Design and Analytic Considerations for Sequential, Multiple-Assignment Randomized Trials with Longitudinal Outcomes

by

Nicholas J. Seewald

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2021

Doctoral Committee:

Research Associate Professor Daniel Almirall, Co-Chair
Professor Kerby Shedden, Co-Chair
Associate Professor Kelley M. Kidwell
Professor Naisyin Wang

Nicholas J. Seewald
nseewald@umich.edu
ORCID iD: 0000-0002-8367-0522

To Jeremy.

ACKNOWLEDGMENTS

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development [grant number R01HD073975]; the National Institute of Biomedical Imaging and Bioengineering [grant number U54EB020404]; the National Institute of Mental Health [grant number R03MH097954]; the National Institute on Alcohol Abuse and Alcoholism [grant numbers P01AA016821, RC1AA019092]; and the National Institute on Drug Abuse [grant numbers R01DA039901, P50DA039838]. The content of this dissertation is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

There is a long list of people who helped get this dissertation across the finish line, and I am tremendously grateful to all of them for their encouragement and support throughout this long process. I especially need to thank my advisor, Danny Almirall, who never failed to pull me out of the weeds and taught me to think more like a scientist than a statistician. Danny, your mentorship, encouragement, and vision have been invaluable.

To Kerby Shedden, my co-Chair, I am grateful for the opportunity to learn from you and the rest of the team at CSCAR, as well for your always-insightful questions. And to Naisyin Wang, a tremendous cheerleader, thank you for your wisdom and encouragement over these many years.

I would not have written this dissertation without Kelley Kidwell, who introduced me to SMARTs and DTRs and built my biostatistical intuition from the ground up. Kelley, you showed me how fun, interesting, and impactful statistics can be, and generously treated me more as a collaborator than a student. I am forever grateful for your kindness, clarity of vision, and support.

To my fiancé Jeremy, thank you for keeping me focused on what's important, and providing an island of calm when I lose sight of that. Thank you for your companionship, unwavering support, and patience. I am especially grateful for your nodding along whenever I start rambling about statistics.

To my family, especially my parents Brian and Terry, my brother Andrew, and my sister-in-law Lexi: thank you for your love and support, from first grade to twenty-fourth. You have given me so much strength and encouragement, and I'm so lucky to have you all. You're all also quite good at nodding along when I talk about statistics, but I believe Jeremy offers lessons if you're interested.

I owe a deep debt of gratitude to Lane Kelemen, for perspective and sparkling conversation.

Brenda Gunderson, Jack Miller, and Elaine Hembree taught me how to teach and trusted me to design parts of their courses, for which I will always be grateful. Susan Murphy saw something in me that I am not sure I saw in myself, and has opened so many doors for me because of that. Susan, I am humbled to call you a mentor. Liza Levina expertly guided me along my non-traditional path through the department, providing encouragement and laughing at my jokes along the way. Judy McDonald, Gina Cornacchia, and Jean McKee were so generous as to let me distract them with my chattiness for far too long. Shawna Smith and Walter Dempsey are incredible friends, mentors, and lab partners.

My years in graduate school were made immeasurably better through friendships I have grown to cherish. Among the Stats folks, I need to thank Charlotte, Laura, Zoe, Vincenzo, Yujia, Sanjana, Kali, Ajanae, and the whole of the GS-JEDI crew for your support, vulnerability, and friendship over the years. To my Biostats family, in particular, I owe my sanity and so much joy. Evan, Riley, Marco, Julie, Allison, Joe, Jed, Emma, Krithika, Anagha, Lauren B., Lauren Z., Megan, and Conrad, I feel like we've known each other for a lifetime. It is rare to meet even a handful of people in one's life who will become lifelong friends; I somehow met 14 in one semester. I remain sorry about the Kolmogorov-Smirnoff Test, but will never repent for the puns.

This is for all of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PROGRAMS

LIST OF APPENDICES

ABSTRACT

Clinicians and researchers alike are increasingly interested in how best to personalize interventions. A dynamic treatment regimen (DTR) is a sequence of pre-specified decision rules which can be used to guide the delivery of a sequence of treatments or interventions that are tailored to the changing needs of the individual. The sequential multiple-assignment randomized trial (SMART) is a research tool which allows for the construction of effective DTRs. SMARTs are multi-stage randomized trials in which some or all participants are randomized more than once, with each randomization corresponding to an open scientific question which will aid in the development of a high-quality DTR. In this dissertation, we develop a suite of tools which aid investigators in the design and analysis of SMARTs with continuous, longitudinal outcomes which are collected throughout the multiple stages of the trial.

We begin by deriving easy-to-use formulae for computing the total sample size for three common two-stage SMART designs in which the primary aim is to compare mean end-of-study outcomes for two embedded DTRs which recommend different first-stage treatments. The formulae are derived in the context of a regression model which leverages information from a longitudinal outcome collected over the entire study. We show that the sample size formula for a SMART can be written as the product of the sample size formula for a standard two-arm randomized trial, a deflation factor that accounts for the increased statistical efficiency resulting from a longitudinal analysis, and an inflation factor that accounts for the design of a SMART. The SMART design inflation factor is typically a function of the anticipated probability of response to first-stage treatment. We review modeling and estimation for DTR effect analyses using a longitudinal outcome from a SMART, as well as the estimation of standard errors. We also present estimators for the covariance matrix for a variety of common working correlation structures. Methods are motivated using the ENGAGE

study, a SMART aimed at developing a DTR for increasing motivation to attend treatments among alcohol- and cocaine-dependent patients.

Randomized trials are often constrained by limited financial resources; SMARTs are no different. The longitudinal deflation factor we develop allows for reduction in sample size requirements via both within-person correlation and the repeated measurements of the outcome over time. We provide guidance on how to balance sample size and the number of measurement occasions to minimize total cost of recruitment and measurement while achieving a target power. Finally, we introduce a procedure to generate data from a longitudinal SMART that will achieve an arbitrary desired covariance structure on potential outcomes, averaged over response status. This procedure, as well as user-friendly sample size tools which solve the cost optimization problems, are available in an R package called longsmart.

CHAPTER 1

An Introduction to Dynamic Treatment Regimens and Sequential, Multiple-Assignment, Randomized Trials

In practice, interventions often involve sequences of treatments that are adapted to an individual's changing needs. A single, fixed treatment may or may not be adequately effective for all individuals at all times; indeed, heterogeneity of treatment effects across people often exists (Longford 1999; Gail and Simon 1985). Chronic conditions which wax and wane in severity may require an intervention strategy which adjusts treatment according to changing severity over time.

Clinical practice typically involves the provision of treatment, some follow-up period, then modification of treatment to better suit the individual's needs, if necessary. However, open questions often remain as to the protocolization of this sequence. For example, "[i]gnorance of whether or how to change psychotherapies is a major and persisting gap in psychiatric knowledge" (Markowitz and Milrod 2015).

Dynamic treatment regimens (DTRs) operationalize clinical decision-making by recommending particular treatments to certain subsets of patients at specific times (Chakraborty and Moodie 2013). DTRs are sequences of pre-specified decision rules leading to courses of treatment which adapt to a patient's changing needs (Kosorok and Moodie 2015). Consider the following example DTR which was designed to increase engagement with an intensive outpatient rehabilitation program (IOP) for patients with alcohol and/or cocaine dependence: "Within a week of the participant becoming non-engaged in the IOP, provide two phone-based sessions focused on helping the patient re-engage in the IOP. At week 8, look back at the participant's engagement pattern over the past eight weeks. If the participant continued to not engage (i.e., did not respond to the intervention),

provide a second pair of phone-based sessions, this time focused on facilitating personal choice (i.e., highlighting various treatment options the patient can choose from in addition to IOP). Otherwise, for those who did engage with the intervention, provide no further contact" (McKay et al. 2015). Notice that the DTR recommends intervention strategies for both engaged and non-engaged participants at week 8. Alternative names for DTRs include adaptive treatment strategies (Wallace and Moodie 2014; Ogbagaber, Karp, and Wahed 2016) and adaptive interventions (Almirall et al. 2014; Nahum-Shani et al. 2012a), among others.

Scientists often have questions about how best to sequence and individualize interventions in the context of a DTR. Sequential, multiple-assignment, randomized trials (SMARTs) are one type of randomized trial design that can be used to answer questions at multiple stages of the development of high-quality DTRs (Lavori and Dawson 2000, 2004; Murphy 2005). The characteristic feature of a SMART is that some or all participants are randomized more than once, often based on previously-observed covariates. Each randomization corresponds to a critical question regarding the development of a high-quality DTR, typically related to the type, timing, or intensity of treatment. SMARTs have been employed in a variety of fields, including oncology (Auyeung et al. 2009; Kidwell 2014; Thall 2015), surgery (Diegidio et al. 2017; Hibbard et al. 2018), substance abuse (Murphy et al. 2007), and autism (Kasari et al. 2014).

In this chapter, we introduce and motivate the study of DTRs and SMARTs. We begin by formally defining DTRs, then discuss how they can be studied using a SMART. We present a variety of SMART designs, and discuss motivations for each.

1.1 Dynamic Treatment Regimens (DTRs)

A DTR is a sequence of functions ("decision rules"), each of which takes as inputs a person's history up to the time of the current decision (including baseline covariates, adherence, responses to previous treatments, etc.) and outputs a recommendation for the next treatment (Murphy 2005). Formally, suppose we wish to construct a DTR which recommends M treatments, $a_1, \ldots, a_M$, to

each individual. After the jth treatment, the DTR will have recommended the sequence $\bar{a}_j = \{a_1, a_2, \ldots, a_j\}$. Let $S_j(\bar{a}_{j-1})$ denote information collected in the period after providing treatment a_{j-1} until immediately prior to the provision of treatment a_j. This includes any outcomes and covariates which may be observed, as well as previous treatment assignments. S_1 contains pre-treatment information. Note that $S_j(\bar{a}_{j-1})$ is indexed by the history of treatment assignments made up to, but not including, the time at which a_j is assigned, reflecting the fact that different values of the covariates may be observed depending on the assigned sequence of treatments. We use $\bar{S}_j(\bar{a}_{j-1}) = \{S_1, S_2(a_1), \ldots, S_{j-1}(\bar{a}_{j-2}), S_j(\bar{a}_{j-1})\}$ to represent the "history" until the time at which a_j is provided.

A decision rule φ_j is a function of $\bar{S}_j(\bar{a}_{j-1})$ which outputs a recommendation for subsequent treatment a_j. An M-stage dynamic treatment regimen is a sequence of M decision rules $\{\varphi_1, \ldots, \varphi_M\}$ (Murphy 2005). The times in a patient's care when a decision is made is called a decision point. These can occur at scheduled intervals, after a specific number of clinic visits, or be event-based, such as the point at which a patient fails to respond or adhere to a treatment. The timing of decision points should be based on scientific or practical considerations which inform when treatment may need to be modified.

The information $S_j(\bar{a}_{j-1})$ often contains covariates which inform the recommendation to subsequent treatment a_j. These covariates are called "tailoring variables." These could be static characteristics (e.g., demographic factors, history of prior treatment, etc.) or time-varying participant information, such as disease severity, which may vary based on $\bar{a}_{j-1}$.

Consider the example two-stage DTR above. The clinician experiences two decision points: the first is at treatment initiation and the second occurs after eight weeks, at which point the individual is identified as a "responder" or "non-responder" based on their engagement. There is a single treatment option at the first decision point, and two options at the second decision point (motivational interviewing focused on facilitating personal choice, or no further contact). The recommended first-stage treatment for all patients is a phone-based session with a focus on re-engagement with the IOP; $\phi_1(S_1)$ is constant in S_1. At week 8, each participant's history of engagement is assessed, and an

appropriate second-stage treatment is recommended. For participants who have shown a pattern of continued non-engagement (non-responders), the recommended second-stage treatment is a second phone-based session focusing on personal choice. For all other participants (responders), the DTR recommends no further contact. Here, the tailoring variable contained in $S_2(a_1)$ is an indicator as to whether or nor the participant demonstrated a pattern of continued non-engagement prior to week 8.

In this dissertation, we will consider only two-stage DTRs. Further, we focus on binary tailoring variables, which we will abbreviate to "response" or "non-response". Since a DTR recommends treatments to both responders and non-responders, we can denote a DTR with a triple of the form (a_1, a_{2R}, a_{2NR}), where a_1 is an indicator for the recommended first-stage treatment, a_{2R} an indicator for the second-stage treatment recommended for responders, and a_{2NR} the second-stage treatment recommended for non-responders.

Researchers interested in developing high-quality DTRs often have unanswered questions that cannot necessarily be answered based on existing literature, or expert clinical opinion. These questions typically concern the relative effectiveness of different DTRs, the relative effectiveness of different DTR components at specific stages, how the intervention components at different stages work with (or against) each other, and questions related to how best to tailor treatment at different stages of intervention. Common questions are about which treatment option the DTR should begin with, how to modify the initial treatment for non-responders, how to best define or monitor individuals for response/non-response, and the timing of decision points and thus interventions. These questions can be addressed using a sequential, multiple-assignment randomized trial, or SMART.

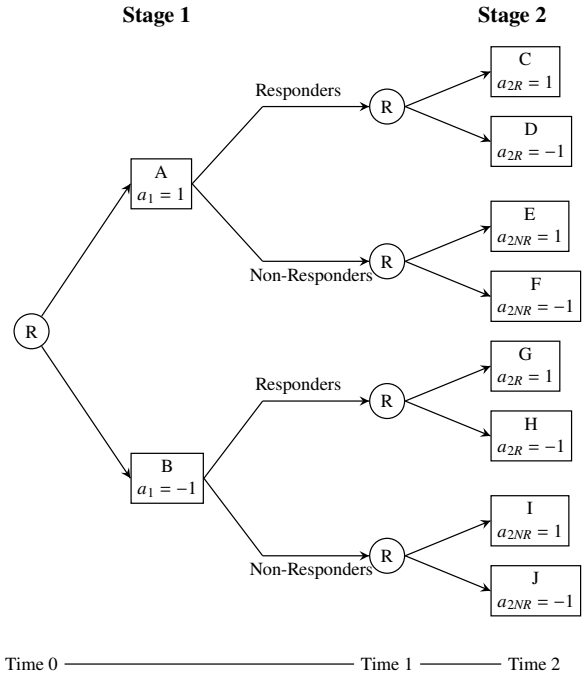
1.2 Sequential Multiple-Assignment Randomized Trials (SMARTs)

A SMART is a type of randomized trial in which some or all participants are randomized more than once, the goal of which is typically to develop a high-quality DTR. In a SMART, all participants move through multiple stages of treatment. At each stage, participants may be randomized to a set of feasible treatment options. These randomizations correspond to scientific questions about the development of an effective DTR. The treatment options to which a participant is randomized at each stage may depend on participant characteristics via a tailoring variable or prior treatment. We consider two-stage SMARTs in which the primary outcome is continuous and repeatedly measured in participants over the course of the study.

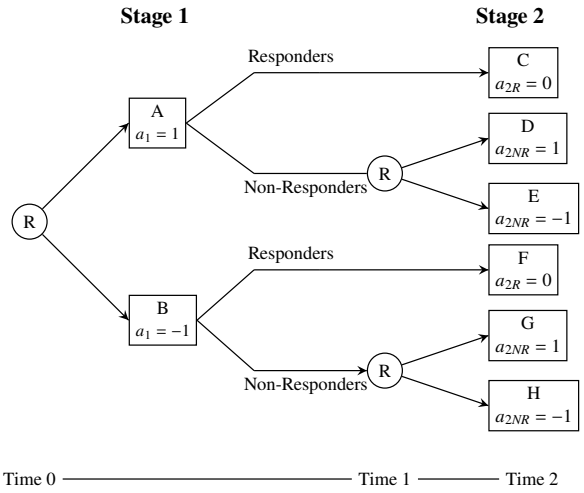
Most SMARTs contain "embedded" DTRs; that is, by design, participants in a SMART may be assigned to treatments which are consistent with recommendations made by one or more DTRs. Often, subsequent randomizations in a SMART are restricted to particular groups of participants based on an embedded tailoring variable, which is chosen based on scientific, ethical, or practical considerations. For example, in oncology, it would be unethical to randomize patients who do not respond to a high dose of chemotherapy to an intervention which would increase the dose beyond a known toxicity threshold. Instead, investigators may choose to not re-randomize these individuals to a higher dose.

We consider SMARTs in which each randomization is between two possible interventions, and where the tailoring variable is binary. In Figure 1.1, we introduce three common two-stage SMART designs which vary in the subsets of participants who are re-randomized after the first stage. To our knowledge, these designs are representative of the majority of the SMARTs in the field to date.

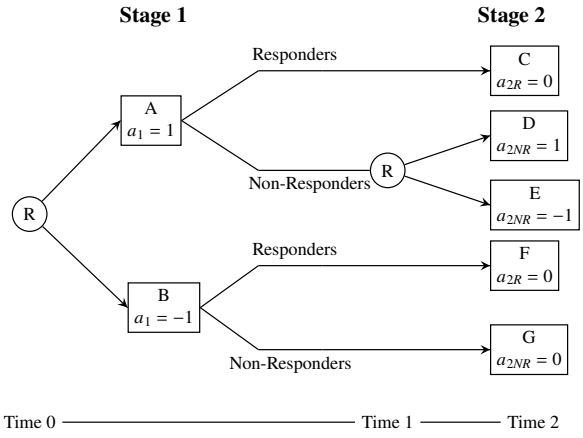
In design I, all participants are re-randomized. There are eight DTRs embedded in this design: for example, the DTR which starts by recommending A, then recommends C for responders and F for non-responders. Using the notation above and the indices in figure 1.1, this DTR would



(I) All participants are re-randomized, regardless of response status.



(II) The second randomization is restricted to only non-responders.



(III) The second randomization is restricted to only non-responders to treatment A.

Figure 1.1: Three commonly-used two-stage SMART designs. Each design varies in choice of which subsets of participants are re-randomized. Circled R indicates randomization, capital letters indicate (potentially non-unique) treatments, and $a_{\_}$ provides a coding system used to index embedded DTRs.

be written (1, 1, -1). SMARTs of this form have been run in the fields of drug dependence (Oslin 2005; Fitzsimons et al. 2015), smoking cessation (Fu et al. 2017), and childhood depression (Eckshtain 2013), among others. Often, motivation for re-randomizing all participants arises out of open scientific questions regarding both a maintenance therapy for responders and a "rescue" intervention for non-responders.

SMARTs using design II restrict the second randomization to only non-responders; that is, only participants who have a certain value of the tailoring variable (here, "non-response") are re-randomized. This might be motivated by an open question regarding second-stage treatment only among non-responders (i.e., the follow-up intervention for responders may be well-established). Design II is perhaps the most common SMART design, and is often referred to as the "prototypical" SMART (NeCamp, Kilbourne, and Almirall 2017). It has been utilized in the study of ADHD (Pelham et al. 2016), adolescent marijuana use (Budney 2014), alcohol and cocaine dependence (McKay et al. 2015), and more. There are four embedded DTRs in this design. Because responders are not re-randomized, a_{2R} is set to zero for all embedded DTRs.

In design III, re-randomization is restricted to only non-responders who receive a particular first-stage treatment. This design might be used when one of the first-stage interventions involves a top-of-the-line treatment that, for practical reasons, cannot be intensified in the second stage. For the individuals randomized to this first-stage treatment, there may only be one option for subsequent intervention, regardless of their value of the tailoring variable. SMARTs of this type have been used to investigate cognition in children with autism spectrum disorder (Kasari et al. 2014; Almirall et al. 2016) and implementation of a re-engagement program for patients with mental illness (Kilbourne et al. 2013). There are three DTRs embedded in this design. Note that, as in design II, responders are not re-randomized, so a_{2R} is set to zero for all embedded DTRs. Furthermore, a_{2NR} is set to zero when $a_1 = -1$, as non-responders to treatment B are not re-randomized.

As stated before, the goal of SMART designs is to aid the development of DTRs. Data collected in a SMART can be used to answer questions concerning which intervention option to provide at critical decision points during care. Common primary aims for SMARTs include a

comparison of first-stage treatment options averaged over subsequent interventions, or a comparison of second-stage intervention options among responders, averaged over the first-stage definition of non-response; similarly for non-responders. Questions can also focus on comparisons of the DTRs embedded in a SMART. An example would be to compare the DTR shown in Figure 1 (embedded DTR 2) to embedded DTR 5 based on proportion of days abstinent from alcohol at the end of the study. This type of comparison may be used to investigate the difference between, say, the most and least intensive DTRs, or the most and least expensive.

Each of the SMART designs discussed above is motivated by a different set of scientific questions at multiple stages of a DTR. Data collected in a SMART can be used to answer questions concerning which intervention option to provide at critical decision points during care. Questions can also focus on comparisons of the DTRs embedded in a SMART. Because each randomization in a SMART corresponds to an open question about subsequent treatment recommendations, and the defining characteristic of a SMART is that some or all participants are randomized more than once, questions that do not involve multiple stages of treatment do not, by themselves, motivate a SMART. Almirall et al. (2018) describe several "singly-randomized" alternatives to SMARTs in the context of research on DTRs.

