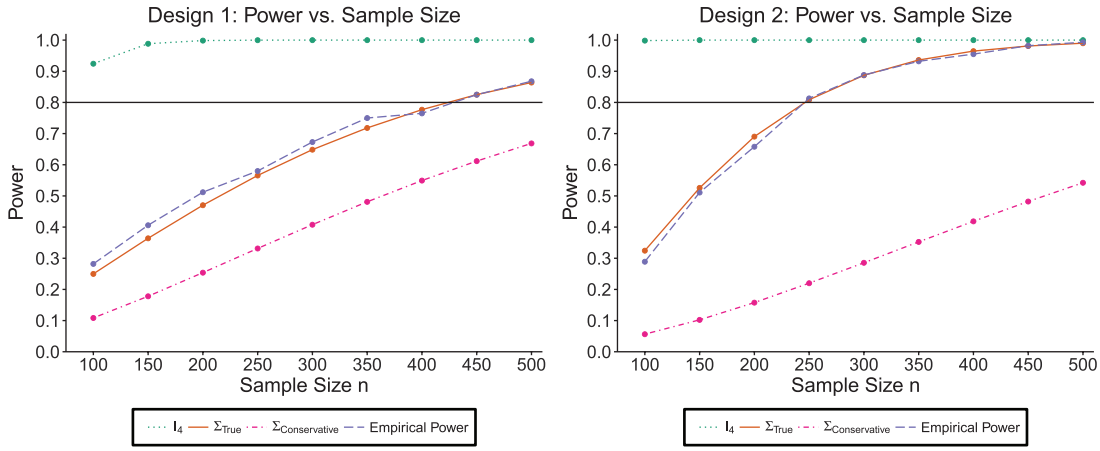

Fig. 3. The plots shows the power as a function of the sample size n with a horizontal line where the power is 80%. The plot shows the power curves for $\boldsymbol{\Sigma} = \boldsymbol{I}_4$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{True}}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{Conservative}}$, and the empirical power curve.

set of best EDTRs using MCB for each data set and sample size n. The empirical power was calculated as the fraction of datasets which excluded all EDTRs with true mean outcome Δ_{min} or more away from the best EDTR (in this case EDTRs 2 and 4), for each n. The true covariance matrix $\boldsymbol{\Sigma}_{\text{True}}$ for this SMART was estimated using AIPW by averaging the estimated covariance matrix of 1000 simulated datasets each of 10000 individuals:

$$\boldsymbol{\Sigma}_{\text{True}} = \begin{pmatrix} 10.50 & 2.52 & 9.83 & 1.85 \\ 2.52 & 7.55 & 1.81 & 6.83 \\ 9.83 & 1.81 & 10.84 & 2.81 \\ 1.85 & 6.83 & 2.81 & 7.79 \end{pmatrix} \tag{5.6}$$

5.1.1. *SMART simulation design 1: results* The simulation results are summarized in the plot on the left-hand side of Figure 3. The plot shows the sample size is sensitive to the choice of $\boldsymbol{\Sigma}$. Choosing $\boldsymbol{\Sigma} = \boldsymbol{I}_4$ will greatly underestimate the required sample size, predicting 72 individuals compared to the true 423 individuals needed to achieve 80% power. We also looked at the power for a covariance matrix $\boldsymbol{\Sigma}_{\text{Conservative}}$ which yields a conservative estimate of power. $\boldsymbol{\Sigma}_{\text{Conservative}}$ had variances chosen to be equal to the true variances and correlations chosen to be equal to zero to achieve a lower bound on the power. The minimum sample size to achieve 80% power for the conservative covariance matrix was 649.