To illustrate ideas, we use ENGAGE, illustrated in figure 1.2. ENGAGE is a SMART designed to study the effects of offering cocaine- and/or alcohol-dependent patients who did not engage in an IOP phone-based sessions either geared toward re-engaging them in an IOP or offering a choice of treatment options (McKay et al. 2015). The study recruited 500 cocaine- and/or alcohol-dependent adults who were enrolled in an IOP and failed to attend two or more sessions in the first two weeks. ENGAGE is modeled on design II. In the context of figure 1.1, treatment A was two phone-based motivational interviews focused on reengaging the participant with the IOP ("MI-IOP"); treatment B was two phone-based motivational interviews geared towards helping the participant choose and engage with an intervention of their choice ("MI-PC"). Participants who exhibited a pattern of continued non-engagement after eight weeks were considered non-responders, and re-randomized to receive either MI-PC (treatments D and G) or no further contact (treatments E

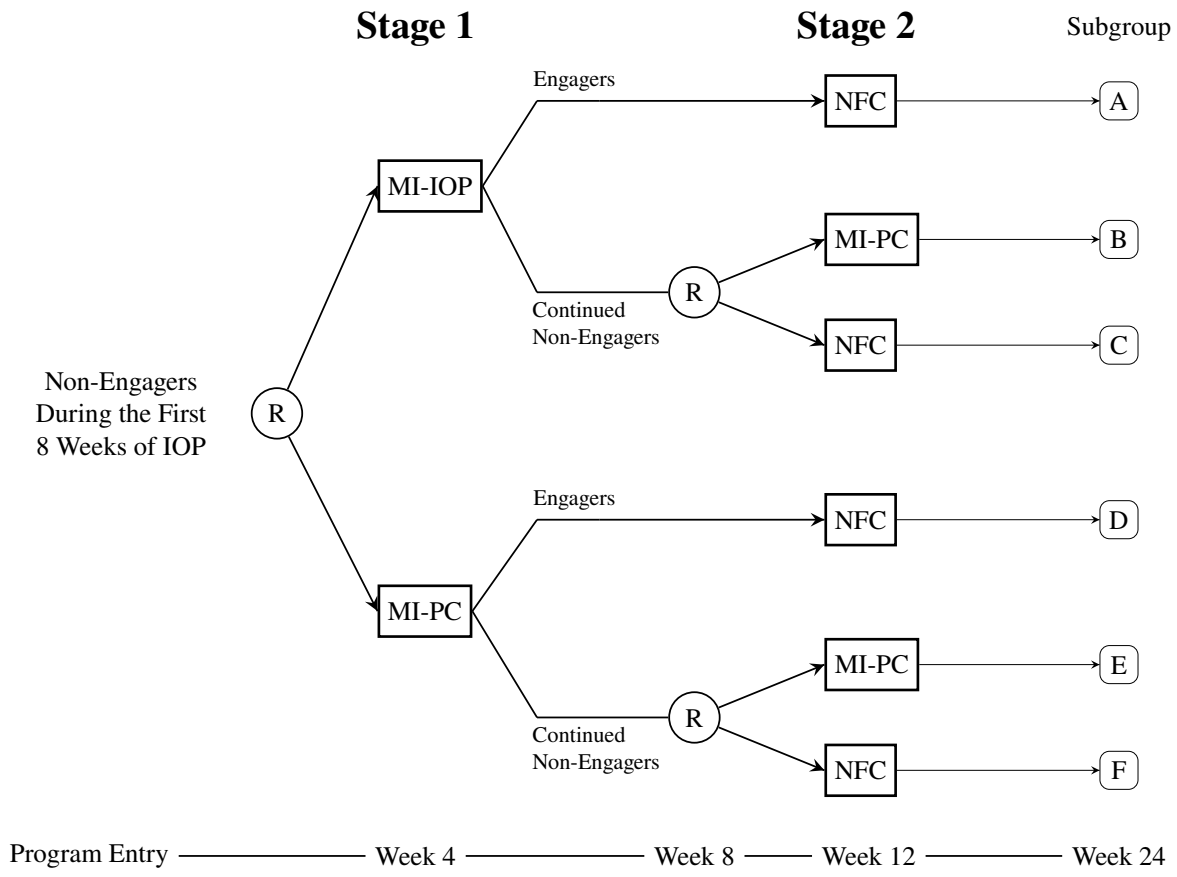


Figure 1.2: The ENGAGE SMART. Circled R indicates randomization with probability 0.5. "MI-IOP" refers to motivational interviewing with focus on intensive outpatient program; "MI-PC" to motivational interviewing with focus on patient choice; "NFC" to no further contact. Subgroups identify particular treatment paths which participants may follow.

and H). Responders were provided no further contact (treatments C and F). Following the coding in figure 1.1, the example DTR on page 1 is labeled (1, 0, 1). The other embedded DTRs are given in table 1.1.

An important continuous outcome in ENGAGE is "treatment readiness". This is a measure of a patient's willingness and ability to commit to active participation in a substance abuse treatment program. The score ranges from 8 to 40 and is coded so that higher scores indicate greater treatment readiness. Measurements are taken at baseline, and 4, 8, 12, and 24 weeks after program entry.

Table 1.1: Embedded DTRs in the ENGAGE SMART. ENGAGE is depicted in figure 1.2. "MI-IOP" refers to motivational interviewing with focus on intensive outpatient program; "MI-PC" to motivational interviewing with focus on patient choice; "NFC" to no further contact. "Subgroups" are in reference to figure 1.2. The final column writes each embedded DTR as a triple (a_1, a_{2R}, a_{2NR}).

DTR	Stage 1 Treatment	Stage 2 Treatment for Responders	Stage-2 Treatment for Non-Responders	Subgroups consistent with DTR	Coding
1	MI-IOP	NFC	MI-PC	A, B	$(1, 0, 1)$
2	MI-IOP	NFC	NFC	A, C	$(1, 0, -1)$
3	MI-PC	NFC	MI-PC	D, E	$(-1, 0, 1)$
4	MI-PC	NFC	NFC	D, F	$(-1, 0, -1)$

1.2.1 Considerations for Designing SMARTs

ENGAGE contains four embedded DTRs. Notice that responders to a particular first-stage intervention are consistent with both embedded DTRs which recommend that intervention. ENGAGE is conceptually similar to a 2×2 (fractional) factorial design (Murphy and Almirall 2009; Collins, Nahum-Shani, and Almirall 2014; Vock and Almirall 2018). The first factor is MI-IOP vs. MI-PC; the second factor is restricted to non-responders and is MI-PC vs. no further contact.

Two key differences from factorial designs are the sequential nature of treatment delivery in a SMART, as well as the possible restriction of certain treatment options to participants based on their response status. Scientific questions which motivate a SMART are asked in the context of a sequence of treatments which are delivered at multiple points in time: this is not typically captured by a standard factorial design. Additionally, SMARTs which contain an embedded tailoring variable usually offer different sets of treatment options to responders and non-responders. Similarly, first-stage treatment assignment may determine whether individuals are re-randomized, as in design III SMARTs. These SMARTs are therefore not fully crossed designs (Nahum-Shani et al. 2012a).

SMARTs often include standard-of-care control groups. Most commonly, this is done by embedding a standard-of-care intervention as one of the DTRs. For instance, in design II, one of the embedded DTRs may be a DTR that is commonly used in practice or could recommend standard-of-care throughout. This type of SMART would allow for comparisons of the other embedded DTRs against this standard-of-care DTR.

An important consideration in the design of a SMART is the choice of embedded tailoring

variable, if included. Embedding a tailoring variable into the trial also embeds it into any DTRs the trial is able to study, so its inclusion should be well-justified based on scientific, ethical, or practical considerations. The tailoring variable is a component of the DTR. As such, its operating characteristics are part of the intervention as well as the trial. Therefore, tailoring variables should be relatively easily measured in a clinical setting and reliably identify responders and non-responders. A variable which may "misclassify" individuals is not a good choice of tailoring variable, as it may make assignment to subsequent treatment unsystematic. This is an issue that should be anticipated and designed around, rather than corrected *post hoc*.

In a SMART, the same cohort of individuals participates in all stages of treatment and a single study consent process is used for all these individuals (prior to the first stage randomization). SMARTs should not employ multiple consents (e.g., one at each randomization point); doing so could severely limit the ability to make inferences about the relative effects of the DTRs embedded in a SMART. Rather, the single consent process should inform participants of all possible treatment sequences to which they may be assigned during the study. Because the goal of a SMART is to develop a high-quality DTR, participants in the trial should experience the DTR as close to a real-world implementation as possible; a re-consent process would detract from this goal. Should they wish, investigators could randomize participants to DTRs at the start of the trial, though this should be carefully blinded to avoid expectancy effects: participants should not have knowledge of their future treatment assignments.

Importantly, SMARTs are typically not adaptive trial designs despite using similar terminology (e.g., adaptive interventions, etc.) (Meurer, Lewis, and Berry 2012). An adaptive trial is a multi-stage study in which ongoing patient information is used to modify the *design* of the trial as data are collected (Dragalin 2006). By contrast, SMARTs are usually fixed designs in which the goal is to identify a sequence of treatments which adapt to the participant's changing needs. Recently, statisticians have begun to develop SMARTs with adaptive randomization (Cheung, Chakraborty, and Davidson 2015).

CHAPTER 2

Estimation and Sample Size for SMARTs with Continuous Longitudinal Outcomes

This work originally appeared as Seewald et al. (2020).

The comparison of two embedded DTRs which recommend different first-stage treatments is a common primary aim for a SMART (Nahum-Shani et al. 2012a). There exist data analytic methods for addressing this aim when the outcome is continuous (Nahum-Shani et al. 2012a), survival (Li and Murphy 2011), binary (Kidwell et al. 2018), cluster-level (NeCamp, Kilbourne, and Almirall 2017), and longitudinal (Lu et al. 2016; Li 2017). A key step in designing a SMART, as with any randomized trial, is determining the sample size needed to be able detect a desired effect with given power. However, there is no existing method for determining sample size for such a comparison when the outcome is continuous and longitudinal.

Often, SMARTs involve repeated measurements of a continuous outcome spaced throughout the trial. This might involve a measurement at baseline, one or more measurements in the first stage of the trial (before assessment of the tailoring variable and subsequent re-randomization), and one or more measurements in the second stage. In this chapter, we begin by reviewing the work of Lu et al. (2016), which developed models and an estimation procedure for SMARTs with longitudinal outcomes. We then extend that work by offering more detailed guidance on the estimation of model parameters used in computing quantities of interest on which to compare two embedded DTRs. Finally, we present sample size formulae for SMARTs in which the primary aim is to compare the mean end-of-study outcomes for two embedded DTRs which recommend different first-stage treatments and which satisfy certain design constraints.

Our primary contribution is tractable sample size formulae for SMARTs with a continuous longitudinal outcome in which the primary aim is an end-of-study comparison of two DTRs which recommend different first-stage treatments. Additionally, we present estimators for parameters in the working covariance matrix used in the analysis methods developed by Lu et al. (2016).

Sections 2.1 and 2.2 review the modeling and estimation procedures introduced by Lu et al. (2016) and extends it by developing estimators for various working covariance structures in section 2.2.3. In section 2.3, we develop and present sample size formulae for SMARTs in which the primary aim is a comparison of two embedded DTRs which recommend different first-stage treatments using a continuous longitudinal outcome. The sample size formulae are evaluated via simulation in section 2.4.

2.1 Marginal Mean Model

Consider a SMART design with embedded DTRs labeled by (a_1, a_{2R}, a_{2NR}). Suppose we have a longitudinal outcome $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,T})$, $i = 1, \ldots, n$, observed such that $Y_{i,j}$ is measured for each of n participants at each of $j = 1, \ldots, T$ measurement occasions $\left\{t_j : j = 1, \ldots, T; t_1 < \ldots < t_T\right\}$. We do not require that these measurements be equally-spaced, though they must be common to all participants in the study. In ENGAGE, for example, $T = 5$, $\left\{t_j\right\} = \{0, 4, 8, 12, 24\}$. There are $T_1 = 3$ measurements in stage 1 (note that this includes baseline) and $T_2 = 2$ in stage 2. Define $t^* = t_{T_1}$ to be the time of the measurement taken immediately before the assessment of response status and second randomization. In ENGAGE, $t^* = t_3 = 8$. Let $\boldsymbol{X}_i$ be a vector of mean-centered baseline covariates, such as age at baseline, sex, etc., for the ith individual.

We are interested in $\mathrm{E}[Y_j^{(a_1,a_{2R},a_{2NR})} \mid \boldsymbol{X}]$, the marginal mean outcome at time t_j under DTR (a_1, a_{2R}, a_{2NR}) conditional on $\boldsymbol{X}$. This is the mean outcome at the jth measurement occasion had all individuals with characteristics $\boldsymbol{X}$ been offered DTR (a_1, a_{2R}, a_{2NR}). Recall that a DTR recommends treatments for both responders and non-responders; therefore, $\mathrm{E}[Y_t^{(a_1,a_{2R},a_{2NR})} \mid \boldsymbol{X}]$ is marginal over response status. Note that $Y_{i,j}^{(a_1,a_{2R},a_{2NR})}$ is a potential outcome, the value of $Y_{i,j}$ that

would be observed at time t_j had participant i been treated according to the DTR (a_1, a_{2R}, a_{2NR}).

We impose a modeling assumption on $\mathrm{E}[Y_j^{(a_1,a_{2R},a_{2NR})} \mid \boldsymbol{X}]$; namely, that $\mathrm{E}[Y_j^{(a_1,a_{2R},a_{2NR})} \mid \boldsymbol{X}] = \mu_j^{(a_1,a_{2R},a_{2NR})}(\boldsymbol{X};\boldsymbol{\theta})$, where $\mu_j^{(a_1,a_{2R},a_{2NR})}(\boldsymbol{X};\boldsymbol{\theta})$ is a marginal structural mean model with unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\eta}^\top, \boldsymbol{\beta}^\top)^\top$. We use $\boldsymbol{\eta}$ to represent a column vector of parameters indexing baseline covariates, and $\boldsymbol{\beta}$ is a column vector of coefficients on terms involving treatment effects; we discuss in more detail below. As noted by Lu et al. (2016), the sequential nature of treatment delivery in SMARTs may suggest constraints on the form of $\mu_j^{(a_1,a_{2R},a_{2NR})}(\boldsymbol{X};\boldsymbol{\theta})$ which depend, in part, on the design of the SMART. For instance, in ENGAGE, at time $t = 0$, no treatments have been assigned, so all DTRs share a common mean. At times $t = 4$ and $t = 8$, the four embedded DTRs differ only by recommended first-stage treatment; thus there are two means of $Y_j^{(a_1,a_{2R},a_{2NR})}$ at each measurement occasion $2 < j \leq T_1$. Finally, for times $t > t^* = 8$ ($T_1 < j \leq T$), each DTR has a different mean $Y_j^{(a_1,a_{2R},a_{2NR})}$.

An example marginal structural mean model for ENGAGE (and, more generally, design II), assuming one baseline covariate X_1, is

$$\begin{aligned}\mu_j^{(a_1,a_{2R},a_{2NR})}(X_1;\boldsymbol{\theta}) = \eta_1 X_1 + \beta_0 + \mathbb{1}_{\{t_j \leq t^*\}}\left(\beta_1 t_j + \beta_2 a_1 t_j\right) \\ + \mathbb{1}_{\{t_j > t^*\}}\left(\beta_1 t^* + \beta_2 t^* a_1 + \beta_3(t_j - t^*) + \beta_4(t_j - t^*)a_1\right. \\ \left. + \beta_5(t_j - t^*)a_{2NR} + \beta_6(t_j - t^*)a_1 a_{2NR}\right),\end{aligned} \tag{2.1}$$

where $\mathbb{1}_{\{E\}}$ is the indicator function for the event E. Similarly, for design I, a saturated marginal structural mean model is of the form

$$\begin{aligned}\mu_j^{(a_1,a_{2R},a_{2NR})}(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{\eta}^\top \boldsymbol{X} + \beta_0 + \mathbb{1}_{\{t_j \leq t^*\}}\left(\beta_1 t_j + \beta_2 t_j a_1\right) \\ + \mathbb{1}_{\{t_j > t^*\}}\left(\beta_1 t^* + \beta_2 t^* a_1 + \beta_3(t_j - t^*) + \beta_4(t_j - t^*)a_1 + \beta_5(t_j - t^*)a_{2R}\right. \\ \left. + \beta_6(t_j - t^*)a_{2NR} + \beta_7(t_j - t^*)a_1 a_{2R} + \beta_8(t_j - t^*)a_1 a_{2NR}\right);\end{aligned} \tag{2.2}$$

for design III, a saturated marginal structural mean model is

$$\mu_j^{(a_1,a_{2R},a_{2NR})}(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{\eta}^\top \boldsymbol{X} + \beta_0 + \mathbb{1}_{\{t_j \le t^*\}}\left(\beta_1 t_j + \beta_2 a_1 t_j\right) + \mathbb{1}_{\{t_j > t^*\}}\left(\beta_1 t^* + \beta_2 t^* a_1 + \beta_3(t_j - t^*) + \beta_4(t_j - t^*)a_1 + \beta_5(t_j - t^*)\mathbb{1}_{\{a_1=1\}}a_{2NR}\right). \quad (2.3)$$

In model (2.1), using contrast coding, i.e., $\{a_1, a_{2NR}\} \in \{-1, 1\}^2$, we can write

$$2\beta_2 = E\left[\frac{Y_j^{(1,0,\cdot)} - Y_k^{(1,0,\cdot)}}{t_j - t_k} - \frac{Y_j^{(-1,0,\cdot)} - Y_k^{(-1,0,\cdot)}}{t_j - t_k} \mid \boldsymbol{X}\right], \quad j,k \le T_1,\ j \ne k. \quad (2.4)$$

This represents the difference in slopes of expected treatment readiness in the first stage of the SMART between DTRs starting with different first-stage treatments (second-stage treatment is arbitrary, as $t < t^*$). Also, we can interpret η_1 as the difference in expected outcome $Y_j^{(a_1,a_{2R},a_{2NR})}$ associated with a one-unit difference in baseline covariate X_1, marginal over all embedded DTRs.

2.2 Estimation

2.2.1 Observed Data

Suppose we have data arising from a SMART with n participants. Let $A_{i,1} \in \{-1, 1\}$ be a random variable which indicates first-stage treatment randomly assigned to participant i ($i = 1, \ldots, n$), and let $R_i \in \{0, 1\}$ indicate whether the ith participant responded to $A_{i,1}$, in which case $R_i = 1$, or not, so $R_i = 0$. Define $A_{i,2} \in \{-1, 1\}$ to be the randomly-assigned second-stage treatment. Throughout, we use uppercase A to denote random treatment assignments; lowercase a's are non-random indices used to denote embedded DTRs.

In design II, since only non-responders are re-randomized, we set $A_{i,2} = 0$ for responders; similarly for design III. We observe a continuous outcome $Y_{i,j}$ for each participant at each measurement occasion t_j, $j = 1, \ldots, T$. In general, the data collected on the ith individual over the course

of the study are of the form

$$\left(\boldsymbol{X}_i, Y_{i,0}, A_{i,1}, \boldsymbol{Y}_{i,1:T_1}, R_i, A_{i,2}, \boldsymbol{Y}_{i,T_1+1:T}\right),$$

where $\boldsymbol{Y}_{i,j:k}$ is a vector consisting of all values of the outcome observed for the i^{th} participant at measurement occasions j through k.

2.2.2 Estimating Equations

Our goal is to estimate and make inferences on $\boldsymbol{\theta}$, the length-p column vector of mean parameters in the marginal structural mean model of interest. For notational convenience, let $\mathcal{D}$ be the set of DTRs embedded in the SMART under study; for instance, in design II,

$$\mathcal{D} = \left\{(a_1, a_{2R}, a_{2NR}) : a_1 \in \{-1, 1\}, a_{2R} = 0, a_{2NR} \in \{-1, 1\}\right\}.$$

We will write $\boldsymbol{Y}^{(d)} := \boldsymbol{Y}^{(a_1, a_{2R}, a_{2NR})}$.

Let $W^{(d)}(A_{i,1}, R_i, A_{i,2})$ be a weight associated with participant i and DTR $d \in \mathcal{D}$ defined as

$$W^{(d)}(A_{i,1}, R_i, A_{i,2}) = \frac{I^{(d)}(A_{i,1}, R_i, A_{i,2})}{P(A_{i,1} = a_1)P(A_{i,2} = a_2 \mid A_{i,1} = a_1, R_i)}, \tag{2.5}$$

where $I^{(d)}(A_{i,1}, R_i, A_{i,2})$ is an indicator of whether participant i is consistent with DTR d. The form of $I^{(d)}(A_{i,1}, R_i, A_{i,2})$ depends on the particular SMART design under study; for each of the designs in figure 1.1, these expressions are shown in table 2.1.

We use $W^{(d)}(A_{i,1}, R_i, A_{i,2})$ to account for the facts that, in some SMARTs (e.g., designs II and III) there is known imbalance in the proportion of responders and non-responders consistent with each DTR, and that that some (or all) participants are consistent with more than one embedded DTR. This imbalance can be corrected using inverse-probability weighting (Nahum-Shani et al. 2012a; Cole and Hernán 2008; Chakraborty and Moodie 2013).

In design II, for example, only non-responders to first-stage treatment are re-randomized; if

Table 2.1: Design-specific indicators for consistency with a given DTR $d \in \mathcal{D}$.

Design	$I^{(d)}(A_{i,1}, R_i, A_{i,2})$
I	$\mathbb{1}_{\{A_{i,1}=a_1\}}\left(\mathbb{1}_{\{A_{i,2}=a_{2R}\}}R_i + \mathbb{1}_{\{A_{i,2}=a_{2NR}\}}(1-R_i)\right)$
II	$\mathbb{1}_{\{A_{i,1}=a_1\}}\left(R_i + \mathbb{1}_{\{A_{i,2}=a_{2NR}\}}(1-R_i)\right)$
III	$\mathbb{1}_{\{A_{i,1}=a_1\}}\left(\mathbb{1}_{\{a_1=-1\}} + \mathbb{1}_{\{a_1=1\}}\left(R_i + \mathbb{1}_{\{A_{i,2}=a_{2NR}\}}(1-R_i)\right)\right)$

all randomizations are with probability 0.5, $W^{(1,0,1)}(1,1,0) = (.5\times 1)^{-1} = 2$ and $W^{(1,0,1)}(1,0,1) = (.5\times .5)^{-1} = 4$. Note that in design I, all participants are re-randomized; hence, all participants receive a weight of 4. The analyst may freely substitute $W^{(d)}(A_{i,1}, R_i, A_{i,2}) = I^{(d)}(A_{i,1}, R_i, A_{i,2})$ in this case.

Define $\boldsymbol{D}^{(d)}(\boldsymbol{X}_i) \in \mathbb{R}^{T\times p}$ to be the Jacobian of $\boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$; i.e., $\boldsymbol{D}^{(d)}(\boldsymbol{X}_i) = \partial\boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$. Let $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau}) \in \mathbb{R}^{T\times T}$ be a working covariance matrix for $\boldsymbol{Y}^{(d)}$, conditional on baseline covariates $\boldsymbol{X}$, under DTR $d \in \mathcal{D}$. Here, $\boldsymbol{\tau} = (\boldsymbol{\sigma}^\top, \boldsymbol{\rho}^\top)^\top$ is a vector of parameters indexing variance ($\boldsymbol{\sigma}$) and correlation ($\boldsymbol{\rho}$) components of the working covariance structure. We discuss $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})$ in detail in section 2.2.3. We estimate $\boldsymbol{\theta}$ by solving the estimating equations

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n}\sum_{d\in\mathcal{D}}\left[W^{(d)}\left(A_{i,1}, R_i, A_{i,2}\right)\cdot \boldsymbol{D}^{(d)}(\boldsymbol{X}_i)^\top \boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})\right)\right]. \tag{2.6}$$

We call the solution to equation (2.6) $\hat{\boldsymbol{\theta}}_n$, and investigate its asymptotic properties in the following propositions; see appendix B for proofs.

Proposition 2.1. *Suppose $\boldsymbol{\mu}^{(d)}(\boldsymbol{X};\boldsymbol{\theta})$ is a correctly-specified model for $\mathrm{E}\left[\boldsymbol{Y}^{(d)} \mid \boldsymbol{X}\right]$. Then $\boldsymbol{\theta}_n$ is consistent for $\boldsymbol{\theta}^*$, the true parameter value.*

Proposition 2.2. *$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)$ converges in distribution to $\mathcal{N}\left(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\right)$, where*

$$\boldsymbol{B} := \mathrm{E}\left[\sum_{d\in\mathcal{D}} W^{(d)}\left(A_{i,1}, R_i, A_{i,2}\right)\boldsymbol{D}^{(d)}(\boldsymbol{X}_i)^\top \boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})^{-1}\boldsymbol{D}^{(d)}(\boldsymbol{X}_i)\right] \in \mathbb{R}^{p\times p} \tag{2.7}$$

and

$$\boldsymbol{M} := \mathrm{E}\left[\left(\sum_{d\in\mathcal{D}} W^{(d)}\left(A_{i,1}, R_i, A_{i,2}\right)\boldsymbol{D}^{(d)}(\boldsymbol{X}_i)^\top \boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})\right)\right)^{\otimes 2}\right] \in \mathbb{R}^{p\times p}, \tag{2.8}$$

with $\boldsymbol{Z}^{\otimes 2} = \boldsymbol{Z}\boldsymbol{Z}^\top$.

Note that $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ regardless of the chosen structure of $\boldsymbol{V}^{(d)}(\boldsymbol{X};\boldsymbol{\tau})$; however, we conjecture that choices of $\boldsymbol{V}^{(d)}(\boldsymbol{X};\boldsymbol{\tau})$ closer to the true covariance matrix $\mathrm{Var}(\boldsymbol{Y}^{(d)})$ will yield more efficient estimates.

2.2.3 Estimation of the Working Covariance Matrix

In general, for an embedded DTR $d \in \mathcal{D}$, $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\sigma},\boldsymbol{\rho})$ takes the form

$$\boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\sigma},\boldsymbol{\rho}) = \boldsymbol{S}\left(\boldsymbol{\sigma}\right)^{1/2}\boldsymbol{R}^{(d)}(\boldsymbol{\rho})\boldsymbol{S}\left(\boldsymbol{\sigma}\right)^{1/2}, \tag{2.9}$$

where $\boldsymbol{S}(\boldsymbol{\sigma})^{1/2} \in \mathbb{R}^{T\times T}$ is a diagonal matrix with diagonal entries $\sigma_1,\ldots,\sigma_T$, and $\boldsymbol{R}^{(d)}(\boldsymbol{\rho})$ is a working correlation matrix for $\boldsymbol{Y}^{(d)}$. Note that this notation allows for different working correlation structures for each DTR, as well as non-constant variances in the repeated-measures outcome.

We propose the following procedure to estimate $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\boldsymbol{\sigma}^\top,\boldsymbol{\rho}^\top)^\top$. First, estimate $\boldsymbol{\theta}$ by solving equation (2.6) using the $T\times T$ identity matrix as $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})$ for all $d\in\mathcal{D}$. Call the solution $\hat{\boldsymbol{\theta}}_{(0)}$. Next, use $\hat{\boldsymbol{\theta}}_{(0)}$ to estimate $\sigma_t^{(d)}$ as follows

$$\hat{\sigma}_t^{(d)} = \frac{\sum_{i=1}^n W^{(d)}\left(A_{i,1},R_i,A_{i,2}\right)\left(Y_{i,t} - \mu_t^{(d)}\left(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}_{(0)}\right)\right)^2}{\sum_{i=1}^n W^{(d)}\left(A_{i,1},R_i,A_{i,2}\right) - p}, \tag{2.10}$$

where p is the dimension of $\boldsymbol{\theta}$. If the scientist believes that this variance is constant over time for each DTR, the estimate in equation (2.10) can be averaged over time; one can also average over DTR if one believes the variance is constant across all embedded DTRs. Estimators for $\boldsymbol{\rho}^{(d)}$

Table 2.2: Correlation estimators for selected working correlation structures. The top entries define estimators under the assumption of constant within-person variance over time; the bottom entries allow for time-varying variances. $d \in \mathcal{D}$ is an embedded DTR, $W_i^{(d)}$ is shorthand for $W^{(d)}(A_{i,1}, R_i, A_{i,2})$, and $\hat{e}_{i,t}^{(d)}(\hat{\boldsymbol{\theta}})$ is the estimated residual $Y_{i,t} - \hat{\mu}_t^{(d)}(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}})$.

Cor. structure	$\text{Cor}(Y_j^{(d)}, Y_k^{(d)})$	Estimator
AR(1)	$\begin{cases} 1 & j = k \\ \left(\rho^{(d)}\right)^{\lvert j-k \rvert} & j \neq k \end{cases}$	$\hat{\rho}^{(d)} = \frac{\sum_{i=1}^n W_i^{(d)} \sum_{m=1}^{T-1} \hat{e}_{i,m}^{(d)}(\hat{\boldsymbol{\theta}}) \hat{e}_{i,m+1}^{(d)}(\hat{\boldsymbol{\theta}})}{\left(\hat{\sigma}^{(d)}\right)^2 \cdot n \cdot (T-1)}$
Exchangeable	$\begin{cases} 1 & j = k \\ \rho^{(d)} & j \neq k \end{cases}$	$\hat{\rho}^{(d)} = \frac{\sum_{i=1}^n W_i^{(d)} \sum_{l<m} \hat{e}_{i,l}^{(d)}(\hat{\boldsymbol{\theta}}) \hat{e}_{i,m}^{(d)}(\hat{\boldsymbol{\theta}})}{\left(\hat{\sigma}^{(d)}\right)^2 \cdot n \cdot T(T-1)/2}$
Unstructured	$\begin{cases} 1 & j = k \\ \rho_{j,k}^{(d)} & j \neq k \end{cases}$	$\hat{\rho}_{j,k}^{(d)} = \frac{\sum_{i=1}^n W_i^{(d)} \hat{e}_{i,j}^{(d)}(\hat{\boldsymbol{\theta}}) \hat{e}_{i,k}^{(d)}(\hat{\boldsymbol{\theta}})}{\left(\hat{\sigma}^{(d)}\right)^2 \cdot n}$

vary with choice of correlation structure $\boldsymbol{R}^{(d)}(\boldsymbol{\rho})$; we present estimators for selected structures in table 2.2 which assume variance is constant in time.

We estimate $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i; \boldsymbol{\tau})$ by plugging appropriately-pooled estimates of $\boldsymbol{\sigma}$ from equation (2.10) and of $\boldsymbol{\rho}$ from table 2.2 into equation (2.9). The form of $\boldsymbol{R}^{(d)}(\boldsymbol{\rho})$ can be chosen according to existing domain knowledge for primary analyses; secondary analyses might use exploratory methods to discover an appropriate working correlation structure.

2.2.4 Iterated Estimation Procedure

After estimating $\boldsymbol{V}^{(d)}(\boldsymbol{X}_i; \boldsymbol{\tau})$, we again solve equation (2.6), this time using $\hat{\boldsymbol{V}}^{(d)}(\boldsymbol{X}_i; \hat{\boldsymbol{\tau}}^{(d)}) = \boldsymbol{S}(\hat{\boldsymbol{\sigma}}^{(d)})^{1/2} \boldsymbol{R}^{(d)}(\hat{\boldsymbol{\rho}}^{(d)}) \boldsymbol{S}(\hat{\boldsymbol{\sigma}}^{(d)})^{1/2}$ as the working covariance matrix. This yields a "one-step" estimator of $\boldsymbol{\theta}$, which we denote by $\hat{\boldsymbol{\theta}}_{(1)}$. This process can be further iterated, as suggested by Liang and Zeger (1986); that is, we can use $\hat{\boldsymbol{\theta}}_{(1)}$ in equation (2.10) to obtain a new estimate for the working covariance structure, and so on until convergence. We call the final estimate of the model parameters $\hat{\boldsymbol{\theta}}$.

Work by Lipsitz et al. (2017) indicates that the one-step estimator is asymptotically equivalent to the "fully-iterated" estimator and is much less computationally intensive when the number of repeated measures is large. Anecdotally, we have found in reasonable simulation models for SMARTs with five or fewer measurement occasions that fully-iterated estimates tend to converge

in L_2-norm within a tolerance of 10^{-8} after about six iterations and do not represent significant computational burden.

2.3 Sample Size Formulae for End-of-Study Comparisons of Embedded DTRs in Two-Stage SMARTs

The estimation procedure presented in section 2.2 is general. The marginal structural mean model $\boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})$ can take any form appropriate for the SMART under analysis, data can be observed at any number of measurement occasions, and the working covariance matrix can have arbitrary structure (Lu et al. 2016).

We now develop sample size formulae for SMARTs in which the primary aim is to compare the mean end-of-study outcomes for two embedded DTRs that recommend different first-stage treatments and which satisfy certain design constraints. For a variety of reasons, there is an interest in collecting repeated-measures outcomes even in settings in which the primary aim is an end-of-study comparison. Because repeated measurements within the same person are often positively correlated, analyses which leverage this information can be more efficient than those which do not (Cook and Ware 1983). This can be especially beneficial in situations with small signal-to-noise ratios. Furthermore, longitudinal data allows investigators to examine trajectories over time, regardless of the primary comparison. This can help tell a fuller story about change over time.

For the sample size methods developed here, we restrict our focus to two-stage SMARTs in which the outcome is observed at three equally-spaced measurement occasions — baseline, just prior to assessment of the tailoring variable and subsequent randomization, and at the end of the study — and in which all randomizations occur with probability 0.5. For simplicity, we ignore baseline covariates; this is a conservative assumption, since it will inflate the variance of the estimates from equation (2.6). Additionally, we consider a saturated, piecewise-linear mean structure $\boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})$ similar to models (2.1) to (2.3).

Let $\boldsymbol{c}$ be some contrast vector so that the primary aim null hypothesis takes the form

$$H_0: \quad \boldsymbol{c}^\top \boldsymbol{\theta} = \mathbf{0},$$

which we will test against an alternative of the form H_1: $\boldsymbol{c}^\top \boldsymbol{\theta} = \Delta$. To compare mean end-of-study outcomes between two embedded DTRs which recommend different first-stage treatments, the estimand of interest is

$$\boldsymbol{c}^\top \boldsymbol{\theta} = \mathrm{E}\left[Y_3^{(1,a_{2R},a_{2NR})} - Y_3^{(-1,a'_{2R},a'_{2NR})}\right], \tag{2.11}$$

for some choice of a_{2R}, a'_{2R}, a_{2NR}, and a'_{2NR}. For example, to test equality of mean end-of-study outcomes for DTRs (1, 0, 1) and (-1, 0, -1) in design II under model (2.1) (assuming no $\boldsymbol{X}$, $\{t_j\} = \{0, 1, 2\}$, $t^* = 1$), the estimand (2.11) can be written as the linear combination $\boldsymbol{c}^\top \boldsymbol{\beta}$, where $\boldsymbol{c}^\top = (0, 0, 2, 0, 2, 2, 0)$.

Because we are interested in a single contrast (i.e., $\boldsymbol{c}$ is a vector, not a matrix) we employ a 1-degree of freedom Wald test. The test statistic is

$$Z = \frac{\sqrt{n}\boldsymbol{c}^\top \hat{\boldsymbol{\theta}}}{\sigma_c},$$

where $\sigma_c = \sqrt{\boldsymbol{c}^\top \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}}$. Under the null hypothesis, by asymptotic normality of $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)$, the test statistic follows a standard normal distribution.

Define δ to be the standardized effect size as described by Cohen (1988) for an end-of-study comparison, i.e.,

$$\delta = \frac{\Delta}{\sigma}, \tag{2.12}$$

where $\sigma = \mathrm{Var}(Y_j^{(d)})$ for arbitrary j (see working assumption A2.2 below).

The sample size formulae will require the response rate $P(R^{(a_1)} = 1) = r_{a_1}$, where $R^{(a_1)}$ is the potential response to first-stage treatment a_1. In order to simplify the form of σ_c and obtain tractable, intepretable sample size formulae, we make the following working assumptions:

A2.1 *Constrained conditional covariance matrices for DTRs under comparison.*

(a) The variability of $Y_t^{(d)}$ around the DTR mean $\mu_t^{(d)}(\boldsymbol{\theta})$ among responders and non-responders is no more than the variance of $Y_t^{(d)}$ unconditional on response, i.e.,

$$\mathrm{E}\left[\left(Y_t^{(d)} - \mu_t^{(d)}(\boldsymbol{\theta})\right)^2 \mid R^{(a_1^{(d)})}\right] \leq \mathrm{E}\left[\left(Y_t^{(d)} - \mu_t^{(d)}(\boldsymbol{\theta})\right)^2\right],$$

for all $t > t^*$ and DTRs $d \in \mathcal{D}$ under study.

(b) For times $t_i \leq t_j \leq t^*$, response status is uncorrelated with products of residuals of Y_{t_i}, i.e,

$$\mathrm{Cov}\left(R^{(a_1^{(d)})}, \left(Y_{t_i}^{(d)} - \mu_{t_i}^{(d)}(\boldsymbol{\theta})\right)\left(Y_{t_j}^{(d)} - \mu_{t_j}^{(d)}(\boldsymbol{\theta})\right)\right) = 0.$$

for DTRs $d \in \mathcal{D}$ under study.

(c) The covariance between the end-of-study measurement and the measurements prior to the second stage among responders is less than or equal to the same quantity among non-responders:

$$\mathrm{Cov}\left(Y_t^{(d)}, Y_2^{(d)} \mid R^{(a_1^{(d)})} = 1\right) \leq \mathrm{Cov}\left(Y_t^{(d)}, Y_2^{(d)} \mid R^{(a_1^{(d)})} = 0\right)$$

for DTRs $d \in \mathcal{D}$ under study and $t = 0, 1$. An additional, related assumption is given in appendix B.2.

A2.2 *Exchangeable marginal covariance structure.* The marginal variance of $\boldsymbol{Y}^{(d)}$ is constant across time and DTR, and has an exchangeable correlation structure with correlation ρ, i.e.,

$$\mathrm{Var}\left(\boldsymbol{Y}^{(d)}\right) = \boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

for all $d \in \mathcal{D}$.

These working assumptions may be seen as overly simplifying; however, we will see in sec-

tions 2.4.2 and 2.5 that formula (2.13) is robust to moderate violations of working assumption A2.1 and that inputs to the formula can be adjusted in a way to accommodate violations of working assumption A2.2. A working assumption similar to A2.1(a) is commonly made in developing sample-size formulae for SMARTs using end-of-study outcomes (Oetting et al. 2011; Kidwell et al. 2018; NeCamp, Kilbourne, and Almirall 2017). Working assumptions A2.1(b) and A2.1(c), as well as A2.2, are necessary for the extension to the setting of a repeated-measures outcome with our proposed estimator.

Working assumption A2.1 arises specifically as a consequence of unequal weights in equation (2.6) (i.e., when there exists imbalance between responders and non-responders, by design); therefore, the assumption is not necessary in design I, and can be relaxed to apply to only the two DTRs in which non-responders are re-randomized in design III. Furthermore, working assumption A2.2 cannot be satisfied in design I if all eight embedded DTRs have unique means.

Under working assumptions A2.1 and A2.2, the minimum-required sample size to detect a standardized effect size δ with power at least $1 - \gamma$ and two-sided type-I error α is

$$n \geq \frac{4\left(z_{1-\alpha/2} + z_{1-\gamma}\right)^2}{\delta^2} \cdot (1 - \rho^2) \cdot \text{DE}, \tag{2.13}$$

where DE is a SMART-specific "design effect" for an end-of-study comparison (see table 2.3). Note that the first term in formula (2.13) is the typical sample size formula for a traditional two-arm randomized trial with a continuous end-of-study outcome and equal randomization probability. The middle term is due to the within-person correlation in the outcome, and is identical to the corresponding correction term for GEE analyses sized to detect a group-by-time interaction when there is no baseline group effect (see, e.g, Fitzmaurice, Laird, and Ware 2011, ch. 20)

The sample size formula presented in formula (2.13) is conservative. It becomes more conservative as ρ approaches $(1 + \sqrt{5})/2 \approx 0.62$. A sharper formula is available in appendix B.2; however, we emphasize formula (2.13) as it is more immediately intepretable.

Table 2.3: SMART-specific design effects for sample size formula 2.13. $r_{a_1} = P(R^{(a_1)} = 1)$ is the response rate to first-stage treatment a_1. The conservative design effect maximizes the sample size requirement by assuming $r_{a_1} = 0$.

Design	Design effect	Conservative design effect
I	2	2
II	$\frac{1}{2}(2 - r_1) + \frac{1}{2}(2 - r_{-1})$	2
III	$\frac{1}{2}(3 - r_1)$	$\frac{3}{2}$

2.4 Simulation Study

We conducted a variety of simulations to assess the performance of sample size formula (2.13). We are interested in the empirical power for a comparison of the DTR which recommends only treatments indicated by 1 and the DTR which recommends only treatments indicated by -1 when the study is sized to detect an effect size of δ. In ENGAGE, this might correspond to a comparison of mean end-of-study outcomes under the DTR which recommends MI-IOP in the first stage, no further contact for engagers, and MI-PC in the second stage for continued non-engagers versus the mean end-of-study outcomes under the regimen which recommends MI-PC in the first-stage, then no further contact for both engagers and non-engagers.

We consider four types of scenarios: first, when no assumptions are violated; second, when each of working assumptions A2.1(a) to A2.1(c) are violated; finally, when the working correlation structure is misspecified, in violation of working assumption A2.2. In each scenario, sample sizes are computed based on nominal power $1 - \gamma = 0.8$ and two-sided type-I error $\alpha = 0.05$.

We believe sample sizes from formula (2.13) will be slightly conservative when all assumptions are satisfied, as formula (2.13) is an interpretable upper bound on a sharper formula given in appendix B.2 and the supplement. For design I, we do not expect power to be affected by violations of working assumption A2.1, as the assumption arises as a consequence of over- or under-representation of responders and non-responders consistent with a particular DTR (see appendix B.2). Since there is no such imbalance in design I, working assumption A2.1 is not applicable. Similarly, in design III, only non-responders to one first-stage treatment are re-randomized, so we expect that empirical power will decrease slightly, but not seriously, when violating working

assumption A2.1. We expect empirical power to suffer most severely when violating this working assumption in design II.

We further conjecture that scenarios in which the true within-person correlation structure of $\boldsymbol{Y}^{(d)}$ is autoregressive, sample sizes from formula (2.13) will be very anti-conservative. Under an AR(1) correlation structure, less information about the end-of-study outcome is provided by the baseline measure than would be under an exchangeable correlation structure. Since, by using formula (2.13), we have assumed more information is available from earlier measurements than is actually the case, we will be underpowered. Similarly, we expect over-estimation of ρ in formula (2.13) to lead to anti-conservative sample sizes.

2.4.1 Data Generative Process

For each simulation, the true marginal mean model is as in models (2.1) to (2.3) for designs I to III, respectively. We do not include baseline covariates $\boldsymbol{X}$; this is a conservative approach, as we believe that adjustment for prognostic covariates typically will increase power: see, eg., Kahan et al. (2014). Estimates of marginal means from ENGAGE were used to inform a reasonable range of "true" means from which to simulate, though the scenarios presented here are not designed to mimic ENGAGE exactly. All simulations take $T = 3$ and values of $\boldsymbol{\beta}$ and σ are chosen to achieve $\delta = 0.3$ or $\delta = 0.5$ ("small" and "moderate" effect sizes, respectively).

Data were generated according to a conditional mean model which, when averaged over response, yields the marginal model of interest. Potential outcomes $Y_{i,j}^{(d)}$ were simulated from appropriately-parameterized normal distributions (see section 4.1 for details); data were "observed" by selecting the potential outcome corresponding to treatment assignment as generated from a Bernoulli(0.5) distribution.

We consider three mechanisms for generating response status. In the first, "$R_{\perp\!\!\!\perp}$", response is generated from a Bernoulli(r_{a_1}) distribution, where $r_{a_1} = P(R^{(a_1)} = 1)$, independently of all previously-observed data. In the second and third scenarios ("R_+" and "R_-", respectively), response status is still generated from a Bernoulli distribution, but each individual is assigned a beta-

distributed probability of response correlated with their observed value of Y_1. These correlations are either positive or negative, depending on the response model. This is intended to mimic different coding choices for $\boldsymbol{Y}$, in the sense of responders tending to have higher or lower values of Y_1 than non-responders.

For each simulation scenario, we compute upper and lower bounds on allowable values of $\text{Var}(Y_2^{(d)} \mid R^{(a_1^{(d)})} = 1)$, beyond which it is not possible to either achieve the desired marginal variance, or which induces violation of working assumption A2.1(a). The results shown in the corresponding column of table 2.4 were generated when responders' variances were set to 75% of the lower bound beyond which the fixed marginal variance forces $\text{E}[(Y_t^{(d)} - \mu_t^{(d)}(\boldsymbol{\theta}))^2 \mid R^{(a_1^{(d)})} = 0] \geq \sigma^2$.

Violation of working assumption A2.1(b) was induced by defining response status as

$$R^{(a_1^{(d)})} = \mathbb{1}_{\left\{Y_1^{(d)} \in \left(-\infty, \kappa_{a_1}^{\text{low}}\right] \cup \left[\kappa_{a_1}^{\text{high}}, \infty\right)\right\}}, \tag{2.14}$$

where κ^{low} and κ^{high} are chosen to be the $r/2$ and $(1 - r/2)$th quantiles of the $\mathcal{N}\left(\mu_1^{(d)}, \sigma^2\right)$ distribution, respectively. This ensures control on response probability while also inducing large positive correlation between $R^{(a_1)^{(d)}}$ and $(Y_1^{(d)} - \mu_1^{(d)})^2$.

Violation of working assumption A2.1(c) was induced by choosing $\text{Cor}(Y_t^{(d)}, Y_2^{(d)} \mid R^{(a_1)} = 1) > \text{Cor}(Y_t^{(d)}, Y_2^{(d)} \mid R = 0)$ while keeping respective variances fixed. In our generative model, it was difficult to exert precise control over these quantities when response was related to prior outcomes; as such, these violations were induced under the $R_{\perp\!\!\!\perp}$ response model.