5.2. *SMART simulation design 2*

In SMART simulation design 2, stage-2 randomization depends on both prior treatment and intermediate outcomes. In particular, individuals are randomized at stage-2 if and only if they are non-responders whose stage-1 treatment option corresponded to $A_1 = -1$ (call this condition B). Individuals are considered responders if and only if $O_{22} > 0$ where O_{22} is an intermediate outcome. We generated 1000 data sets for each sample size $n = 100, 150, ..., 500$ according to the following scheme:

1. (a) Generate $O_{11}, O_{12} \sim N(0, 1)$ (baseline covariates)
   (b) Generate $A_1 \in \{-1, +1\}$ from a Bernoulli distribution with probability 0.5 (stage-1 treatment option indicator)

Table 1. *Extend trial: EDTR outcome estimates and standard errors*

		θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
IPW	Estimate	7.56	9.53	8.05	10.02	7.71	9.68	8.19	10.17
	SD	0.76	0.81	0.71	0.83	0.74	0.80	0.69	0.82
AIPW	Estimate	7.65	9.44	7.83	9.62	8.06	9.85	8.24	10.03
	SD	0.67	0.76	0.70	0.70	0.67	0.77	0.70	0.72

2. (a) Generate $O_{21} \mid O_{11}, A_1 \sim N(0.5O_{11} + 0.5I(A_1 = -1), 1)$ and $O_{22} \mid O_{12}, A_1 \sim N(0.5O_{12} + 0.5I(A_1 = +1), 1)$ (intermediate outcomes)
   (b) Generate $A_2^B \in \{1, 2, 3, 4\}$ from a Multinomial distribution with probability 0.25 (stage-2 treatment option indicator for individuals satisfying condition B)
3. $Y \mid O_{11}, O_{12}, O_{21}, O_{22}, A_1, A_2^B \sim$ Normal with unit variance and mean equal to

$$1 + O_{11} - O_{12} + O_{21} + O_{22} + I(A_1 = -1)(\delta + O_{11}/2) + I(O_{22} < 0)I(A_1 = -1)[-\delta/4I(A_2 = 1) + \delta/2I(A_2 = 2) + 0I(A_2 = 3) + \delta/2O_{21}I(A_2 = 2)]$$

where $\delta = 2.00$

The true $\boldsymbol{\theta}$ value is $(1.500, 3.501, 3.251, 4.251, 3.501)$. The vector of effect sizes $\boldsymbol{\Delta}$ is $(2.751, 0.750, 1.000, 0.000, 0.750)$, and the minimum detectable effect size $\Delta_{\min}$ was set to 0.7. The set of best was computed for each data set. For each sample size, the empirical power is the fraction of 1000 data sets which exclude EDTRs 1, 2, 3, and 5. The true covariance matrix $\boldsymbol{\Sigma}_{\text{True}}$ for this SMART was estimated using AIPW by averaging the estimated covariance matrices of 1000 simulated datasets each of 10000 individuals:

$$\boldsymbol{\Sigma}_{\text{True}} = \begin{pmatrix} 9.50 & 1.25 & 1.19 & 1.76 & 1.24 \\ 1.25 & 17.26 & 13.55 & 13.85 & 13.25 \\ 1.19 & 13.55 & 18.32 & 13.96 & 13.55 \\ 1.76 & 13.85 & 13.96 & 23.06 & 13.85 \\ 1.24 & 13.25 & 13.55 & 13.85 & 17.27 \end{pmatrix}. \tag{5.7}$$

5.2.1. *SMART simulation design 2: results* Our simulation results are summarized in the plot on the right-hand side of Figure 3. The power plot shows the predicted power is similar to the empirical power when assuming the correct $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{True}}$. The anticipated sample size is 246 individuals for $\boldsymbol{\Sigma}_{\text{True}}$. Choosing $\boldsymbol{\Sigma} = \boldsymbol{I}_5$ yields overestimated power for each sample size, predicting 40 individuals necessary to achieve 80% power. Conversely, choosing a conservative covariance matrix $\boldsymbol{\Sigma}_{\text{Conservative}}$ underestimates the power. The $\boldsymbol{\Sigma}_{\text{Conservative}}$ is a diagonal matrix with variances set to the true variances and the correlation set to 0. The sample size for the conservative covariance matrix is 786 to achieve 80% power. The loss of power when assuming the conservative covariance matrix compared with the true covariance is due to there being a high correlation between EDTR outcomes.