There exist natural constraints on how much larger than $\text{Cov}(Y_t^{(d)}, Y_2^{(d)} \mid R = 0)$ the responders' covariance can be while ensuring that (1) all conditional covariance matrices are positive definite and (2) $\text{Cov}(Y_t^{(d)}, Y_2^{(d)} \mid R^{(a_1^{(d)})} = 0) \geq 0$ for $t = 0, 1$. These constraints vary with ρ. We choose $\text{Cor}(Y_t^{(d)}, Y_2^{(d)} \mid R^{(a_1^{(d)})} = 1)$ such that $\text{Cov}(Y_t^{(d)}, Y_2^{(d)} \mid R^{(a_1^{(d)})} = 1)$ is the midpoint between the minimum covariance for which the assumption is violated and the maximum covariance allowed by the aforementioned constraints.

2.4.2 Simulation Results

Simulation results based on 3,000 simulated data sets are compiled in table 2.4. We find that sample size formula (2.13) performs as expected when all assumptions are satisfied. Empirical power is not significantly less than the target power of 0.8, per a one-sided binomial test with level 0.05. The sample size is, as expected, often conservative, particularly when within-person correlation is high.

Table 2.4: Sample sizes and empirical power results for an end-of-study comparison of the DTR recommending only treatments indexed by 1 and that which recommends only treatments indicated by -1. δ is the true standardized effect size as defined in equation (2.12), r is the common probability of response to first-stage treatment, and ρ is the true exchangeable within-person correlation. n is computed using formula (2.13) with $\alpha = 0.05$ and $\gamma = 0.2$. $R_{\perp\!\!\!\perp}$ refers to a generative model in which response status is independent of all prior outcomes; R_+ and R_- refer to generative models in which response is positively or negatively correlated with Y_1, respectively. All violation scenarios assume the R_+ generative model, except working assumption A2.1(c). Results are the proportion of 3000 Monte Carlo simulations in which we reject $H_0 : \boldsymbol{c}^\top \boldsymbol{\theta} = 0$ at the 5% level.

					Empirical power						
					A2.1 and A2.2 satisfied			Violation of A2.1			Violation of A2.2
Design	δ	r	ρ	n	$R_{\perp\!\!\!\perp}$	R_+	R_-	A2.1(a)	A2.1(b)	A2.1(c)	True AR(1)
I	0.3	0.4	0.0	698	0.798	0.807	0.803	0.798	0.796	‡	‡
			0.3	635	0.819	0.817	0.800	0.820	0.804	0.815	0.780*
			0.6	447	0.815	0.862	0.773*	0.865	0.817	0.827	0.728*
			0.8	252	0.835	0.925	0.733*	†	†	0.840	0.721*
		0.6	0.0	698	0.796	0.799	0.806	0.800	0.791	‡	‡
			0.3	635	0.808	0.813	0.792	0.824	0.805	0.807	0.775*
			0.6	447	0.833	0.856	0.798	0.859	0.831	0.838	0.727*
			0.8	252	0.827	0.901	0.758*	†	†	0.835	†
	0.5	0.4	0.0	252	0.799	0.801	0.798	0.798	0.801	‡	‡
			0.3	229	0.813	0.815	0.797	0.814	0.811	0.814	0.771*
			0.6	161	0.824	0.872	0.789	0.868	0.833	0.843	0.742*
			0.8	91	0.843	0.931	0.734*§	0.926	†	0.839§	0.725*
		0.6	0.0	252	0.796	0.797	0.810	0.792	0.802	‡	‡
			0.3	229	0.817	0.815	0.808	0.811	0.823	0.823	0.771*
			0.6	161	0.838	0.859	0.790	0.861	0.832	0.837	0.749*
			0.8	91	0.835§	0.896	0.765*§	0.896	†	0.859	†
II	0.3	0.4	0.0	559	0.801	0.801	0.808	0.778*	0.803	‡	‡
			0.3	508	0.804	0.813	0.831	0.800	0.797	0.798	0.795
			0.6	358	0.817	0.819	0.834	0.807	0.759*	0.788	0.811
			0.8	201	0.836	0.814	0.836	0.809	†	0.792	0.806
		0.6	0.0	489	0.804	0.796	0.793	0.736*	0.810	‡	‡
			0.3	445	0.797	0.804	0.818	0.758*	0.795	0.780*	0.804
			0.6	313	0.824	0.831	0.844	0.793	0.752*	0.770*	0.824
			0.8	176	0.845	†	†	0.754*	†	0.776*	0.842
	0.5	0.4	0.0	201	0.801	0.800	0.802	0.768*	0.794	‡	‡

continued

					Empirical power						
					A2.1 and A2.2 satisfied			Violation of A2.1			Violation of A2.2
Design	δ	r	ρ	n	$R_{\perp\!\!\!\perp}$	R_+	R_-	A2.1(a)	A2.1(b)	A2.1(c)	True AR(1)
			0.3	183	0.813	0.800	0.819	0.790	0.813	0.796	0.803
			0.6	129	0.814	0.828	0.833	0.810	0.763*	0.799	0.815
			0.8	73	0.839	0.841	0.852	0.829	†	0.795	0.804
		0.6	0.0	176	0.807	0.799	0.796	0.733*	0.808	‡	‡
			0.3	160	0.816	0.815	0.821	0.767*	0.808	0.802	0.812
			0.6	113	0.829	0.830	0.837	0.792	0.765*	0.770*	0.817
			0.8	64	0.845§	†	†	0.783*§	†	0.789§	†
III	0.3	0.4	0.0	454	0.806	0.813	0.806	0.782*	0.794	‡	‡
			0.3	413	0.815	0.809	0.814	0.789	0.800	0.800	0.775*
			0.6	291	0.821	0.811	0.818	0.794	0.783*	0.787*	0.687*
			0.8	164	0.824	0.812	0.839	0.812	†	0.802	0.637*
		0.6	0.0	419	0.813	0.814	0.817	0.781*	0.769*	‡	‡
			0.3	381	0.823	0.812	0.808	0.776*	0.791	0.795	0.771*
			0.6	268	0.823	0.817	0.844	0.807	0.750*	0.754*	0.709*
			0.8	151	0.820	†	†	0.803	†	0.784*	†
	0.5	0.4	0.0	164	0.808	0.804	0.795	0.776*	0.802	‡	‡
			0.3	149	0.822	0.815	0.827	0.811	0.791	0.805	0.789
			0.6	105	0.811	0.810	0.812	0.810	0.798	0.785*	0.698*
			0.8	59	0.838	†	0.823	0.845	†	0.817§	0.684*
		0.6	0.0	151	0.798	0.809	0.803	0.778*	0.772*	‡	‡
			0.3	138	0.812	0.809	0.814	0.800	0.782*	0.799	0.778*
			0.6	97	0.803§	0.812	0.826§	0.826§	0.762*	0.774*§	0.705*§
			0.8	55	0.826§	†	†	0.837§	†	0.797§	†

* Statistically significantly less than 0.8 at the 5% level.
† Our data generative model could not accommodate this scenario.
‡ Violation of this working assumption is not applicable when $\rho = 0$.
§ Fewer than 3000 simulations generated data in which all treatment sequences were observed.

There may be some concern that, for high within-person correlation, formula (2.13) is overly conservative; should this concern arise, we recommend use of the sharper formulae presented in the supplement. The difference between the sharper formulae and formula (2.13) is maximized when $\rho = (1+\sqrt{5})/2 \approx 0.62$, so we expect to see the largest differences in power between formula (2.13) and the sharp formula when we set $\rho = 0.6$.

When all working assumptions are satisfied, we see that empirical power for R_+ and R_- scenarios are similar or slightly higher than under the $R_{\perp\!\!\!\perp}$ model. In general, there do not appear to be practical differences in empirical power between the response models.

As conjectured, violating working assumption A2.1(a) does not impact empirical power in design I (compare the results to column "R_+"). For design II, empirical power is consistently

less than the nominal value when working assumption A2.1(a) is violated. However, while the empirical power is often statistically significantly less than 0.8, for practical purposes the loss of power is relatively small. For design III, we notice small reductions in power relative to scenarios in which both working assumptions A2.1 and A2.2 are satisfied, though the conservative nature of formula (2.13) appears to protect against more severe loss of power. This suggests that our sample size formula is moderately robust to reasonable violations of A2.1(a).

For small ρ, we see no meaningful change in empirical power when violating working assumption A2.1(b). However, as ρ increases, this also leads to increased correlation between response and the other products of first-stage residuals, which increases the severity of the violation. For $\rho = 0.6$, we see noticeable, but not extreme, departures from nominal power. When $\rho = 0.8$, our generative model was not able to violate working assumption A2.1(b) without also violating working assumption A2.1(a); as such, we omit those results.

Interestingly, as can be seen in the supplement, defining *non*-response as in equation (2.14) (i.e., replacing $R^{(a_1)}$ with $1 - R^{(a_1)}$) leads to higher-than-nominal power. When there exists *negative* correlation between response and products of squared first-stage residuals, the form of σ_c^2 derived in appendix B.2 is more conservative, leading to increased power.

Simulation results show that our sample size formula is quite robust to violations of working assumption A2.1(c) for low-to-moderate within-person correlations; at high correlations, the empirical power is statistically significantly less than 0.8. However, as with working assumption A2.1(a), the practical reduction in power is relatively small.

The final column of table 2.4 suggests that formula (2.13) is highly sensitive to violations of working assumption A2.2 in regards to the true correlation structure. In particular, when the true correlation structure is not exchangeable with correlation ρ and is instead AR(1) with correlation ρ, empirical power is substantially lower than the target of 0.8, particularly as ρ increases. This is unsurprising: as our assumed exchangeable ρ increases, the difference between the assumed and actual correlation between the end-of-study measurement and earlier measurements increases, leading to more severe loss of power.

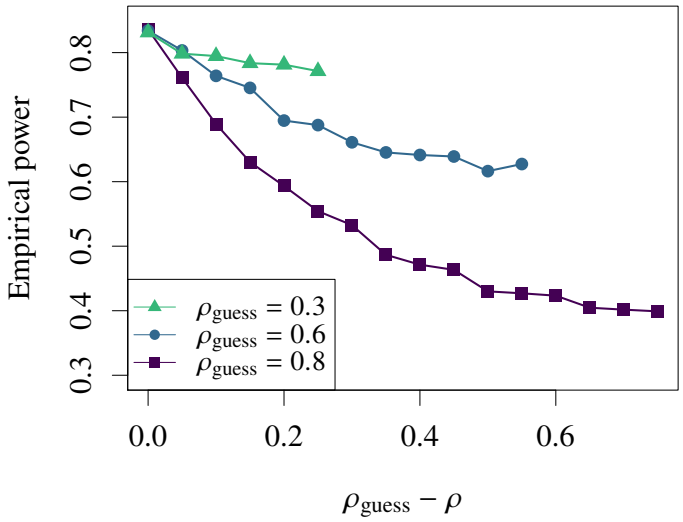


Figure 2.1: Empirical power under misspecified within-person correlation. Simulated power is plotted against the difference between the true within-person correlation ρ and hypothesized correlation ρ_{guess} used to compute sample size. Results are shown for design II with a hypothesized response rate of 0.4, and sample size was chosen to detect standardized effect size $\delta = 0.3$ for the comparison of DTRs (1, 0, 1) and (-1, 0, -1). Each point is based on 3000 simulations with target power 0.8 and significance level 0.05. Results are similar for designs I and III and different values of δ and r.

Note that when within-person correlation is high, sample size becomes rather small. Since the method presented here is based on asymptotic normality, we caution the reader that small sample sizes (e.g., $n < 100$) provided by formula (2.13) may be quite sensitive to violation of the working assumptions.

In figure 2.1, we examine the effect on empirical power of misspecifying the within-person correlation. Analytically, we see from formula (2.13) that if the assumed ρ is smaller than the true within-person correlation, the sample size will be conservative. On the other hand, when the assumed ρ in formula (2.13) is larger than the true correlation, the sample size will be anti-conservative. Figure 2.1 shows plots of empirical power against the difference between the assumed within-person correlation ρ_{guess} and the true ρ. For small ρ_{guess}, formula (2.13) appears to be quite robust to misspecification of ρ; however, as ρ_{guess} increases, the formula becomes highly sensitive to such a violation of working assumption A2.2. This is supported analytically, since formula (2.13)

is a function of ρ^2_{guess}.

2.5 Discussion

We have derived sample size formulae for SMART designs in which the primary aim is a comparison of two embedded DTRs that begin with different first-stage treatments on a continuous, longitudinal outcome observed at three measurement occasions. We derived the formulae for three common SMART designs.

The sample size formula is the product of three components: (1) the formula for the minimum sample size for the comparison of two means in a standard two-arm trial (see, e.g., Friedman, Furberg, and DeMets (2010) page 147), (2) a deflation factor of $1 - \rho^2$ that accounts for the use of a longitudinal outcome, and (3) a SMART-specific "design effect", an inflation factor that accounts for the SMART design.

The SMART design effect can be interpreted as the cost of conducting the SMART relative to conducting a standard two-arm randomized trial of the two DTRs which comprise the primary aim. The benefit of conducting a SMART (relative to the standard two-arm randomized trial) is the ability to answer additional, secondary questions that are useful for constructing effective DTRs. For example, such questions may focus on one or more of the other pairwise comparisons between DTRs, on whether the first- and second-stage treatments work synergistically to impact outcomes (e.g., a test of the null that $\beta_6 = 0$ in model (2.1)), or may focus on hypothesis-generating analyses that seek to estimate more deeply-tailored DTRs (Watkins 1989; Nahum-Shani et al. 2012b; Zhang et al. 2015).

The formulae are expected to be easy-to-use for both applied statistical workers and clinicians. Indeed, inputs α, γ, and Δ are as in the sample size formula for a standard z-test. Furthermore, estimates of ρ, r_{a_1}, and σ are often readily available from the literature or can be estimated using data from prior studies (e.g., prior randomized trials, or external pilot studies).

We make a number of recommendations concerning the use of the formulae; in particular,

how best to use the formulae conservatively in the absence of certainty concerning prior estimates of ρ, r_{a_1}, and/or the structure of the variance of the repeated measures outcome. First, in designs II and III, if there is uncertainty concerning the response rate (e.g., response rate estimates are based on data from smaller prior studies), one approach is to err conservatively by assuming a smaller-than-estimated response rate. In both designs, the most conservative approach is to assume a response rate of zero.

Second, as in standard randomized trials in which the primary aim is a pre-post comparison, the required sample size decreases as the hypothesized within-person correlation increases (Zhang, Cao, and Ahn 2014). Therefore, if the hypothesized ρ is larger than the true ρ, the computed sample size will be anti-conservative, resulting in an under-powered study. Indeed, we see this in the results of the simulation experiment (see figure 2.1). Here, again, one approach is to err conservatively towards smaller values of ρ.

Finally, working assumption A2.2 (concerning the variance of the repeated measures outcome) may be seen as overly restrictive in the imposition of an exchangeable correlation structure. For example, studies with a continuous repeated measures outcome may observe an autoregressive correlation structure. However, the exchangeable working assumption can be employed conservatively in the following way: if the hypothesized structure is not exchangeable, one approach is to set ρ in formula (2.13) to the smallest plausible value (e.g., the within-person correlation between the baseline and end-of-study measurements for an autoregressive structure). Because this approach utilizes a lower bound on the value of the true within-person correlations, it is expected to yield a larger than needed (more conservative) sample size. Similarly, if the true within-person correlation is expected to differ by DTR, one approach is to employ the smallest plausible ρ. As with the third recommendation, these recommendations are not unique to SMARTs; indeed, these strategies may also be used to size standard two-arm randomized trials with repeated measures outcome.

In the case where $\text{Var}(Y_j^{(d)})$ varies with time and/or DTR, we conjecture that power will suffer if a pooled estimate of σ^2 is used when the variance decreases with time. To see this, consider that the standardized effect size δ defined in equation (2.12) has as a denominator the pooled standard

deviation of $Y_2^{(d)}$ across the groups under comparison. Should the estimate of pooled standard deviation be larger than the true value, the variance of $\boldsymbol{c}^\top \hat{\boldsymbol{\theta}}$ will increase; since the estimate will be less efficient than hypothesized, power will be lower than expected. Conversely, we also conjecture that when $\text{Var}(Y_j^{(d)})$ increases with t, the sample size will be conservative using similar reasoning.

CHAPTER 3

Balancing Sample Size and Measurement Occasions in Longitudinal SMARTs

Monetary costs are a key consideration of any study; clinical trials are no exception. Martin et al. (2017) found that "much of the variability in [trial] costs is related to trial protocol design choices and factors such as the number of [participants], sites, and visits". Myers et al. (2019) discuss costs and challenges associted with recruiting participants for a study of major depressive disorder in adults with type 2 diabetes mellitus, reporting an average of $1358 spent per patient recruited. Strategies for recruitment included outreach to physicians, pharmacists, and community diabetes education programs, as well as direct advertising on Facebook, radio, and television. Sertkaya et al. (2016) found, in a systematic review of multi-site clinical trials in medical research, that while per-patient costs represent less than half of the total costs associated with these large trials, recruitment, retention, and intervention expenditures represent an important component of trial costs.

Longitudinal between-groups analyses have the advantage of reduced sample size requirements compared to a cross-sectional analysis due to within-person correlation (Hedeker, Gibbons, and Waternaux 1999). It therefore stands to reason that sample size requirements can be further reduced with more frequent measurement of the outcome. There is a broad literature exploring the selection of sample size and number of measurement occasions in randomized trials. To our knowledge, however, this problem has not been addressed in the context of a SMART.

Overall and Doyle (1994) provided a variety of sample size formulae for "[ANOVA] tests of significance in a two-group repeated measurements design", considering a variety of possible comparisons with and without adjustment for baseline covariates. They discovered that, for the

these comparisons, the total number of repeated measurements is less meaningful for power than the number of individuals enrolled in the study. In particular, at least for ANOVA tests, "the analysis of simple difference scores at endpoint is blind to the number of intervening repeated measurements" (Overall and Doyle 1994).

Maxwell (1998) showed that two-arm longitudinal trials in which the outcome is measured five or more times, sized for a comparison of slopes using ANOVA, yield important sample size reductions relative to a pretest-posttest design with analyzed via analysis of covariance. This was confirmed and extended by Arndt et al. (2000), who explored the problem in terms of "precision"; specifically, the relative contributions of additional sample size versus more measurement occasions to the standard error of the estimate for mean change over time. The idea of optimizing for the standard error of estimates is interesting; although, for a fixed effect size, target power, and type-I error, this can be equivalent to optimizing for power.

Raudenbush and Xiao-Feng (2001) considered this problem in two-level hierarchical models, representing power for a detecting a treatment effect as a function of study duration, measurement frequency, and sample size. The authors focused primarily on "group differences in polynomial change", modeling individual trajectories as polynomial functions in time. Furthermore, they conceptualized the effect size as a standardized mean difference in polynomial trend. This is a useful framework to consider, as linear trends may be overly restrictive.

Zhang and Ahn (2011a) consider tradeoffs between adding participants or measurements in the context of a test for a group-by-time interaction using a GEE estimator under a cost constraint. Their results are useful for two-arm randomized trials in which the primary aim is a comparison of slopes. Further work by the same authors extended the results of Overall and Doyle (1994) by considering a regression model for time-averaged outcomes across groups (Zhang and Ahn 2011b).

The sample size formula presented in Chapter 2 focuses specifically on SMARTs in which the longitudinal outcome is measured three times. This allowed us to narrow our focus to a saturated model, which greatly simplified computations. However, such an approach is overly simplistic. Consider the ENGAGE trial (Figure 1.2) in which outcomes were collected 4, 8, 12, and 24 weeks

after the baseline assessment (McKay et al. 2015). We conjectured in Chapter 2 that the three-timepoint simplification would yield a conservative sample size for designs which measure the outcome more than three times; here, we investigate that conjecture further.

In this chapter, we extend the method in chapter 2 to accommodate a general number of measurement occasions, with the goal of investigating relationships between the frequency and timing of measurements and sample size. We present a more general sample size formula in section 3.2, then discuss balancing sample size and repeated measurements subject to monetary considerations in section 3.3 .We introduce a simple cost function to describe per-participant expenditures like recruitment and measurement costs, and explore how to allocate resources between sample size and number of measurement occasions to both minimize cost and achieve a desired target power. We implement this optimization in an R package, described in chapter 4. Our primary contribution is a reframing of conversations about sample size between clinicians and statisticians by accounting for budget considerations in the design stages of the trial, discussed in detail in section 3.4

3.1 Modeling and Estimation

As in section 2.3, we wish to develop sample size formulae for longitudinal SMARTs in which the primary aim is the end-of-study comparison of two embedded DTRs which recommend different first-stage treatments. As before, we consider piecewise-linear models which respect the sequential randomization in a SMART, as in models (2.1) to (2.3). With three measurement occasions, these models are fully saturated: the model simply estimates $n_{\text{DTR}} + 3$ means, where n_{DTR} is the number of dynamic treatment regimens embedded in the SMART.

Consider a SMART in which the outcome is measured at T occasions $t_1 < t_2 < \cdots < t_T$. Define t^* as the time of the last measurement prior to re-randomization, T_1 as the number of measurements in the first stage, from baseline (t_1) through $t^* := t_{T_1}$, and T_2 as the number of measurements after t^* (i.e., in the second stage of the SMART), such that $T_1 + T_2 = T$.

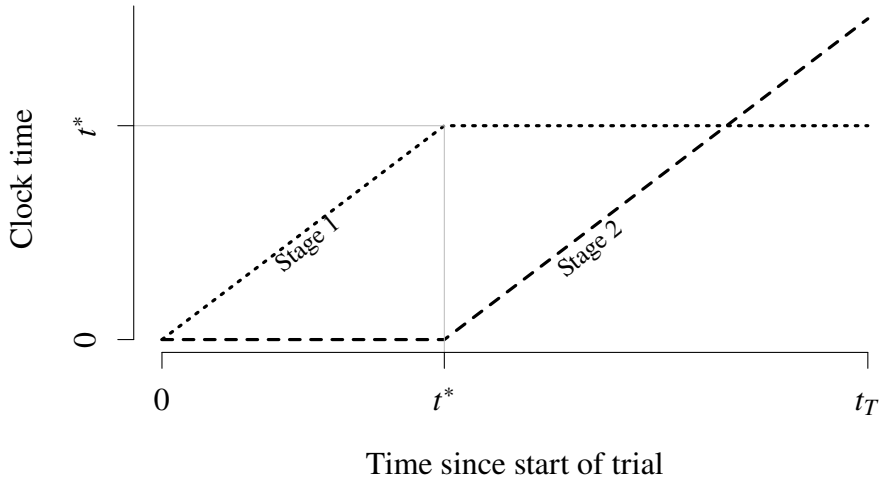


Figure 3.1: Depiction of clocks for time in the first and second stages of a longitudinal SMART. The dotted line labeled Stage 1 shows u_1 and the dashed line labeled Stage 2 shows u_2.

Consider model (2.1), a simplified version of which (i.e., without baseline covariates) is reproduced below:

$$\begin{aligned}\mu_j^{(d)}(\boldsymbol{\beta}) = \beta_0 &+ \mathbb{1}_{\{t_j \le t^*\}} \left(\beta_1 t_j + \beta_2 a_1 t_j\right) \\ &+ \mathbb{1}_{\{t_j > t^*\}} \left(t^*\beta_1 + t^*\beta_2 a_1 + \beta_3(t_j - t^*) + \beta_4(t_j - t^*)a_1\right. \\ &\left. + \beta_5(t_j - t^*)a_{2NR} + \beta_6(t_j - t^*)a_1 a_{2NR}\right),\end{aligned} \tag{2.1 revisited}$$

where $\mu_j^{(d)}$ is a marginal structural mean model for $\mathrm{E}[Y_j^{(d)}]$, the expected value of the outcome Y at time t_j had an individual been treated according to DTR $d \in \mathcal{D}$.

It will be helpful to consider the notion of separate "clocks" for each stage. Define $u_{1j} = \min\left(t_j, t^*\right)$ and $u_{2j} = \max\left(t_j - t^*, 0\right)$. u_{1j} is the clock for the first stage of the trial, starting at t_1 and continuing until t^*, at which point it remains t^* for the remainder of the trial time. The second-stage clock, u_{2j}, is zero through the first stage, then counts time since t^* through the end of the trial. A visual depiction of the clocks is given in figure 3.1.

Using the clock notation, we can easily re-write this model in terms of u_{1j} and u_{2j}:

$$\mu_j^{(d)}(\boldsymbol{\beta}) = \beta_0 + \beta_1 u_{1j} + \beta_2 u_{1j} a_1^{(d)} + \beta_3 u_{2j} + \beta_4 u_{2j} a_1^{(d)} + \beta_5 u_{2j} a_{2NR}^{(d)} + \beta_6 u_{2j} a_1^{(d)} a_{2NR}^{(d)}. \quad (3.1)$$

We can similarly re-write models (2.2) and (2.3) as

$$\mu_j^{(d)}(\boldsymbol{\beta}) = \beta_0 + \beta_1 u_{1j} + \beta_2 u_{1j} a_1^{(d)} + \beta_3 u_{2j} + \beta_4 u_{2j} a_1^{(d)} + \beta_5 u_{2j} a_{2R}^{(d)} + \beta_6 u_{2j} a_{2NR}^{(d)} + \beta_7 u_{2j} a_1^{(d)} a_{2R}^{(d)} + \beta_8 u_{2j} a_1^{(d)} a_{2NR}^{(d)}, \quad (3.2)$$

and

$$\mu_j^{(d)}(\boldsymbol{\beta}) = \beta_0 + \beta_1 u_{1j} + \beta_2 u_{1j} a_1^{(d)} + \beta_3 u_{2j} + \beta_4 u_{2j} a_1^{(d)} + \beta_5 u_{2j} \mathbb{1}_{\left\{a_1^{(d)}=1\right\}} a_{2NR}^{(d)}, \quad (3.3)$$

respectively.

As in chapter 2, we estimate $\boldsymbol{\beta}$ by solving equation (2.6):

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^{n} \sum_{d \in \mathcal{D}} \left[W^{(d)} \left(A_{i,1}, R_i, A_{i,2} \right) \cdot \left(\boldsymbol{D}^{(d)} \right)^{\top} \left(\boldsymbol{V}^{(d)} \right)^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}(\boldsymbol{\beta}) \right) \right]. \quad \text{(2.6 revisited)}$$

The (robust or sandwich) variance of $\hat{\boldsymbol{\beta}}$ is given by $\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$, where $\boldsymbol{B}$ and $\boldsymbol{M}$ are as defined in equations (2.7) and (2.8), respectively. Following proposition 2.1, the solution $\hat{\boldsymbol{\beta}}$ to the estimating equations is consistent for $\boldsymbol{\beta}^*$, the true, causal parameter vector, provided that the model is correctly specified (along with usual regularity conditions). Proposition 2.2 also still applies; $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ converges in distribution to $\mathcal{N}\left(0, \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\right)$.

3.2 Sample Size Formulae for End of Study Comparisons

We remain interested in an end-of-study comparison of two embedded DTRs which recommend different first-stage treatments. Using the potential outcomes notation developed in section 2.1, the

estimand of interest is

$$\mathrm{E}\left[Y_T^{(1,a_{2R},a_{2NR})} - Y_T^{(-1,a'_{2R},a'_{2NR})}\right]. \tag{3.4}$$

Under model (3.1), we can write this estimand as a linear combination of regression parameters $\boldsymbol{c}^\top\boldsymbol{\beta}$, where

$$\boldsymbol{c}^\top = \left(0, 0, 2u_{1T}, 0, 2u_{2T}, u_{2T}\left(a_{2NR} - a'_{2NR}\right), u_{2T}\left(a_{2NR} + a'_{2NR}\right)\right). \tag{3.5}$$

Notice that the contrast involves the stage-1 clock u_1, despite the estimand being a comparison of mean end-of-study outcomes. This is a consequence of the DTRs under study recommending different first-stage treatments. It highlights the fact that our regression-based approach to estimating quantity (3.4) uses information collected throughout the SMART: when the contrast is applied to $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$, the $2u_{1T}$ component will "pick up" variability from the first-stage of the study. We will see that this is important for understanding the behavior of the sample size formula in section 3.3.

As in chapter 2, we wish to size the SMART for a test of the null hypothesis

$$\mathrm{H}_0 : \boldsymbol{c}^\top\boldsymbol{\beta} = 0$$

to detect the alternative $\mathrm{H}_1 : \boldsymbol{c}^\top\boldsymbol{\beta} = \Delta$ with power $1 - \gamma$. The test statistic is, again,

$$Z = \frac{\sqrt{n}\boldsymbol{c}^\top\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}^\top\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}}};$$

developing a useful sample size formula depends on obtaining a tractable expression for $\boldsymbol{c}^\top\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}$, the variance of the contrast $\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}$.

We make the following working assumptions to aid in the development of the formulae:

A3.1 *Constrained conditional variability.*

(a) For all DTRs $d \in \mathcal{D}$, the element-wise difference between the matrix of variability of

responders' outcomes around the DTR mean and the marginal covariance matrix of the outcome is positive semi-definite,

$$\mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)^{\otimes 2} \mid R_i^{(a_1^{(d)})} = 1\right] - \mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] \overset{L}{\geq} \mathbf{0},$$

where $A \overset{L}{\geq} B$ implies that $A - B$ is positive semi-definite; L refers to the Loewner partial order (see appendix B.3).

(b) For all DTRs $d \in \mathcal{D}$, the element-wise difference between the marginal covariance matrix of the outcome, inflated by response probability, and the matrix of variability of responders' outcomes around the DTR mean is positive semi-definite,

$$\frac{1}{P\left(R_i^{(a_1^{(d)})} = 1\right)} \mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] - \mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)^{\otimes 2} \mid R_i^{(a_1^{(d)})} = 1\right] \overset{L}{\geq} \mathbf{0}.$$

A3.2 *Constrained conditional means.* For every DTR $d \in \mathcal{D}$ and all DTRs $d' \neq d$ with $a_1^{(d)} = a_1^{(d')}$,

$$\left(\mathrm{E}\left[\boldsymbol{Y}^{(d)} \mid R^{(a_1^{(d)})} = 1\right] - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{\mu}^{(d)} - \boldsymbol{\mu}^{(d')}\right)^{\top}$$

is "small".

A3.3 *Equal spacing.* Measurement occasions are equally-spaced in both stages, which are of fixed duration t^* and $t_T - t^*$, respectively. More specifically, we write

$$t_j = \begin{cases} t^* \cdot \frac{j-1}{T_1 - 1} & j = 1, 2, \ldots, T_1 \\ t^* + (t_T - t^*) \cdot \frac{j - T_1}{T_2} & j = T_1 + 1, \ldots, T \end{cases}.$$

We continue to make working assumption A2.2, which assumes an exchangeable within-person correlation structure marginal over response status, and constant marginal variances across time and DTR. We also restrict $\rho \in [0, 1]$.

Working assumption A3.1 is a generalization of working assumption A2.1. Indeed, A2.1 is an interpretable way to describe A3.1 in the three timepoint setting. Unfortunately, A3.1 is not particularly interpretable and may be difficult to assess intuitively. It is currently unclear what "small" means in working assumption A3.2; however, the idea is that the mean outcomes for responders are relatively close to marginal mean trajectories.

We claim that working assumption A3.3 is a realistic simplification. In appendix C, we investigate the behavior of $\omega\left(\rho, T, T_2\right)$ without equally-spaced measurements and show that, generally, the trends discussed below are maintained, though the function is noticeably less smooth. Finally, under working assumption A2.2, a negative within-person correlation would imply that individuals experience noticeable fluctuation in the outcome over time, which is typically not the case in behavioral science settings. In appendix C, we present a more general version of equation (3.7) which does not require working assumption A3.3 and is implemented in software described in chapter 4.

Under working assumptions A3.1 and A2.2, the minimum-required sample size to detect a standardized effect size δ with power at least $1-\gamma$ with a two-sided level-α test is

$$n \geq \frac{4\left(z_{1-\alpha/2}+z_{1-\gamma}\right)^2}{\delta^2} \cdot \mathrm{DE} \cdot \omega\left(\rho, T, T_2\right), \tag{3.6}$$

where z_p is the pth quantile of the standard normal distribution, DE is the design effect from table 2.3, and $\omega\left(\rho, T, T_2\right)$ is a deflation factor which accounts for longitudinal measurements and within-person correlation. See appendix B for more details, including a derivation.

As with formula (2.13), formula (3.6) can be decomposed into the standard sample size formula for a two-group comparison of means, a SMART-specific inflation factor DE, and a longitudinal deflation factor. $\omega\left(\rho, T, T_2\right)$ depends on the exchangeable within-person correlation, T, the total number of measurement times, and T_2, the number of measurements in the second stage of the SMART. Under working assumption A3.3, we can write

$$\omega\left(\rho, T, T_2\right) = \frac{f(\rho, T, T_2)}{g(\rho, T, T_2)}, \tag{3.7}$$

where

$$f(\rho, T, T_2) = 6(1-\rho)(T-1)\left(\rho(T-1)\left((T-1)T_2 - T_2^2 + 2\right) + 4T_2(T - T_2 - 1) + 2\right)$$

and

$$g(\rho, T, T_2) = (T_2+1)\left(2\left(T^2(4T_2+2) - T\left(T_2(5T_2+9)+1\right) + T_2(T_2+2)^2\right) + \right.$$
$$\left. \rho(T-1)(T-T_2-2)\left(2TT_2 + T - 2T_2(T_2+2)\right)\right)$$

A key challenge in understanding formula (3.6) is understanding the behavior of equation (3.7) as ρ and the number and timing of measurement occasions change. Note that, under working assumption A3.3, neither the numerator nor the denominator of equation (3.7) depend on the durations of each of the stages i.e., the expression is free of t^* and t_T. The time at which re-randomization occurs, t^*, and t_T, the full duration of the study, are dictated by scientific considerations. In particular, t^* is determined by the length of time needed to identify individuals as responders or non-responders to first stage treatment. When measurement occasions are equally spaced, these scientific factors do not have an impact on the sample size requirement, nor does the choice of how time is coded (for example, the study duration could be normalized to 1 without consequence for sample size).

Discovery of a more interpretable upper bound on equation (3.7) has proven intractable. Instead, we investigate the behavior of $\omega\left(\rho, T, T_2\right)$ numerically. In figure 3.2, we plot $\omega\left(\rho, T, T_2\right)$ against T and T_2 for various exchangeable within-person correlations ρ. We consider $T \in \{3, \ldots, 15\}$ and $T_2 \in \{1, \ldots, T-2\}$. Possible values T_2 are constrained so that there are always a minimum of two measurements in the first stage of the trial: one at baseline, and a second immediately prior to re-randomization at time t^*.

As expected, $\omega\left(\rho, T, T_2\right)$ is in fact a deflation factor and is bounded above by 1 on its domain. There is no deflation when $\rho = 0$ and $T_2 = 1$, i.e., $\omega\left(0, T, 1\right) = 1$ for any choice of T. In this case,

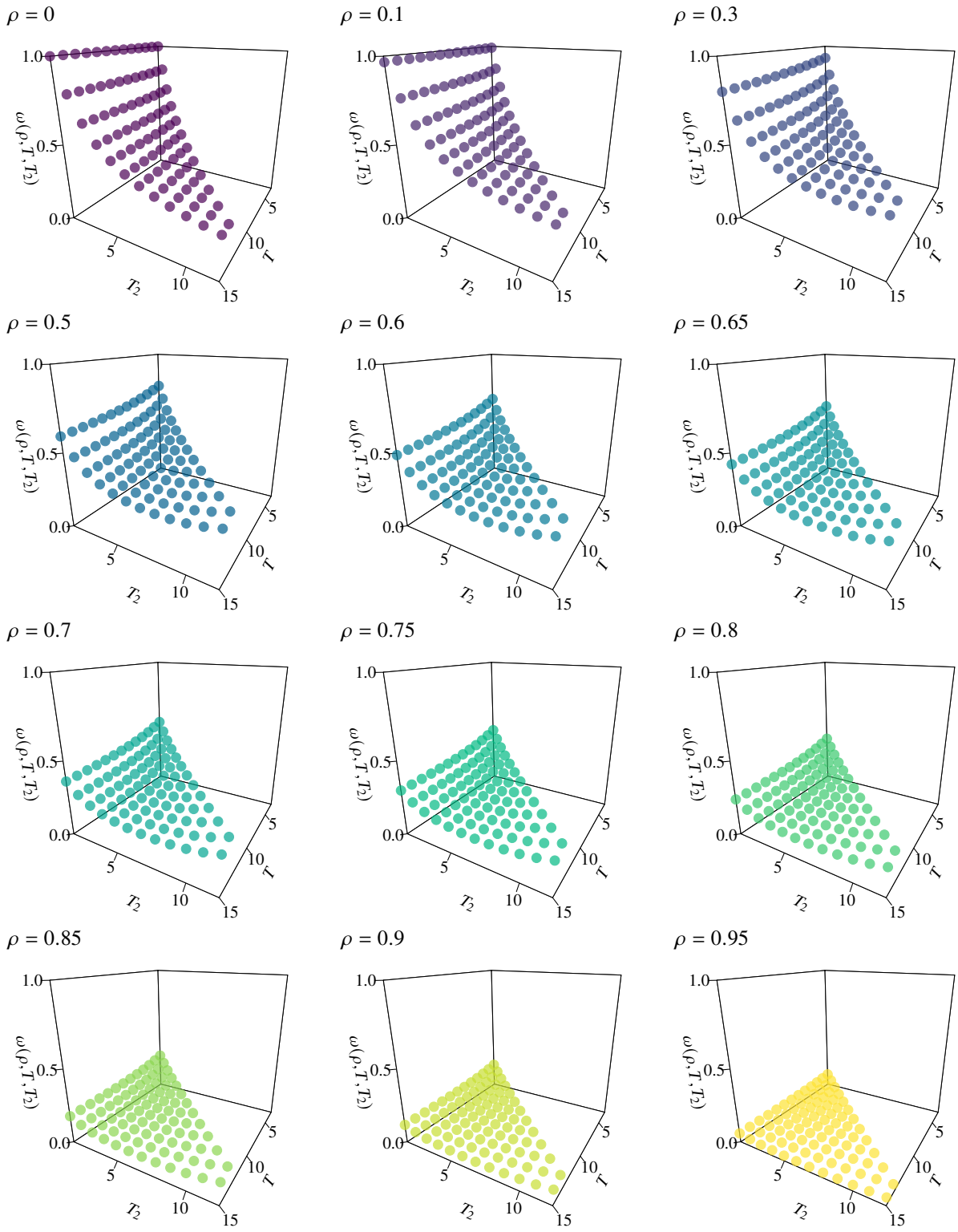


Figure 3.2: Within-person deflation factor $\omega\left(\rho, T, T_2\right)$. T is the total number of measurement occasions, T_2 of which are in stage 2 of the SMART, with $T \in \{3, \ldots, 15\}$ and $T_2 \in \{1, \ldots, T-2\}$. The function is bounded above by 1 for all ρ, T, and T_2, demonstrating that it is in fact a deflation factor. The function tends to decrease with T, and is increasingly non-monotone in T_2 for higher within-person correlations ρ.

Table 3.1: Example sample sizes for design II SMARTs with more than three measurement occasions. Provided sample sizes are chosen to detect an effect size δ with 80% power with a two-sided type-I error rate of 0.05, assuming an (exchangeable) within-person correlation ρ, and total number of measurement occasions T in a design II SMART. For each T, we set $T_2 = \lfloor T/2 \rfloor$. Response rate is 0.4 for both first-stage treatments.

		Sample Size			
δ	T	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.8$
0.3	3	559	508	358	201
	5	462	427	296	164
	7	382	358	245	134
	9	323	307	208	113
0.5	3	201	183	129	73
	5	167	154	107	59
	7	138	129	89	49
	9	116	111	75	41

the primary aim is a comparison of two means estimated by simple weighted averaging, so we do not expect to benefit from repeated measurements in any way. This is easily shown analytically; see appendix C for details. If $\rho = 1$, the deflation factor is zero, as $f(1, T, T_2) = 0$ for any choice of T and T_2. Furthermore, in the three-timepoint setting, we have $\omega(\rho, 3, 1) = (1 - \rho^2)$, the deflation factor found in formula (2.13).

For a given $\rho \in [0, 1)$, ω tends to decrease with T, though with diminishing returns. In settings with high ρ (say, $\rho \geq 0.5$) and large T (say, $T \geq 7$), ω becomes increasingly concave in T_2: for a fixed T, ω achieves a minimum over T_2 in the interior of the domain, rather than on the boundary. This is discussed in greater detail in section 3.3.1.

We provide example sample sizes in table 3.1 for design II SMARTs in which $T_2 = \lfloor T/2 \rfloor$ and where the probability of response is equal across both first-stage treatments, $r_1 = r_{-1} = 0.4$. This choice of T_2 achieves balance across the stages; if T is odd, we slightly favor stage 1 because we count baseline in stage 1. As expected, the sample sizes for SMARTs with three measurement occasions correspond to those given in table 2.4. As within-person correlation and the total number of measurement occasions increase, the sample size requirement decreases, though at a decreasing rate. For fixed ρ, the benefits of adding T diminish for higher values of T.

This pattern, in which higher within-person correlation and more measurement occasions

decreases sample size requirements, is a well-established phenomenon. Raudenbush and Xiao-Feng (2001) found that the sample size requirement for detecting a between-group polynomial effect decreases as the frequency of measurement occasions increases. Zhang and Ahn (2011b) found similar results for the comparison of time-averaged responses in a two-arm randomized trial, noting especially that the reduction in sample size achieved by adding measurement occasions diminishes as T increases. As in chapter 2, larger values of ρ are generally associated with smaller sample size requirements as well: because the correlation is within-person and the analysis is between groups, efficiency improves with larger ρ (Hedeker, Gibbons, and Waternaux 1999).

3.2.1 Simulation Study

We investigate performance of formula (3.6) using a simulation study. We hypothesize that the formula achieve nominal power or greater (i.e., be conservative) when working assumptions A3.1 to A2.2 are satisfied.

Data were generated using the longsmart R package as described in chapter 4. We focus on a design II SMART with either 3 or 5 measurement occasions. As in section 2.4, results from ENGAGE informed the parameter selections for the marginal mean and variance model, but the results are not representative of that trial. Potential response status was generated using a threshold-based criterion such that simulated individuals with potential outcomes $Y_{T_1}^{(d)}$ greater than some fixed value $\kappa_{a_1^{(d)}}$ were identified as potential responders; otherwise the individual was a non-responder. For each $d \in \mathcal{D}$, $\kappa_{a_1^{(d)}}$ was chosen to achieve the specified response probability. All errors are assumed to be normally-distributed.

Results of 1000 Monte Carlo simulations are given in table 3.2. In general, we achieve 80% target power or higher for $T = 3$ and $T = 5$ for small response rates. When one or more response rates is 0.6 and $T = 5$, empirical power tends to be significantly less than 0.8, but not worryingly so. In these scenarios, it is non-trivial to find parameters which do not violate working assumption A3.1(a); the results show reduced power as a result of this assumption being violated. Further investigation is required to evaluate the performance of the sample size method under a

Table 3.2: Sample sizes and empirical power results for **??** SMARTs with three or more measurement occasions. Sample sizes were chosen for an end-of-study comparison of DTRs $(1, 0, 1)$ and $(-1, 0, -1)$ in a design II SMART with three or five measurement occasions. δ is the true standardized effect size as defined in equation (2.12), r is the common probability of response to first-stage treatment, and ρ is the true exchangeable within-person correlation. n is computed using formula (2.13) with $\alpha = 0.05$ and $\gamma = 0.2$. Results are the proportion of 1000 Monte Carlo simulations in which we reject $H_0 : \boldsymbol{c}^\top \boldsymbol{\theta} = 0$ at the 5% level.

				$T = 3$		$T = 5, T_2 = 2$	
δ	ρ	r_1	r_{-1}	n	Power	n	Power
0.3	0	0.4	0.4	559	0.804	462	0.788
		0.4	0.6	524	0.813	434	0.767*
		0.6	0.4	524	0.804	434	0.760*
		0.6	0.6	489	0.825	405	0.758*
	0.3	0.4	0.4	508	0.803	427	0.804
		0.4	0.6	477	0.809	400	0.770*
		0.6	0.4	477	0.790	400	0.794
		0.6	0.6	445	0.810	373	0.770*
	0.6	0.4	0.4	358	0.833*	296	0.818
		0.4	0.6	335	0.799	278	0.788
		0.6	0.4	335	0.825	278	0.736*
		0.6	0.6	313	0.818	259	0.738*
	0.8	0.4	0.4	201	0.858*	164	0.842*
		0.4	0.6	189	0.860*	154	0.789
		0.6	0.4	189	0.862*	154	0.817
		0.6	0.6	176	0.815	144	†

* Statistically significantly different from 0.8 at the 5% level.
† Our data generative model could not accommodate this scenario.

variety of scenarios, including when assumptions are violated in a principled way.