6. Illustration: EXTEND retrospective power calculation

In this section, we examine how much power there was to distinguish between EDTRs $\Delta_{\min}$ away from the best in the EXTEND trial. Please see Section 2.1 for more details about EXTEND and Figure 1 for a diagram depicting the trial. The true sample size was 250. The outcome of interest was the Penn Alcohol

Craving Scale (PACS) and lower PACS were considered better outcomes. The covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the vector of EDTR outcomes $\boldsymbol{\theta}$ were estimated using both IPW and AIPW. See Table 1 for the mean EDTR outcome estimates. The covariance matrices are:

$$\hat{\boldsymbol{\Sigma}}_{\text{IPW}} = \begin{pmatrix} 145.86 & 54.88 & 77.24 & -13.74 & 101.55 & 10.57 & 32.93 & -58.05 \\ 54.88 & 163.02 & 1.13 & 109.27 & 12.66 & 120.80 & -41.09 & 67.05 \\ 77.24 & 1.13 & 125.54 & 49.42 & 33.34 & -42.77 & 81.64 & 5.53 \\ -13.74 & 109.27 & 49.42 & 172.43 & -55.55 & 67.46 & 7.62 & 130.63 \\ 101.55 & 12.66 & 33.34 & -55.55 & 138.36 & 49.47 & 70.16 & -18.73 \\ 10.57 & 120.80 & -42.77 & 67.46 & 49.47 & 159.71 & -3.87 & 106.37 \\ 32.93 & -41.09 & 81.64 & 7.62 & 70.16 & -3.87 & 118.86 & 44.84 \\ -58.05 & 67.05 & 5.53 & 130.63 & -18.73 & 106.37 & 44.84 & 169.94 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_{\text{AIPW}} = \begin{pmatrix} 113.35 & 32.52 & 82.01 & 1.19 & 103.80 & 22.97 & 72.46 & -8.36 \\ 32.52 & 143.74 & -13.93 & 97.28 & 25.91 & 137.12 & -20.55 & 90.67 \\ 82.01 & -13.93 & 123.63 & 27.69 & 72.32 & -23.63 & 113.94 & 17.99 \\ 1.19 & 97.28 & 27.69 & 123.78 & -5.58 & 90.52 & 20.92 & 117.02 \\ 103.80 & 25.91 & 72.32 & -5.58 & 112.10 & 34.21 & 80.62 & 2.73 \\ 22.97 & 137.12 & -23.63 & 90.52 & 34.21 & 148.36 & -12.39 & 101.76 \\ 72.46 & -20.55 & 113.94 & 20.92 & 80.62 & -12.39 & 122.09 & 29.08 \\ -8.36 & 90.67 & 17.99 & 117.02 & 2.73 & 101.76 & 29.08 & 128.11 \end{pmatrix}$$

The EDTR outcome vectors $\hat{\boldsymbol{\theta}}_{\text{IPW}}$ and $\hat{\boldsymbol{\theta}}_{\text{AIPW}}$ are summarized in Table 1. The vector of effect sizes for IPW is $\boldsymbol{\Delta}_{\text{IPW}} = (0.00, 1.97, 0.49, 2.46, 0.15, 2.12, 0.63, 2.61)$ and for AIPW is $\boldsymbol{\Delta}_{\text{AIPW}} = (0.00, 1.79, 0.18, 1.97, 0.41, 2.20, 0.59, 2.38)$. The set of best when performing estimation using AIPW excluded EDTRs 6 and 8. The set of best when using IPW failed to exclude any of the inferior EDTRs. In order to evaluate the power there was to exclude EDTRs 6 and 8, we set the minimum detectable effect size $\Delta_{\min}$ to 2.15.

At an α level of 0.05, the power to exclude all EDTRs inferior to the best by $\Delta_{\min}$ or more was 34% for IPW and 46% for AIPW. AIPW yields greater power than IPW because AIPW yields smaller standard errors compared with IPW (Ertefaie *and others*, 2015). Our method estimates that a total of 644 individuals would need to be enrolled to achieve a power of 80% using IPW and a total of 482 individuals would need to be enrolled when using AIPW.

In the left-hand plot of Figure 4, we computed the power over a grid of $\Delta_{\min}$ values to see how the power changes as a function of effect size. In the right-hand side of Figure 4, we show how the power changes as a function of a uniform effect size. Specifically, we assume EDTR 1 is the best and set the effect sizes of EDTRs 2, 3,..., 8 to be equal. We then vary this uniform effect size. In this case, we ignore the actual effect sizes of the true EDTR estimates $\hat{\boldsymbol{\theta}}$. We see the trend that AIPW yields greater power when compared with IPW.