3.3 Cost Considerations for Longitudinal SMARTs

Managing trial costs are a key reality of designing experiments, and this is no different for SMARTs. Here, we develop tools which can aid clinicians and applied statisticians in choosing both sample size and the number of measurement occasions to achieve minimum total cost, subject to a constraint on statistical power.

There are a variety of costs involved in conducting any study, including personnel and staffing

costs, costs of administering the interventions, data management costs, etc. For our purposes, we consider only costs related to recruiting individuals into the trial and measuring their outcomes. Note that we consider only costs directly related to participants: overhead and other costs (e.g., data management, salaries, etc.) are taken as sunk.

We examine this problem given a specific scientific context; that is, we consider the intervention options in the SMART, as well as the desired type-I error rate, target power, and target standardized effect size, to be fixed *a priori*. This also fixes ρ and the response rates r_1 and r_{-1}, as these are characteristics of the interventions in the trial and not design choices.

3.3.1 Minimizing Recruitment Costs

Suppose an investigator is interested primarily in minimizing the cost of recruiting individuals into a longitudinal SMART, or, equivalently, minimizing the sample size requirement. This may be of particular interest when the target population is hard to reach, for example. As seen in figure 3.2, the sample size computed by formula (3.6) decreases with the number of measurement occasions T; the most naive strategy for minimizing sample size is to measure the outcome as many times as is feasible. This is not a very practical recommendation, however: a large number of measurement occasions may be quite burdensome to participants, potentially leading to dropout. We therefore proceed assuming that the investigator has chosen T *a priori*.

The multi-stage nature of SMARTs introduces a question of how best to allocate measurement occasions across stages. For instance, formula (3.6) is for an end-of-study comparison; because of this, we expect that adding measurement occasions in the second stage of the trial will yield greater reductions in sample size requirements than will adding measurements in the first stage. Here, we investigate the allocation of measurement occasions in the first and second stages of a SMART which produces the smallest sample size requirement (and therefore the smallest total recruitment

cost) for a fixed T. Specifically, we solve the following integer optimization problem:

$$\begin{aligned} \underset{T_2}{\text{minimize}} \quad & \frac{4\left(z_{1-\alpha/2} + z_{1-\gamma}\right)^2}{\delta^2} \cdot \text{DE} \cdot \omega\left(\rho, T, T_2\right) \\ \text{subject to} \quad & T_2 \in \{1, \ldots, T-2\}. \end{aligned} \tag{3.8}$$

As above, we require at least two measurements in stage 1. Minimizing the sample size requirement for a fixed design is therefore equivalent to minimizing $\omega\left(\rho, T, T_2\right)$, so optimization problem (3.8) becomes

$$\begin{aligned} \underset{T_2}{\text{minimize}} \quad & \omega\left(\rho, T, T_2\right) \\ \text{subject to} \quad & T_2 \in \{1, \ldots, T-2\}. \end{aligned} \tag{3.9}$$

We call the solution to optimization problem (3.9) T_2^{n}. Since equation (3.7) is free of stage duration under working assumption A3.3, we set $t_1 = 0$, $t^* = 1$, and $t_T = 2$ without loss of generality. Optimization problem (3.9) is difficult to solve analytically. However, the feasible set of T_2 is relatively small, so solutions can be quickly and easily obtained using grid search. For a given ρ, T, and T_2, we generate equally-spaced measurement occasions $\boldsymbol{t}$, compute $\omega\left(\rho, T, T_2\right)$, and find T_2^{n} as the choice of T_2 which yields the smallest value.

Our exploration of $\omega\left(\rho, T, T_2\right)$ in section 3.2 provides evidence for our earlier conjecture about T_2^{n}: as T_2 increases toward $T-2$ (i.e., as more of the total measurement occasions in the trial are placed in the second stage) ω tends to decrease. However, careful inspection of figure 3.2 reveals that, for larger ρ and T, the deflation factor $\omega\left(\rho, T, T_2\right)$ achieves a minimum at $T_2 < T-2$. This suggests that it is sub-optimal to measure the outcome as many times as possible in the second stage: some balance across stages is favored.

In figure 3.3, we plot T_2^{n} against T for various choices of ρ. In each plot, we include the lines $T_2^{\text{n}} = T-2$ (upward-facing triangles) and $T_2^{\text{n}} = \lfloor T/2 \rfloor$ (downward-facing triangles). Recall that T is fixed *a priori*. The former line represents the design strategy of including as many measurement occasions in stage 2 as possible; the latter corresponds to balancing measurements across stages.

For smaller ρ and/or smaller T, $T_2^{\text{n}} = T-2$: the optimal allocation strategy is to maximize the

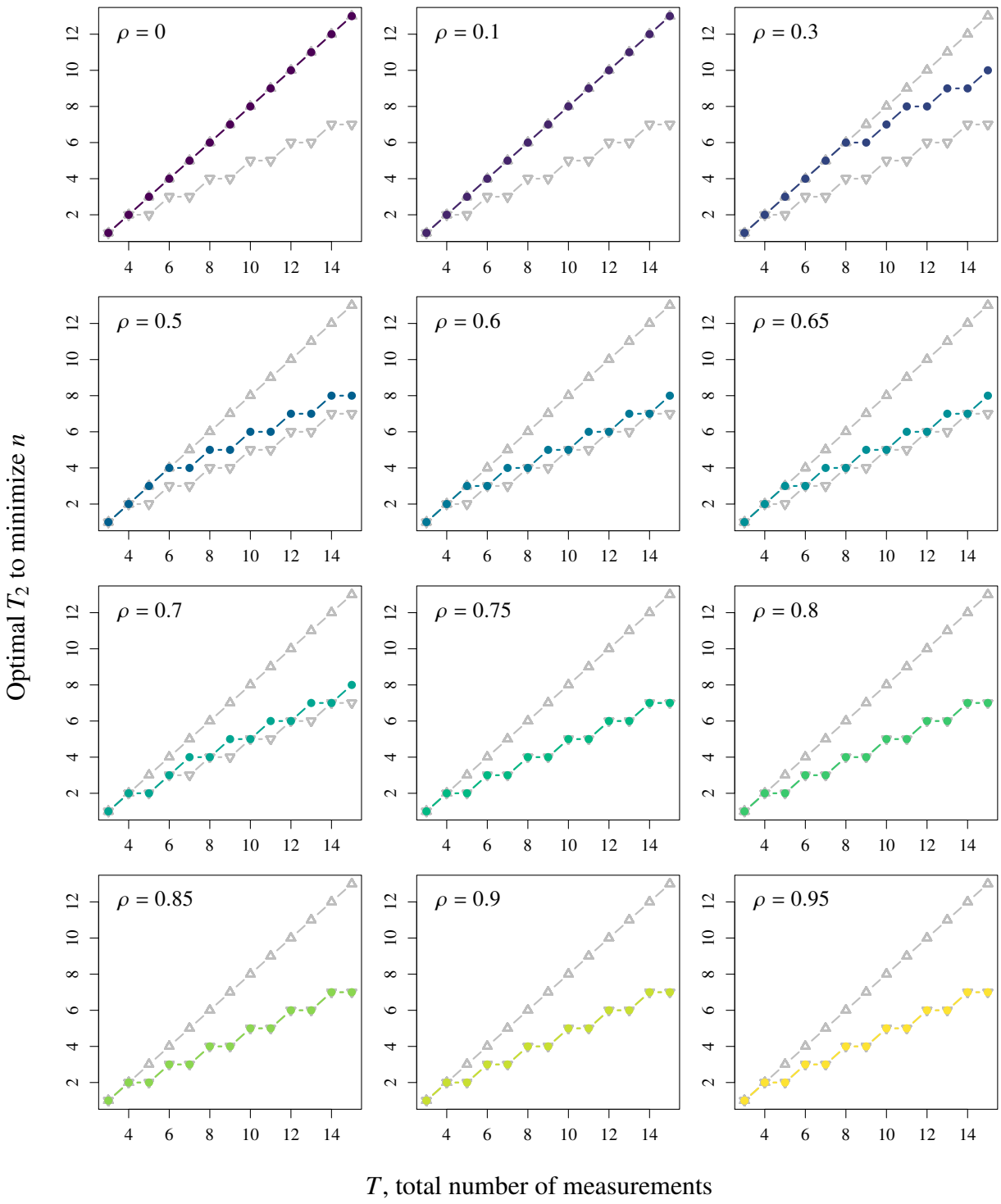


Figure 3.3: Optimal allocation of equally-spaced measurement occasions in stage 2 to minimize sample size for various within-person correlations ρ. The trajectory with filled circles represents T_2^n, the optimal allocation, for a given T. The trajectory with upward-facing triangles is the line $T_2 = T - 2$, the maximum number of measurements allowed in stage 2. The trajectory with downward-facing triangles is $T_2 = \lfloor T/2 \rfloor$, representing balanced allocation of measurements across stages. For $T > 4$, it is optimal to allocate more measurements to stage 1 than to stage 2.

number of measurement occasions in the second stage of the trial. For $\rho = 0$, it is easy to show that the partial derivatives of ω in T and T_2 are strictly negative for $T \geq 3$ and $T \leq T - 2$, implying that it is always better to add measurements in stage 2. This remains true for small, nonzero values of ρ.

Recall that model (3.1) is unsaturated for $T > 3$ and $T_2 > 1$. For an end-of-study comparison, we believe that gains in power relative to the three-measurement setting are achieved primarily through smoothing the mean model in the second stage. The sample size requirement is directly proportional to the variance of the contrast (see appendix B). When $\rho = 0$, any efficiency gain in estimating the contrast (meaning, any reduction in the sample size requirement) cannot be attributed to repeated measurements on the same individual; instead, the reduction must be derived from improved model fit. As the number of measurement occasions in stage 2 increases, we are able to estimate the mean end-of-study difference with greater precision. We believe this is driven primarily by increased effective sample size: each additional measurement occasion provides more data with which to estimate regression parameters.

For higher values of ρ, it becomes sub-optimal to maximize the number of measurements in stage 2 as the total number of measurements T increases. This is reflected in the concavity seen in figure 3.2, as well as the deviation of the filled circles from the $T = T - 2$ trajectory in figure 3.3. For larger ρ, each additional measurement of an individual yields more information about that individual's outcome trajectory, which in turn lowers the variance of the model parameters. Therefore, we expect diminishing returns of adding (exchangeable) measurements in the second stage of the SMART for moderate to large ρ. Because the contrast (3.5) involves first-stage quantities, it becomes advantageous to also smooth the model in the first stage of the SMART for large ρ and large T.

3.3.2 Minimizing Per-Patient Costs

Often, concerns about large sample sizes for trials are related to overall study costs: recruitment of participants can be expensive, and so larger sample sizes correspond to more expensive trials. As

we have seen, the repeated measures in a longitudinal SMART allow for gains in efficiency which can reduce the sample size requirement. If an investigator's interest is in reducing overall study expenditures, we may be able to trade off between sample size and the number of measurement occasions in order to achieve a minimum cost.

Suppose the cost of recruiting one individual into the study is C_R, and let C_1 and C_2 be the costs of measuring their outcomes in the first and second stages of the SMART, respectively. For a SMART with n participants and T total measurement occasions, T_2 of which are in stage 2, the total per-participant cost of the trial is

$$C(n, T, T_2) = n\left(C_R + (T - T_2)C_1 + T_2C_2\right). \tag{3.10}$$

C_R, C_1, and C_2 can include a variety of participant-specific costs, including those related to advertising for recruitment and incentives for measuring the research outcome. While equation (3.10) does not directly accommodate variable incentives within each stage, investigators could average these costs across measurement occasions.

Our goal is to solve the optimization problem

$$\begin{aligned} \underset{n,T,T_2}{\text{minimize}} \quad & C(n, T, T_2) \\ \text{subject to} \quad & 1 - \gamma \geq 0.8 \\ \text{subject to} \quad & T \in \left\{3, 4, \ldots, T^{\max}\right\}, \\ & T_2 \in \{1, 2, \ldots, T - 2\} \end{aligned} \tag{3.11}$$

where $1 - \gamma$ is the power of the end-of-study comparison of embedded DTRs for which the SMART is sized. The power constraint is satisfied by choosing n according to formula (3.6). Therefore, we

can re-write optimization problem (3.11) as

$$\begin{aligned} \underset{T,T_2}{\text{minimize}} \quad & \left\lceil \frac{4\left(z_{1-\alpha/2}+z_{0.8}\right)^2}{\delta^2} \cdot \text{DE} \cdot \omega(\rho,T,T_2) \right\rceil \left(C_R + (T-T_2)C_1 + T_2C_2\right) \\ \text{subject to} \quad & T \in \left\{3,4,\ldots,T^{\max}\right\}, \\ & T_2 \in \left\{1,2,\ldots,T^{\max}-2\right\}. \end{aligned} \tag{3.12}$$

This is a nonlinear integer program for which it is difficult to find a general solution. As in section 3.3.1, we investigate numerically across a variety of scenarios. We also make the simplification of equally-spaced measurement occasions in both stages, as per working assumption A3.3.

Notice first that the objective function in optimization problem (3.12) involves a number of terms which are constant in T and T_2. The cost-optimal number of measurement occasions, therefore, does not depend on features of the hypothesis test (significance level, target power), the target effect size, response rate, or even the design of the SMART (through the design effect DE). Certainly the total costs required will change based on these quantities since they change the sample size requirement; however, in order to understand the behavior of the objective function with regard to the solutions of optimization problem (3.12), we need only consider ρ, T, and T_2. Because the objective function uses the form of n from formula (3.6), the power constraint is guaranteed to be satisfied under working assumptions A3.1 to A3.3.

The feasible set for optimization problem (3.12) is theoretically unbounded; practically, it is not. Most SMARTs with longitudinal outcomes, to our knowledge, employ a relatively small number of measurement occasions. Three measurement occasions are common (Naar-King et al. 2016), as are five (McKay et al. 2015; Kilbourne et al. 2018). To our knowledge, SMARTs with more than 15 measurement occasions are rare, and often driven by questions about the effects of interventions on relatively fine time scales. Typically, end-of-study comparisons do not motivate this type of (intensive longitudinal) data collection, so we restrict our focus to SMARTs in which the outcome is measured 15 times or fewer (Walls and Schafer 2006).

In table 3.3, we compile the total number of measurement occasions T^{cost} and the number of occasions in stage 2, T_2^{cost}, which solve optimization problem (3.12) for a variety of costs and within-person correlations ρ. We consider $T \in \{3, \ldots, 15\}$. Because the objective function is linear in C_R, C_1, and C_2, their exact values only affect the value of the minimum cost, not the values of the minimizers. We therefore focus on the costs of recruiting one individual relative to that of measuring an individual once by considering the ratios C_R/C_1 and C_R/C_2.

When it is relatively cheap to recruit, meaning that C_R is similar to C_1 and C_2, the objective function strongly favors recruiting more individuals in favor of adding measurement occasions. When C_R, C_1, and C_2 are similar, T^{cost} is often 3: the most cost-effective way to achieve the target power in a longitudinal SMART is to measure the outcome infrequently in a larger number of participants. When $\rho = 0$, T^{cost} and T_2^{cost} are the maxima of their domains when recruitment becomes even slightly more expensive than second-stage measurements. For even small non-zero values of ρ, the preference for $T^{\text{cost}} = 3$ is maintained generally, except when C_R is two to four times C_2 and similar to C_1. The majority of remaining settings examined favor maximizing the total number of measurement occasions.

The solutions in table 3.3 suggest that there is little middle ground in the trade-off between sample size and number of measurement occasions when optimizing for cost. In most situations, the solution to optimization problem (3.12) is either $T^{\text{cost}} = 3$, the minimum, or $T^{\text{cost}} = 15$, the maximum number of measurements we are willing to consider. For small-to-moderate non-zero within-person correlations ρ in which C_R is not too much larger than C_1 or C_2, the trade-off is more balanced between n and T: T^{cost} is neither the maximum nor minimum of its domain.

Curiously, the allocation of measurement occasions across stages, i.e., T_2^{cost} exhibits more variability than does T^{cost}. Note that, in general, $T_2^{\text{cost}} \neq T_2^{\text{n}}$; i.e., the solutions to optimization problems (3.9) and (3.12) do not always coincide. Recall that T_2^{n} minimizes the sample size requirement given equal spacing as in working assumption A3.3 and fixed ρ and T. The fact that $T^{\text{cost}} \neq T^{\text{n}}$ for $\rho \neq 0$ makes the n-versus-T trade-off more apparent: it is not always cost-optimal to minimize sample size, even when recruitment is much more expensive than measurement. This is

Table 3.3: Total number of measurement occasions T^{cost} and number of second-stage measurements T_2^{cost} (in parentheses) which minimize trial cost for a design II SMART. Solutions are obtained by solvingoptimization problem (3.12) for $T \in \{3, \ldots, 15\}$ and $T_2 \in \{1, \ldots, T-2\}$. C_R is the cost of recruiting one participant into the SMART, C_1 the cost of measuring one participant once in stage 1, and C_2 the cost of measuring one participant once in stage 2. As C_R becomes large relative to C_1 and, particularly, C_2, costs are minimized by measuring the outcome as many times as is feasible.

			T^{cost} (T_2^{cost})			
C_R	C_1	C_2	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
1	1	1	3 (1)	3 (1)	3 (1)	3 (1)
	1	0.75	15 (13)	3 (1)	3 (1)	3 (1)
	1	0.5	15 (13)	5 (3)	3 (1)	3 (1)
	0.75	1	3 (1)	3 (1)	3 (1)	3 (1)
	0.5	1	3 (1)	3 (1)	3 (1)	3 (1)
2	1	1	15 (13)	3 (1)	3 (1)	3 (1)
	1	0.75	15 (13)	3 (1)	3 (1)	3 (1)
	1	0.5	15 (13)	7 (5)	6 (4)	3 (1)
	0.75	1	15 (13)	3 (1)	3 (1)	3 (1)
	0.5	1	3 (1)	3 (1)	3 (1)	3 (1)
5	1	1	15 (13)	7 (5)	5 (3)	15 (7)
	1	0.75	15 (13)	8 (6)	15 (9)	15 (8)
	1	0.5	15 (13)	12 (10)	15 (10)	15 (9)
	0.75	1	15 (13)	6 (4)	15 (8)	15 (7)
	0.5	1	15 (13)	15 (8)	15 (7)	15 (6)
10	1	1	15 (13)	15 (10)	15 (8)	15 (7)
	1	0.75	15 (13)	15 (11)	15 (9)	15 (8)
	1	0.5	15 (13)	15 (12)	15 (10)	15 (9)
	0.75	1	15 (13)	15 (9)	15 (8)	15 (7)
	0.5	1	15 (13)	15 (8)	15 (7)	15 (7)
100	1	1	15 (13)	15 (10)	15 (8)	15 (7)
	1	0.75	15 (13)	15 (10)	15 (8)	15 (8)
	1	0.5	15 (13)	15 (10)	15 (8)	15 (8)
	0.75	1	15 (13)	15 (10)	15 (8)	15 (7)
	0.5	1	15 (13)	15 (10)	15 (8)	15 (7)

evident in figure 3.4: as C_R grows relative to C_1 and C_2, or for higher within-person correlation, the objective function becomes increasingly concave.

Recall that our intuition around the results in section 3.3.1 and figure 3.3 is that the benefit of allocating more measurements in the second stage is primarily related to smoothing: the efficiency gained by better estimating the second-stage trajectory is important for reducing the sample size requirement for small ρ. For larger ρ, we favor more balanced allocation of measurements across stages, presumably due to the within-person correlation providing more information about the trajectories in each stage. In table 3.3, the benefits of the within-person correlation become apparent at lower ρ as well: the cost optimization favors balance across stages.

3.4 Practical Implications for Designing Longitudinal SMARTs

Conversations between statisticians and investigators about sample size can be challenging, for a variety of reasons. Boen and Zahn (1982), in describing their experiences as statistical consultants, remark that investigators typically have firm upper (and sometimes lower) bounds on sample size which are often dictated by personal experience, disciplinary traditions, or budget constraints. Financial considerations are often of particular importance At the same time, that a study is sized to detect a relevant effect with at least 80% power is often an implicit or explicit requirement for many funding agencies.

The methods presented in section 3.3 allow for a reframing of conversations about sample size for longitudinal SMARTs by introducing the ability to balance sample size and the number of measurement occasions to achieve specified power while minimizing cost. This requires elicitation of more parameters on the part of the statistician, which is often nontrivial but opens the door to more collaborative discussions of trial design (Lenth 2001).

Consider an investigator who wishes to design a longitudinal SMART powered for the primary aim of comparing two embedded DTRs which recommend different first-stage treatments. As a secondary aim, the investigator wishes to use the longitudinal outcome to examine mean trajectories

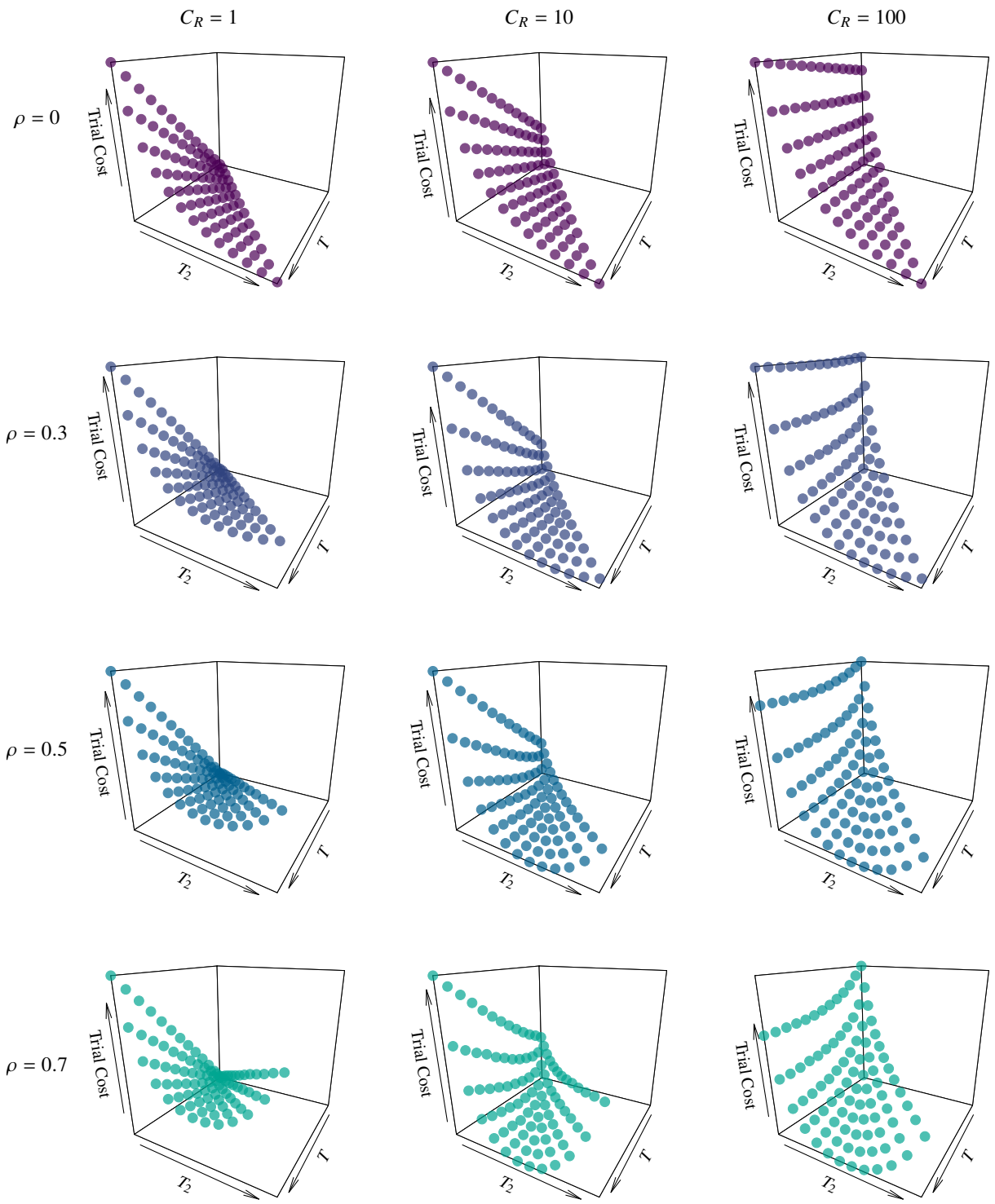


Figure 3.4: Scaled objective function $\omega\left(\rho, T, T_2\right) \cdot C(n, T, T_2)$ for minimizing per-participant trial costs. Each plot is of the (scaled) objective function in optimization problem (3.12) for various choices of within-person correlation ρ and recruitment cost C_R. For all scenarios, stage-1 and stage-2 measurement costs are set to 1: $C_1 = C_2 = 1$. For larger ρ, the objective function becomes increasingly concave over the optimization domain (much more so than in figure 3.2). T_2^{cost} tends to be smaller than T_2^{n} from figure 3.3.

over time for each of the DTRs, perhaps to investigate delayed effects of treatments (Nahum-Shani et al. 2020). Their initial plan is to measure the outcome four times: once at baseline, once at the end of the first stage, and twice in the second stage.

Formula (3.6) was designed to involve only quantities which are relatively easy to find from existing literature or domain knowledge, so the statistician's work with this investigator may begin fairly typically. The statistician will need to elicit a target effect size, estimated probability of response to first-stage treatments, and within-person correlation. Formula (3.6) enables easy computation of sample size given these quantities. However, the result may be practically infeasible due to budget constraints or other realities of running a trial.

The trade-offs discussed in section 3.3 allow the statistician to work collaboratively with the investigator to modify the number and timing of measurement occasions to minimize trial costs. By eliciting additional information such as the maximum number of times the investigator would be willing to measure the outcome, approximate costs of those measurements, and the cost of recruitment, the statistician can shift the conversation from one about sample size to one about trial design more broadly.

Table 3.3 and figure 3.4 suggest that, for most realistic scenarios in which recruitment is much more expensive than measurement, overall costs are lower in trials with more measurement occasions. By working with the investigator to identify the largest number of measurement occasions they would consider, the statistician can address the investigator's feasibility concerns with regard to sample size and total cost of the trial, while still maintaining target power. This will allow the statistician and investigator to collaboratively maximize what can be done with the trial's finite resources.

This is explicitly a different framing of the cost problem than that considered by, for example, Bloch (1986), Liu and Colditz (2017), and Zhang and Ahn (2011a). These authors all examine financial considerations in randomized trials, but with an eye towards maximizing statistical power given a fixed budget. This approach has the advantage of recognizing and working within monetary constraints provided by the investigator, but may under-power the trial. Our approach acknowledges

the reality that achieving 80% power is an important convention in trial design and imposes that as the primary constraint. We do not guarantee that the solutions to optimization problem (3.12) will yield a total cost that is lower than a given budget.

The cost function equation (3.10) is easily extensible. C_1 and C_2 might, for instance, incorporate some notion of burden on behalf of the participant or clinical staff. The function could also be designed to incorporate non-constant costs of measurement within stages, which might be the case if incentives for participants change over time. Furthermore, while an investigator may be willing to measure the outcome $T^{\max}$ times, they may have a strong preference against including more than $T^{\text{pref}} < T^{\max}$ occasions. This preference can be incorporated by imposing a penalty on $T > T^{\text{pref}}$ in the cost function. In chapter 4, we describe how alternative cost functions can be accommodated via software.

CHAPTER 4

Software for Designing Longitudinal SMARTs

As interest in SMARTs grows, and as the design is extended to accommodate more complex scientific settings, there is a clear need for the development of software that would allow domain scientists and applied statisticians to perform simulation-based sample size and power calculations. An important challenge here is to make the software general enough to be used across a number of different types of SMART designs (e.g., three stages of randomization), yet not so flexible that it is difficult to use. The benefit of this is the ability to examine the power for various different scientific questions given a single data generative model and for many other types of SMARTs.

In this chapter, we develop a framework for simulating data from a SMART with a continuous longitudinal outcome and introduce the R package longsmart, which implements the data generative model as well as the analytic methods described in chapter 3. We discuss the specifics of the generative model in section 4.1, then illustrate how longsmart can be used to design and simulate longitudinal SMARTs in section 4.2

4.1 A Data-Generative Procedure for Longitudinal SMARTs

The analytic methods for SMARTs described in chapters 2 and 3 allow for inference marginal over the tailoring variable. Models (2.1) to (2.3) can be used to estimate mean potential outcomes $\mathrm{E}[Y^{(a_1,a_{2R},a_{2NR})}]$ for a DTR (a_1, a_{2R}, a_{2NR}) averaging over response status. However, data are observed conditionally on response: a single participant can only be a responder or a non-responder. This is a key challenge in developing data-generative models for SMARTs.

When a continuous outcome is observed once at the end of a SMART, this challenge is easily overcome using the laws of total probability and variance; similarly for other types of outcomes (Ogbagaber, Karp, and Wahed 2016; Kidwell et al. 2018). Generating longitudinal data presents a more complex challenge: the outcome is observed across stages, the tailoring variable is likely related to previously-observed outcome measurements, and it may be necessary to precisely control the marginal covariance structure of the outcomes. The challenge related to the tailoring variable, in particular, is non-trivial.

The primary goal of the data generative procedure described here is to enable simulation of longitudinal SMARTs with known marginal means and covariances for each embedded DTR. Because the sample size methods described in chapters 2 and 3 assume a particular marginal covariance structure, it is important for testing these methods that the data generative model satisfies this assumption. The procedure requires a mean model $\boldsymbol{\mu}^{(d)}(\boldsymbol{\beta})$ for each embedded DTR d, a target marginal variance structure $\boldsymbol{\Sigma}^{(d)}$ for each embedded DTR, as well as second-stage means and covariance "components" (see below) for responders to each first-stage treatment. The procedure also requires methods for identifying responders and non-responders, and for computing means and covariances of stage-1 outcomes conditional on response. We write $\nu_j^{(d)}(r) = \mathrm{E}[Y_j^{(d)} \mid R^{(a_1^{(d)})} = r]$ and $\boldsymbol{\Xi}^{(d)}(r) = \mathrm{Var}\left(\boldsymbol{Y}^{(d)} \mid R^{(a_1^{(d)})} = r\right)$, with (j,k)th element $\xi_{jk}^{(d)}$.

In order to develop a realistic data generative model, we attempt to follow as closely as possible the way in which data is accumulated in a SMART. Let $\boldsymbol{Y}_{i,j:k}$ be the vector of the ith individual's outcomes at measurement occasions $j, j+1, \ldots, k$, for $k > j$. Recall from section 2.2.1 that the data collected from the ith participant in the SMART is of the form

$$\left(\boldsymbol{X}_i, Y_{i,1}, A_{i,1}, \boldsymbol{Y}_{i,2:T_1}, R_i, A_{i,2}, \boldsymbol{Y}_{i,T_1+1:T}\right),$$

where $T_1 = T - T_2$ is the number of measurement occasions in stage 1. For simplicity, we currently ignore baseline covariates $\boldsymbol{X}$; future work will extend the generative model described below to accommodate baseline covariates.

4.1.1 Simulation of Potential Outcomes

To guarantee that consistency (identifiability assumption I2; see appendix A) is satisfied, we generate potential outcomes for each "participant" in the trial. Consider a marginal mean model $\mu_t^{(d)} = \mathrm{E}[Y_t^{(d)}]$ for some DTR $d \in \mathcal{D}$ ($\mathcal{D}$ as defined in section 2.2.2). Suppose we wish to generate data with an arbitrary covariance structure $\boldsymbol{\Sigma}^{(d)}$ such that $\boldsymbol{\Sigma}_{j,j}^{(d)} = \mathrm{Var}(Y_{i,j}^{(d)}) = (\sigma_j^{(d)})^2$ and $\boldsymbol{\Sigma}_{j,k}^{(d)} = \mathrm{Cov}(Y_{i,j}^{(d)}, Y_{i,k}^{(d)}) = \rho_{jk}^{(d)}\sigma_j^{(d)}\sigma_k^{(d)}$ for $j \neq k$.

In the first stage, we can express the ith individual's potential outcome under DTR d at time j as

$$Y_{i,j}^{(d)} = \mu_j^{(d)} + \sum_{k=1}^{j-1} b_{jk}^{(d)} \left(Y_{i,k}^{(d)} - \mu_k^{(d)} \right) + \epsilon_{i,j}^{(d)}, \quad j = 1, \ldots, T_1, \tag{4.1}$$

where $b_{jk}^{(d)}$ are constants chosen to achieve the desired marginal covariance structure and $\epsilon_{j,i}$ is mean-zero noise with variance

$$\mathrm{Var}\left(\epsilon_{i,j}^{(d)}\right) = \left(\sigma_j^{(d)}\right)^2 - \left(\boldsymbol{b}_j^{(d)}\right)^\top \boldsymbol{\Sigma}_{1:j-1,1:j-1}^{(d)} \boldsymbol{b}_j^{(d)}. \tag{4.2}$$

Note that for the baseline measurement ($j = 1$), we set the sum from $k = 1$ to 0 to zero.

Equation (4.1) induces within-person correlation by explicitly making each potential outcome a function of previous potential outcomes. We choose $\boldsymbol{b}_j^{(d)} = \left(b_{j,1}^{(d)}, \ldots, b_{j,j-1}^{(d)}\right)^\top$ to achieve the desired correlation structure; we discuss this in more detail below. Note that the choice of $\boldsymbol{c}_j^{(d)}$ does not affect the mean of $Y_{i,j}^{(d)}$, as the summand in equation (4.1) is mean-zero.

Potential response status is generated as $R_i^{(a_1)} = g_{a_1}(\boldsymbol{Y}_{i,1:T_1}^{(d)})$, where $g_{a_1} : \mathbb{R}^{T_1} \to \{0, 1\}$ is any function of stage-1 outcomes that returns 1 if the individual is a responder and 0 otherwise. As an example, consider a "threshold"-based response function, such that

$$R_i^{(a_1^{(d)})} = g_{a_1}(\boldsymbol{Y}_{i,1:T_1}^{(d)}) = \mathbb{1}_{\left\{Y_{i,T_1}^{(d)} > \kappa_{a_1}\right\}}. \tag{4.3}$$

Here, an individual is a (potential) responder to first-stage treatment a_1 if their (potential) outcome at time $t_{T_1} = t^*$ exceeds some threshold κ_{a_1} and a non-responder otherwise. We discuss this in more

detail in section 4.1.3.

In the second stage, we generate data condtitionally on response status, such that, for $j = T_1 + 1, \ldots, T$,

$$Y_{i,j}^{(d)} = \nu_j^{(d)}(R_i^{(a_1^{(d)})}) + \sum_{k=1}^{j-1} b_{jk}^{(d)} \left(Y_{i,k}^{(d)} - \nu_k^{(d)}(R_i^{(a_1^{(d)})}) \right) + \zeta_{i,j}^{(d)} \left(r^{(a_1^{(d)})} \right), \tag{4.4}$$

where $\zeta_{i,j}^{(d)}(R_i^{(a_1^{(d)})})$ has mean zero and variance

$$\text{Var}\left(\zeta_{i,j}^{(d)}(r) \right) = \left(\xi_{j,j}^{(d)}(r) \right)^2 - \left(\boldsymbol{b}_j^{(d)} \right)^\top \boldsymbol{\Xi}^{(d)}(r)_{1:j-1,1:j-1} \boldsymbol{b}_j^{(d)}. \tag{4.5}$$

Note that, given the marginal mean model $\boldsymbol{\mu}^{(d)}$, we need only specify either $\boldsymbol{\nu}^{(d)}(1)$ or $\boldsymbol{\nu}^{(d)}(0)$ for each first-stage treatment; the other is fixed by the law of total expectation:

$$\boldsymbol{\mu}^{(d)} = P\left(R^{(a_1^{(d)})} = 1 \right) \boldsymbol{\nu}^{(d)}(1) \left(1 - P\left(R^{(a_1^{(d)})} = 1 \right) \right) \boldsymbol{\nu}^{(d)}(0). \tag{4.6}$$

It remains to show how to choose $\boldsymbol{b}_j^{(d)}$ to achieve the desired marginal covariance structure. We do this in stage 1 (i.e., $j = 1, \ldots, T_1$) by solving

$$\boldsymbol{\Sigma}_{1:j-1,1:j-1}^{(d)} \boldsymbol{b}_j^{(d)} = \boldsymbol{\Sigma}_{1:j-1,j}^{(d)}. \tag{4.7}$$

As an example, if $\boldsymbol{\Sigma}^{(d)} = (\sigma^{(d)})^2 \mathbf{Exch}_T(\rho^{(d)})$, i.e., the true correlation structure is exchangeable with correlation $\rho^{(d)}$, then

$$b_{j,k}^{(d)} = \frac{\rho^{(d)}}{\left(1 + \rho^{(d)}\right)\left(1 + (j-1)\rho^{(d)}\right)}$$

for all $k = 1, \ldots, j-1$.

In the second stage of the SMART, data is generated conditionally on response status, so equation (4.7) becomes, for $j = T_1 + 1, \ldots, T$,

$$\boldsymbol{\Xi}^{(d)}(r)_{1:j-1,1:j-1} \boldsymbol{b}_j^{(d)} = \boldsymbol{\Xi}^{(d)}(r)_{1:j-1,j} \tag{4.8}$$

for all $d \in \mathcal{D}$ and $r \in \{0, 1\}$. As we discuss below, there is not a closed-form expression for $\boldsymbol{\Xi}(r)^{-1}$, so we cannot give an expression for $c_{j,k}^{(d)}$; however, equation (4.8) is easy to solve computationally.

We now describe how the generative model elicits and computes variances for second-stage potential outcomes. We begin by partitioning $\boldsymbol{\Xi}^{(d)}(r)$ as

$$\boldsymbol{\Xi}^{(d)}(r) = \left[\begin{array}{c|c} \boldsymbol{\Xi}_{11}^{(d)}(r) & \boldsymbol{\Xi}_{12}^{(d)}(r) \\ \hline \left(\boldsymbol{\Xi}_{12}^{(d)}(r)\right)^{\top} & \boldsymbol{\Xi}_{22}^{(d)}(r) \end{array}\right]_{T \times T}, \tag{4.9}$$

where $\boldsymbol{\Xi}_{11}^{(d)}(r) \in \mathbb{R}^{T_1 \times T_1}$ and $\boldsymbol{\Xi}_{22}^{(d)}(r) \in \mathbb{R}^{T_2 \times T_2}$. Note that $\boldsymbol{\Xi}_{11}^{(d)}(r) = \text{Var}(\boldsymbol{Y}_{1:j-1}^{(d)} \mid R^{(d)} = r)$ is determined by the choice of response status and is therefore fixed. In general, $\boldsymbol{\Xi}^{(d)}(r)$ does not respect the marginal covariance structure, and is typically unstructured.

Both $\boldsymbol{\Xi}_{12}^{(d)}(r)$ and $\boldsymbol{\Xi}_{22}^{(d)}(r)$ need to be specified for either $r = 1$ or $r = 0$. This involves identifying stage-2 and "cross-stage" covariances, both conditional on response. By design, the re-randomizations in SMARTs produce subsets of participants consistent with more than one DTR; recall that, for example, responders in design II are consistent with both embedded DTRs which recommend the same first-stage treatment. This means that we need only specify $\boldsymbol{\Xi}_{12}^{(d)}(r)$ and $\boldsymbol{\Xi}_{22}^{(d)}(r)$ for one DTR which recommends each first-stage treatment; all others are fixed by the law of total variance:

$$\begin{aligned} \boldsymbol{\Sigma}^{(d)} = P(R^{(a_1^{(d)})} = 1)\boldsymbol{\Xi}^{(d)}(1) + \left(1 - P(R^{(a_1^{(d)})} = 1)\right)\boldsymbol{\Xi}^{(d)}(0) \\ + P(R^{(a_1^{(d)})} = 1)\left(1 - P(R^{(a_1^{(d)})} = 1)\right)\left(\boldsymbol{\nu}^{(\boldsymbol{d})}(1) - \boldsymbol{\nu}^{(d)}(0)\right)^{\otimes 2}. \end{aligned} \tag{4.10}$$

In sum, the process of generating potential outcomes in a longitudinal SMART involves the following steps:

1. For each embedded DTR d, specify $\boldsymbol{\mu}^{(d)}$, the marginal mean outcome at all measurement occasions and $\boldsymbol{\Sigma}^{(d)}$, the marginal variance of the outcome.

2. Specify a response function $g_{a_1}(\boldsymbol{Y}_{1:T_1}^{(d)})$ which assigns (potential) response status based on

potential outcomes in the first stage of the SMART.

3. For one DTR that recommends each first-stage treatment and one response status r, specify $\boldsymbol{\Xi}_{12}^{(d)}(r)$ and $\boldsymbol{\Xi}_{22}^{(d)}(r)$.

4. For the jth measurement occasion in stage 1, $j = 1, \ldots, T_1$, find $\boldsymbol{c}_j^{(d)}$ by solving equation (4.7). Simulate mean-zero noise $\epsilon_j^{(d)}$ with variance (4.2) (using, e.g., a normal distribution), then generate potential outcomes $Y_{i,j}^{(d)}$ using equation (4.1).

5. For each first-stage treatment option, compute potential response status $R_i^{(a_1)^{(d)}} = g_{a_1}(\boldsymbol{Y}_{i,1:T_1}^{(d)})$ for each simulated individual.

6. For each treatment path, find $\boldsymbol{\nu}^{(d)}(r)$ and $\boldsymbol{\Xi}^{(d)}(r)$ using equations (4.6) and (4.10). For each $j = T_1 + 1, \ldots, T$, compute $\boldsymbol{b}_j^{(d)}$ by solving equation (4.8). Simulate mean-zero noise $\zeta_{i,j}^{(d)}(R_i^{(a_1^{(d)})})$ with variance as in equation (4.5) and generate second-stage potential outcomes using equation (4.4).

In table 4.1, we show that the target marginal variance structures are achieved using this data generative model for three measurement occasions with an exchangeable marginal correlation structure.

4.1.2 "Observing" Potential Outcomes

Once the potential outcomes data is generated following the procedure in section 4.1.1, we will "observe" a subset of those outcomes for each participant in the SMART according to simulated treatment assignment. Recall that the potential baseline measure, $Y_{i,1}^{(d)}$, is the same for all DTRs d,

Table 4.1: Target and estimated marginal variance matrices from the data generative model described in section 4.1.1. The "unstructured estimate" is produced by estimating the variance at each timepoint and for each DTR, and correlation for each DTR using the unstructured estimate in table 2.2, then averaging over DTRs. The "exchangeable estimate" is computed by assuming variance is constant over time and DTR, and using the exchangeable estimate of ρ from table 2.2, averaged over DTRs. The exchangeable estimate is used in simulations assuming working assumption A2.2 is satisifed.

Design	Target Structure	Unstructured Estimate	Exchangeable Estimate
I	$\begin{pmatrix} 64 & 19.2 & 19.2 \\ 19.2 & 64 & 19.2 \\ 19.2 & 19.2 & 64 \end{pmatrix}$	$\begin{pmatrix} 63.9 & 19.3 & 19.1 \\ 19.3 & 63.8 & 18.7 \\ 19.1 & 18.7 & 62.5 \end{pmatrix}$	$\begin{pmatrix} 63.4 & 18.9 & 18.9 \\ 18.9 & 63.4 & 18.9 \\ 18.9 & 18.9 & 63.4 \end{pmatrix}$
II	$\begin{pmatrix} 36 & 10.8 & 10.8 \\ 10.8 & 36 & 10.8 \\ 10.8 & 10.8 & 36 \end{pmatrix}$	$\begin{pmatrix} 35.9 & 10.9 & 11.0 \\ 10.9 & 35.9 & 11.1 \\ 11.0 & 11.1 & 35.8 \end{pmatrix}$	$\begin{pmatrix} 35.9 & 10.9 & 10.9 \\ 10.9 & 35.9 & 10.9 \\ 10.9 & 10.9 & 35.9 \end{pmatrix}$
III	$\begin{pmatrix} 64 & 19.2 & 19.2 \\ 19.2 & 64 & 19.2 \\ 19.2 & 19.2 & 64 \end{pmatrix}$	$\begin{pmatrix} 63.9 & 19.4 & 19.9 \\ 19.4 & 63.7 & 21.3 \\ 19.9 & 21.3 & 63.6 \end{pmatrix}$	$\begin{pmatrix} 63.8 & 20.0 & 20.0 \\ 20.0 & 63.8 & 20.0 \\ 20.0 & 20.0 & 63.8 \end{pmatrix}$

since it is pre-treatment. The observed data

$$
\begin{aligned}
Y_{i,1} &= Y_{i,1}^{(d)} \\
A_{i,1} \mid Y_{i,1} &\sim 2 * \text{Bernoulli}(\pi_1) - 1 \\
\boldsymbol{Y}_{i,2:T_1} \mid A_{i,1}, Y_{i,1} &= Z_1\left(A_{i,1}, \boldsymbol{Y}_i^{(d)}\right) \\
R_i \mid A_{i,1}, \boldsymbol{Y}_{i,1:T_1} &= \mathbb{1}_{\{A_{i,1}=1\}} R_i^{(1)} + \mathbb{1}_{\{A_{i,1}=-1\}} R_i^{(-1)} \\
A_{i,2} \mid R_i, A_{i,1}, \boldsymbol{Y}_{i,1:T_1} &\sim 2 * \text{Bernoulli}(\pi_2) - 1 \\
\boldsymbol{Y}_{i,T_1+1:T} \mid A_{i,2}, R_i, A_{i,1}, \boldsymbol{Y}_{i,1:T_1} &= Z_2(A_{i,1}, R_i, A_{i,2}, \boldsymbol{Y}_i^{(d)}),
\end{aligned}
$$

where Z_k is a function which maps observed treatment history and (if applicable) response status to the potential outcome $\boldsymbol{Y}_{\boldsymbol{t},i}^{(d)}$ consistent with that history. The form of Z_k is given in table A.1, and varies by SMART design.