7. GUIDELINES FOR CHOOSING THE COVARIANCE MATRIX $\boldsymbol{\Sigma}$

We saw in Sections 4 and 5 that the power is sensitive to $\boldsymbol{\Sigma}$. However, the dependence of power on the covariance matrix is not unique to MCB. We argue this is a necessary feature of power analysis in SMART designs because it entails comparisons of correlated EDTR outcomes. We demonstrate the sensitivity of the power to the covariance matrix when sizing a SMART to detect differences in pairwise comparisons in Appendix B of the supplementary material available at *Biostatistics* online.

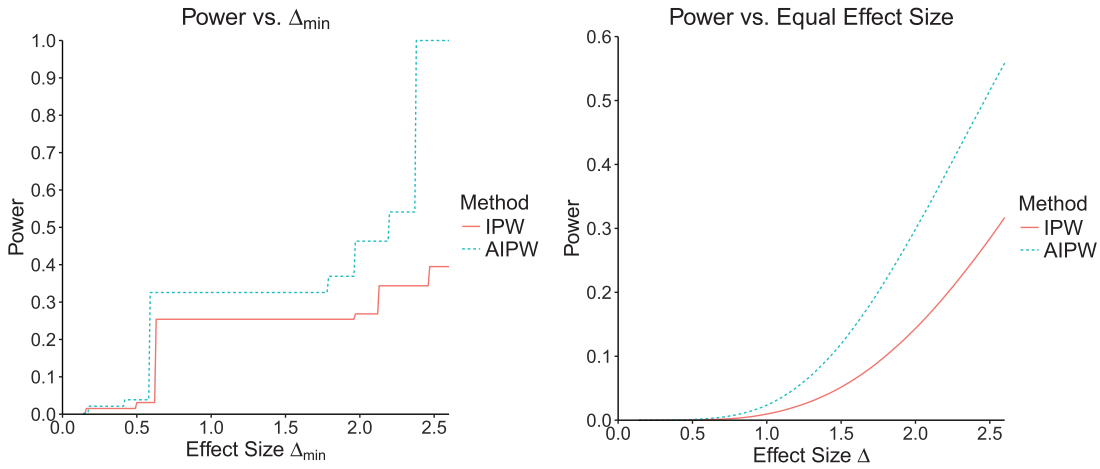

Fig. 4. The left-hand plot shows the power as a function of $\Delta_{\min}$ in the EXTEND trial when performing estimation with IPW and AIPW, respectively. The right-hand side of the plot shows the power as a function of the uniform effect size in the EXTEND trial when performing estimation with IPW and AIPW, respectively.

We focus on how to choose the covariance when the variances can be estimated (or an upper-bound given). We consider the situation in which the correlations are unknown in the absence of pilot SMART data and the situation in which pilot SMART data are available to estimate the correlations. In the absence of information about the correlation between EDTR outcomes, it is reasonable to assume all correlations are equal. Figure 2 illustrates that the correlation between the best EDTR and the non-best EDTR outcomes has a greater impact on power than the correlation between two non-best EDTR outcomes. Therefore, the correlation between the best EDTR and second best EDTR is important while the other correlations do not have as great an impact on the power, so we may make the working assumption that the correlations are equal. For example, the conservative covariance matrices in the simulation studies have equal correlations. Theorem 4.1 shows that the power for an exchangeable matrix is a monotone increasing function of the common correlation and a decreasing function of the variances. A similar monotone trend can be seen in Figure 2 for non-exchangeable covariance matrices. Specifically, larger variances and smaller correlations are more conservative. This is rather intuitive as if there is less variation, then it will be easier to distinguish between EDTRs.

When only an upper bound can be obtained for the variance of EDTR outcomes, one may assume a conservative exchangeable covariance matrix in which the diagonal elements are all equal to the upper bound on the variance and the correlation is set to the smallest plausible value. Information about the variances of outcomes for each EDTR may be obtained from prior non-SMART studies that provide the variation in outcomes for the treatments embedded in each EDTR. In this case, one may assume a matrix in which the diagonal elements equal the known variances and the correlation is set to the smallest plausible value. If a negative correlation between EDTR outcomes is implausible, a diagonal matrix may be chosen to obtain a conservative power estimate. For a covariance matrix in which the correlations are equal, the minimum negative correlation is bounded below by $-1/(N-1)$ for the covariance matrix to be positive definite (Tong, 2012).

As an alternative to sizing SMARTs based off a conservative covariance matrix, we propose conducting a pilot SMART to estimate the correlations in $\boldsymbol{\Sigma}$ in order to fine-tune power calculations. In addition to assisting in sample size calculations, pilot SMARTs are able to answer questions about the feasibility of the investigators to carry out the SMART and acceptability of the treatments by patients

(Almirall *and others*, 2012; Gunlicks-Stoessel *and others*, 2016; Kim *and others*, 2016). If estimates of the variances of each EDTR outcome are known (by choosing the largest plausible values based off knowledge of the variance of response to treatments embedded in the EDTRs), the pilot SMART may be used to estimate the correlations by first estimating the full covariance matrix using AIPW and then transforming to a correlation matrix. The covariance matrix with given variances may then be obtained by left and right multiplying the correlation matrix by the square root of the diagonal matrix whose elements consist of the variances of EDTR outcomes. We propose the following algorithm: (i) conduct a pilot SMART; (ii) bootstrap K times to obtain K estimates of the covariance matrix using an estimation procedure such as AIPW; (iii) transform the covariance matrix estimates to correlation matrices and then use the variances of EDTR outcomes obtained from prior study data to transform back to covariance matrices; (iv) compute the sample size for each bootstrapped covariance matrix and choose the maximum sample size (or 97.5th percentile, for example). When planning the pilot SMART, it is necessary to choose the pilot SMART sample size sufficiently large so that there are patients in each EDTR in order for the covariance to be estimated. It is the subject of future work to develop methods for sizing pilot SMARTs to estimate the unknown covariance matrix to a specified accuracy. For now, we refer readers to Kim *and others* (2016) and Almirall *and others* (2012) for sizing a pilot SMART. In Appendix B of the supplementary materials available at *Biostatistics* online, we demonstrate the above algorithm for two simulated pilot SMARTs with 50 individuals.

8. FINAL COMMENTS

If practitioners size a SMART using MCB, the study may be underpowered for conducting all pairwise comparisons since MCB yields greater power compared with approaches which entail a greater number of comparisons. Such confidence intervals obtained by all pairwise comparisons might not be sufficiently narrow. An important point is that MCB does not provide a p-value, so practitioners may wish to apply a method such as the global test for equality of EDTR outcomes (Ogbagaber *and others*, 2016). Sizing a SMART based off our method may overpower such an approach.

9. DISCUSSION

One important goal of SMARTs is determination of an optimal EDTR. It is hence crucial to enroll a sufficient sample size to be able to detect the best EDTR and exclude EDTRs inferior to the best one by a clinically meaningful quantity. We introduced a novel method for carrying out power analyses for SMART designs which leverages multiple comparison with the best and Monte Carlo simulation. We saw the power is sensitive to the covariance matrix and have provided guidelines for choosing it. We illustrated our method on the EXTEND SMART to see how much power there was to exclude inferior EDTRs from the set of best and the necessary sample size to achieve 80% power.

Other work has focused on estimating the optimal DTR (not embedded DTR) based on tailoring variables not embedded in the SMART. Such methods include Q-learning (Watkins, 1989; Chakraborty and Moodie, 2013; Ertefaie *and others*, 2016). These analyses are exploratory in nature and are typically not the primary goal of SMARTs. Future work will involve developing methods for sizing a SMART for such exploratory aims (Laber *and others*, 2016; Kidwell *and others*, 2018).

SOFTWARE

The R package `smartsizer` implementing Algorithms 1 and 2 is available to download at the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/smartsizer/). The R code used in this manuscript is also available to download at https://github.com/wilart/SMART-Sizer-Paper.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

Conflict of Interest: None declared.

Funding

The NIAAA (R01 AA019092, R01 AA014851, RC1 AA019092, and P01 AA016821) (in part) and also R01 DA039901 (NIH/NIDA) and K24 DA029062 (NIDA, national institute on drug abuse). The project described in this publication was partially supported by the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Almirall, D., Compton, S. N., Gunlicks-Stoessel, M., Duan, N. and Murphy, S. A. (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in Medicine* **31**, 1887–1902.

Chakraborty, B. (2011). Dynamic treatment regimes for managing chronic health conditions: a statistical perspective. *American Journal of Public Health* **101**, 40–45.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes*. New York: Springer.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and its Application* **1**, 447–464.

Crivello, A. I., Levy, J. A. and Murphy, S. A. (2007a). Evaluation of sample size formulae for developing adaptive treatment strategies using a smart design. *Technical Report* No. 07-81. University Park, PA: The Pennsylvania State University, The Methodology Center.

Crivello, A. I., Levy, J. A. and Murphy, S. A. (2007b). Statistical methodology for a smart design in the development of adaptive treatment strategies. *Technical Report* No. 07-82. University Park, PA: The Pennsylvania State University, The Methodology Center.

Ertefaie, A., Shortreed, S. and Chakraborty, B. (2016). Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Statistics in Medicine* **35**, 2221–2234.

Ertefaie, A., Wu, T., Lynch, K. G. and Nahum-Shani, I. (2015). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics* **17**, 135–148.

Gunlicks-Stoessel, M., Mufson, L., Westervelt, A., Almirall, D. and Murphy, S. (2016). A pilot smart for developing an adaptive treatment strategy for adolescent depression. *Journal of Clinical Child & Adolescent Psychology* **45**, 480–494.

Hsu, J. C. (1981). Simultaneous confidence intervals for all distances from the "best". *The Annals of Statistics* **9**, 1026–1034.

Hsu, J. C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics* **12**, 1136–1144.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. London: CRC Press.

KIDWELL, K. M., SEEWALD, N. J., TRAN, Q., KASARI, C. AND 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.

KIM, H., IONIDES, E. AND ALMIRALL, D. (2016). A sample size calculator for smart pilot studies. *SIAM Undergraduate Research Online*. DOI: http://dx.doi.org/10.1137/15S014058.

LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. AND MURPHY, S. A. (2014). Dynamic treatment regimes: technical challenges and applications. *Electronic Journal of Statistics* **8**, 1225–1272.

LABER, E. B., ZHAO, Y., REGH, T., DAVIDIAN, M., TSIATIS, A., Stanford, J. B., ZENG, D., SONG, R. AND KOSOROK, M. R. (2016). Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Statistics in Medicine* **35**, 1245–1256.

LAVORI, P. W., DAWSON, R. AND RUSH, A. J. (2000). Flexible treatment strategies in chronic disease: clinical and research implications. *Biological Psychiatry* **48**, 605–614.

LEI, H., NAHUM-SHANI, I., LYNCH, K. G., OSLIN, D. AND MURPHY, S. A. (2012). A SMART design for building individualized treatment sequences. *Annual Review of Clinical Psychology* **8**, 21–48.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 331–355.

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

MURPHY, S. A., van der LAAN, M. J., ROBINS, J. M. AND Conduct Problems Prevention Research Group. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* **96**, 1410–1423.

Nahum-SHANI, I., ERTEFAIE, A., LU, X., LYNCH, K. G., McKAY, J. R., Oslin, D. W. AND ALMIRALL, D. (2017). A smart data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction* **112**, 901–909.

Nahum-SHANI, I., QIAN, M., ALMIRALL, D., PELHAM, W. E., GNAGY, B., FABIANO, G. A., WAXMONSKY, J. G., YU, J. AND MURPHY, S. A. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods* **17**, 457–477.

OGBAGABER, S. B., KARP, J. AND WAHED, A. S. (2016). Design of sequentially randomized trials for testing adaptive treatment strategies. *Statistics in Medicine* **35**, 840–858.

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In: *Proceedings of the Second Seattle Symposium in Biostatistics.* New York, NY: Springer, pp. 189–326.

TONG, Y. L. (2012). *The Multivariate Normal Distribution*. Springer Series in Statistics. New York: Springer.

WATKINS, C. J. C. H. (1989). Learning from delayed rewards [Ph.D. Thesis]. Cambridge: King's College.

[*Received April 23, 2018; revised September 21, 2018; accepted for publication October 7, 2018*]