4.1.3 Threshold-Based Response Status

A challenging aspect of generating data for longitudinal SMARTs is the fact that analyses are (typically) conducted marginally over response status. When generating data, however, it necessary to understand means and variances of stage-1 outcomes conditional on response status. This is non-trivial, as it is, in some sense, conditioning on "the future": response status is often a function of stage-1 outcomes. Our need for this is driven primarily by a need to ensure that the correct marginal variance structure is achieved, and because working assumption A3.1 depends on these quantities. Here, we discuss how these conditional means and variances are computed for a "threshold-based" response status, as defined in equation (4.3). Here, we assume errors $\epsilon_j^{(d)}$ are jointly normally distributed in the first stage.

At $t^* = t_{T_1}$, the measurement time to which the threshold is applied, the potential outcomes follow a truncated normal distribution conditional on response, with density

$$f_{Y_{T_1}^{(d)}|R^{(d)}}(y \mid r) = \frac{\phi\left(\frac{y-\mu_{T_1}^{(d)}}{\sigma_{T_1}^{(d)}}\right)}{1-\Phi\left(\frac{\kappa_{a_1}-\mu_{T_1}^{(d)}}{\sigma_{T_1}^{(d)}}\right)}, \tag{4.11}$$

where ϕ is the standard normal density and Φ is the standard normal cumulative distribution function.

For times $t_j < t^*$, we rely on joint normality of the errors, so that

$$\begin{bmatrix} Y_{i,j}^{(d)} \\ Y_{i,T_1}^{(d)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_j^{(d)} \\ \mu_{T_1}^{(d)} \end{bmatrix}, \begin{bmatrix} \left(\sigma_j^{(d)}\right)^2 & \rho_{j,T_1}^{(d)}\sigma_j^{(d)}\sigma_{T_1}^{(d)} \\ \rho_{j,T_1}^{(d)}\sigma_j^{(d)}\sigma_{T_1}^{(d)} & \left(\sigma_{T_1}^{(d)}\right)^2 \end{bmatrix}\right).$$

By properties of the bivariate normal distribution,

$$\mu_{T_1|j}^{(d)}(y_j) := \mathrm{E}\left[Y_{T_1}^{(d)} \mid Y_j^{(d)} = y_j\right] = \mu_{T_1}^{(d)} + \frac{\rho_{j,T_1}^{(d)}\left(\sigma_{T_1}^{(d)}\right)^2}{\left(\sigma_j^d\right)^2}\left(y_j - \mu_j^{(d)}\right)$$

and Bayes' theorem,

$$f_{Y_j^{(d)}|R^{(d)}}(y \mid r) = \frac{\phi\left(\frac{y-\mu_j^{(d)}}{\sigma_j^{(d)}}\right)\left(1-\Phi\left(\frac{\kappa_{a_1}-\mu_{T_1|j}^{(d)}(y)}{\left(1-\left(\rho_{T_1,j}^{(d)}\right)^2\right)\left(\sigma_{T_1}^{(d)}\right)^2}\right)\right)}{1-\Phi\left(\frac{k_{a_1}-\mu_{T_1}^{(d)}}{\sigma_{T_1}^{(d)}}\right)} \tag{4.12}$$

For $j = 1, \ldots, T_2$, we can find the mean and variance of the potential outcoems under each first-stage treatment conditional on response using equation (4.12).

Using similar ideas, we can find the conditional density of products of first-stage potential outcomes $Y_j^{(d)}Y_k^{(d)}$ given response status, which is necessary to compute conditional covariances. The details of this computation are not particularly illuminating; we refer the interested reader to the documentation for the companion R package longsmart. Similar derivations are required for other definitions of response to ensure that working assumption A3.1 is satsified.

4.2 The longsmart R Package

The longsmart package for R (available at https://github.com/nseewald1/longsmart) implements the generative model described in section 4.1 as well as the methods for sample size described in chapter 3. The package enables relatively easy, customizable simulation of data arising from longitudinal SMARTs and provides users with tools to design such trials by choosing sample size and the number and timing of measurement occasions, keeping trial budget in mind.

A key idea that runs throughout the package is the identification of a SMART design via randomization probabilities. We assert that a SMART design can be uniquely identified by a set of randomization probabilities π. In a two-stage SMART with at most two treatment options at each randomization, $\pi = \{\pi_1, \boldsymbol{\pi}_{2R}, \boldsymbol{\pi}_{2NR}\}$, where $\pi_1 = P(A_1 = 1)$,

$$\boldsymbol{\pi_{2R}} = \left(P(A_{2R} = 1 \mid A_1 = 1, R = 1), P(A_{2R} = 1 \mid A_1 = -1, R = 1)\right)^\top,$$

and

$$\boldsymbol{\pi_{2NR}} = \left(P(A_{2NR} = 1 \mid A_1 = 1, R = 0), P(A_{2NR} = 1 \mid A_1 = -1, R = 0)\right)^\top.$$

We will say that $P(A_{2NR} = 1 \mid A_1 = 1, R = 0) = 0$ if responders to first-stage treatment $A_1 = 1$ are not re-randomized; similarly with other elements of $\boldsymbol{\pi}_{2R}$ and $\boldsymbol{\pi}_{2NR}$. For example, we could identify a design II SMART with equal randomizations wiht $\pi = \left\{0.5, (0,0)^\top, (0.5, 0.5)^\top\right\}$.

We first discuss longsmart's implementation of the design methods from chapter 3, then how it can be used to simulate data.

4.2.1 Tools for Designing SMARTs

An important function available in longsmart is the `smart_size()` function, which implements the general sample size formula for longitudinal SMARTs in formula (3.6). The package implements a version of the within-person deflation factor ω which allows for measurement occasions which are not equally spaced, given as equation (B.26) in appendix B.3.

The `smart_size()` function takes a variety of inputs which describe the SMART for which sample size is desired. Users specify the design of the SMART using the `randomization` argument, which encodes π as discussed above. The function uses this to compute the appropriate design effect DE in formula (3.6). Other design-related inputs include `mTimes`, the vector of measurement times, and `tStar`, the time measurement time immediately after which participants are re-randomized. `smart_size()` also elicits information about the interventions under study through `rho`, the within-person correlation, and `pR`, a vector of probabilities of response.

Results from `smart_size()` are of class `power.htest`, so the output is presented to the user in a familiar way, similar to built-in power functions such as `power.t.test()`. An example is in program 4.1. As with other power functions in base R, the user can alternatively specify `power = NULL` or `sig.level = NULL` to have the function compute power or significance level given a sample size `n`. The default `randomization` argument yields a prototypical SMART with equal randomization throughout.

Program 4.1: Use of the `smart_size()` function to compute sample size for a longitudinal SMART. The target standardized effect size is $\delta = 0.3$, the outcome is measured at times 0-4, and re-randomization occurs after measurement time 2. We assume $\rho = 0.3$, and 40% response rates to both first-stage treatments. The minimum-required sample size to compare two embedded DTRs with different first-stage treatments is 427.

```
library(longsmart)
smart_size(n = NULL, delta = 0.3, mTimes = c(0, 1, 2, 3, 4),
           tStar = 2, power = 0.8, pR = c(0.4, 0.4), rho = .3)

#      Longitudinal SMART power calculation
#
#               n = 427
#           delta = 0.3
#       sig.level = 0.05
#           power = 0.8
#      alternative = two.sided
#       meas.times = 0, 1, 2, 3, 4
#           t.star = 2
#              rho = 0
#               pR = 0.4, 0.4
#
# NOTE: Power for a SMART in which the probability of
# randomization to first-stage treatment A1 = 1 is 0.5;
# responders are not re-randomized; non-responders are
# re-randomized to second-stage treatment A2 = 1 with
# probability 0.5.
```

Program 4.2: Use of the `optimize_cost()` function to find the number and allocation of measurement occasions which minimize per-participant trial costs. The investigator wishes to detect an effect size $\delta = 0.4$ and is planning a 16-week study in which each stage is 8 weeks long. They will consider at most 8 measurement occasions.

```
optimize_cost(delta = 0.4, tStar = 8, tMax = 16, numTimesMax = 8,
              rho = 0.36, pR = c(0.4, 0.5), cost_recruit = 300,
              cost_meas = 20)

#     Cost-optimal measurement allocation for longitudinal SMART
#
# Call:
# optimize_cost(delta = 0.4, tStar = 8, tMax = 16, numTimesMax = 8,
#     rho = 0.36, pR = c(0.4, 0.5), cost_recruit = 300,
#     cost_meas = 20)
#
# Optimal total number of measurements:  8
# Optimal number of measurements in stage 2:  5
# Sample size required:  160
# Total cost:  73,600
```

A second useful design tool in longsmart is the `optimize_cost()` function, which implements grid search to solve optimization problem (3.12). Given the same SMART design-related arguments as `sample_size()` (i.e., `delta`, `rho`, `pR`, and `randomization`), the maximum number of measurement occasions the user is willing to consider in in the SMART (`numTimesMax`), and costs of recruitment (`cost_recruit`) and measurement (`cost_meas`), `optimize_cost()` will identify the combination of T and T_2 which minimize total cost of the trial.

Consider an example in which an investigator will run a 16-week design II SMART with re-randomization after week 8. They are willing to consider at most 8 measurement occasions. The hypothesized response rate to first-stage treatments $A_1 = 1$ and $A_1 = -1$ are 0.4 and 0.5, respectively, and previous literature suggests an exchangeable within-person correlation of 0.36. The cost of recruiting one participant is expected to be \$300; the cost of each measurement is \$20 for both stages. The study should be sized to detect an effect size of 0.4 with at least 80% power using a two-sided level-0.05 test. The solution to optimization problem (3.12), shown in program 4.2, is to use $T^{\text{cost}} = 8$ total measurements, placing $T_2^{\text{cost}} = 5$ in the second stage. The investigator should measure the outcomes at $\boldsymbol{t} = \{0, 4, 8, 9.6, 11.2, 12.8, 14.4, 16\}$ weeks. The required sample size is

160 participants. This is the cheapest of all possible configurations of T^{cost} and T_2^{cost} that achieves 80% power; the total cost of recruitment and measurement is $73,600.

Note that if the investigator is interested solely in minimizing the sample size requirement, as in section 3.3.1, they can simply set `cost_meas` to 0 in `optimize_cost()`.

4.2.2 Tools for Simulating Longitudinal SMARTs

The longsmart package is designed with a particular eye towards simplifying the process of simulating data from longitudinal SMARTs. Working with the generative model in section 4.1 can be challenging; longsmart attempts to meet that challenge by creating user-friendly interfaces.

The primary function used for simulating data from a longitudinal SMART is `design_smart()`. This function creates an object of class `longsmartDesign`, and is the foundation of all simulation-related functions in longsmart. A `longsmartDesign` object is a list which completely describes the SMART the user wishes to simulate. These properties include the randomization set π, the times at which the outcome is measured, and marginal and conditional means and variances for all embedded DTRs and treatment paths at all measurement occasions. By default, `design_smart()` is to use the threshold-based response status described in section 4.1.3, but the user can specify any function with appropriate inputs and returned objects (see below).

Consider an investigator wishing to run a design II SMART in which the outcome is measured five times, two of which are in the second stage of the trial. They might begin the simulation process by using `mean_model_prototypical()` to identify the marginal means at each measurement time from regression parameters $\boldsymbol{\beta}$ as in model (3.1), which they have stored in R as a length-7 vector called `betas`.

```
means <- mean_model_prototypical(mTimes = 0:4, tStar = 2,
                                 marginalCoefs = betas)
```

This returns a list containing marginal means $\boldsymbol{\mu}^{(d)}$ for all embedded DTRs d as well as a data frame identifying those embedded DTRs by first- and second-stage treatment recommendations.

The use of `design_smart()` requires the user specify $\boldsymbol{\Xi}_{12}^{(a_1,0,1)}(1)$ and $\boldsymbol{\Xi}_{22}^{(a_1,0,1)}(1)$ for both

Program 4.3: Creation of a `longsmartDesign` object.

```
smart <- design_smart(
    mTimes = 0:4,
    tStar = 2,
    marginalMeans = means,
    marginalVariances = 36 *
        cormat(rho = 0.3, p = 5,
            corstr = "exchangeable"),
    responderMeans = list(c(33, 32), c(35, 34)),
    responderVariances = list(
        list(matrix(rep(10.8, 6), nrow = 3),
             36 * cormat(0.3, 2, "exch")),
        list(matrix(rep(10.8, 6), nrow = 3),
             36 * cormat(0.3, 2, "exch"))),
    threshold = c(32, 33)
)
```

first-stage treatments a_1. The `cormat()` function can be of use for specifying $\boldsymbol{\Xi}_{22}^{(a_1,0,1)}(1)$: given a correlation and dimension `p` $= T_2$, `cormat()` will return a correlation matrix with the given structure:

```
36 * cormat(rho = 0.3, p = 2, corstr = "exchangeable")

#      [,1] [,2]
# [1,] 36.0 10.8
# [2,] 10.8 36.0
```

Putting everything together, the user create a `longsmartDesign` object following the example in program 4.3. The `responderMeans` argument takes a list of vectors $\boldsymbol{\nu}_{T_1+1:T}^{(a_1,0,1)}(1)$, one element per first-stage treatment, and `responderVariances` is a list of lists, where the first list contains $\boldsymbol{\Xi}_{12}^{(1,0,1)}(1)$ and $\boldsymbol{\Xi}_{22}^{(1,0,1)}(1)$ for DTRs which recommend $a_1 = 1$, and the second is similar for $a_1 = -1$. By default, `design_smart()` uses threshold-based tailoring (see section 4.1.3); the thresholds for first-stage treatments $a_1 = 1$ and $a_1 = -1$ are specified in the `threshold` argument.

`design_smart()` is highly flexible with regard to specification of a tailoring variable. The optional `responseFun` argument allows the user to specify a function which describes response status, taking, at minimum, three arguments: `stage1Data`, a data frame containing the baseline outcome measurement as well as potential outcomes for both first-stage treatments at all measurement times prior to t^*; `meanModel`, an object describing measurement occasions and marginal means

Program 4.4: Simulation of data from a longitudinal SMART.

```
d <- generate_smart(n = 300, smart = smart)
head(d$obsData, 5)

#   id       Y0 A1       Y1       Y2 R A2       Y3       Y4
# 1  1 35.54661 -1 29.26066 24.89015 0 -1 37.34615 45.55055
# 2  2 26.11318  1 26.04476 36.76698 1  0 42.73083 33.07181
# 3  3 34.70027 -1 40.37163 29.46754 0 -1 28.96139 41.86827
# 4  4 38.06908 -1 31.93310 24.99966 0  1 43.43282 32.37043
# 5  5 25.60596 -1 36.09325 38.32931 1  0 23.78595 39.34885
#   weight  dtr1  dtr2  dtr3  dtr4
# 1      4 FALSE FALSE FALSE  TRUE
# 2      2  TRUE  TRUE FALSE FALSE
# 3      4 FALSE FALSE FALSE  TRUE
# 4      4 FALSE FALSE  TRUE FALSE
# 5      2 FALSE FALSE  TRUE  TRUE
```

for all embedded DTRs at all measurement occasions; and `marginalVariance`, a list of marginal variance matrices. The function then must return a data frame with potential response statuses for each observation in `stage1Data`, as well as probabilities of response to each first-stage treatment. When specifying a custom `responseFun`, the arguments `conditionalMeanFun` and `conditionalVarFun` must also be provided. These return conditional means and variances, respectively, for first-stage outcomes given response status. For threshold-based tailoring, these functions work by integrating over the densities given in section 4.1.3.

Once a `longsmartDesign` object is created using `design_smart()`, the user can generate data for `n` participants from the designed SMART using `generate_smart()`. An example is shown in program 4.4. The returned object is of class `longsmart`, and contains all potential outcomes, the observed data, and information about the SMART design from which the data were generated as a `longsmartDesign` object. The observed data is shown above in wide format and includes both weights and indicators for consistency with each embedded DTR (see table 2.1). `d$obsData` is ready to use with the analysis method of the user's choice.

CHAPTER 5
Conclusions and Future Work

In this dissertation, we have developed a variety of tools for designing and analyzing sequential, multiple-assignment randomized trials with continuous longitudinal outcomes. We pay particular attention to the case in which the trial is designed to compare two embedded dynamic treatment regimens which recommend different first-stage treatments. A key goal is to reduce barriers to implementing SMARTs among clinicians and applied statisticians. To that end, we have been keenly interested in ease-of-use.

Formulae (2.13) and (3.6) have been designed to require relatively few parameters. Those required inputs are, we believe, relatively easy to estimate from the literature or pilot studies. Indeed, our method requires only one additional input (ρ) relative to formulae for SMARTs in which the outcome is measured once at the end of the study. We described in section 2.5 that formula (2.13) is conservative when ρ is underrestimated; similarly for formula (3.6). In this way, the methods are able to accommodate uncertainty in the investigator's guess of the exchangeable within-person correlation by selecting the lowest of the possible values of ρ.

We acknowledge and accommodate the practical realities of clinical trial design by incorporating financial considerations into the methods developed in chapter 3. We describe a search-based approach to minimizing the total cost of recruiting participants and measuring the outcome. The approach is based on a simple cost function, but could easily be extended to accommodate more complex situations like incentives that change over time. The method is implemented in an easy-to-use function in the R package longsmart. A future goal is to build a web-based sample size tool as a companion to longsmart, which will allow us to better guide clinicians through the planning stages of a longitudinal SMART.

The main contribution of this dissertation is the development of sample size formulae for SMARTs in which the primary aim is an end-of-study comparison of two embedded DTRs which recommend different first-stage treatments (so-called "separate-path" DTRs; Kidwell and Wahed (2013)). It is possible, though, that some trialists may have interest in sizing a SMART for an end-of-study comparison of "shared-path" DTRs; that is, two DTRs which recommend the same first-stage treatment. We believe that, for the comparison of shared-path DTRs, investigators are better set to use a standard sample size calculation to compare the second-stage treatments (conditional on response) which differ between the DTRs, then upweighting the result by the proportion of participants expected to be in these groups.

It is important to note that while the methods described in this dissertation allow for a more varied conversation about the design of longitudinal SMARTs, the focus of this conversation must always be on the science. SMARTs, like other randomized trials, should be designed to address particular scientific questions: the trial's design should be chosen based on those questions, not the other way around. Sample size calculations should be similarly principled. Target effect sizes, for example, should be specified prior to choosing a sample size. This is reinforced in the design of `design_smart()` in longsmart: unlike base R functions like `power.t.test()`, we do not allow the user to find a detectable effect size given a sample size and power.

Statisticians should be careful to use the optimization approaches in section 3.3 in a way that serves the investigator's scientific interests, and not cherry-pick components of the sample size formulae which minimize cost and/or sample size. Effort should be made to pre-specify reasonable ranges of values for ρ and response rates based on either pre-existing evidence about the interventions under study in the SMART, or domain knowledge. Uncertainty in the choices of these values are of course to be expected, and power curves over a range of choices are a valuable tool; the ranges of these parameters, as well as the maximum number of measurement occasions under consideration, should be determined by scientific, ethical, or practical considerations.

There are a number of interesting ways to build on this dissertation in future methodological work. First, some scientists may be interested in a primary aim comparison that involves other

features of the marginal mean trajectory, such as the area under the curve (AUC). Future work could develop formulae for these other primary aim comparisons. An important challenge here is in whether and how to define the standardized effect size δ. We believe this would be best implemented through software tools with graphical user interfaces that allow investigators to interactively build and explore models for AUC. The extension of, say, formula (3.6) to other estimands is a matter of specifying a new contrast of regression parameters, meaning we can rely on most of the derivation in appendix B.

A second extension would be to build methods for SMARTs with intensive longitudinal outcomes. In contrast to more traditional repeated-measures data, intensive longitudinal data (ILD) is observed (potentially much) more frequently and can provide more detailed information on an individual's trajectory over time. This allows researchers to study the dynamics of behavioral or disease processes on a much finer scale compared to a more conventional setting in which relatively few observations are made (Hamaker and Wichers 2017).

Because of the SMART's usefulness in constructing DTRs, which are decision rules leading to a sequence of treatments tailored to an individual's changing needs over time, there is increasing interest in collecting intensive longitudinal data throughout a SMART. This would enable more detailed assessment of the impact of treatment over time, as well as any delayed effects of treatment that may arise as a consequence of the sequencing of interventions within a DTR. In some settings, these effects may be quite proximal; ILD can potentially capture brief changes. An example of the use of ILD in SMARTs is given in section 5.3 of Lu et al. (2016), in which the authors model the outcome using regression splines. This work, particularly the details of how to incorporate design features of a SMART into the model, is discussed only briefly, and could serve as a starting point for meaningful future projects.

This work has been largely guided by a focus on the investigator. Ultimately, the design of a trial is driven by the scientific questions it seeks to address. The methods in this dissertation aim to help reduce barriers for investigators seeking to design appropriate, efficient SMARTs to address pressing questions in their field.

APPENDIX A

Identifiability Assumptions

We make the following assumptions in order to show that equation (2.6) has mean zero.

I1 *Positivity.* The probabilities $P(A_1 = 1)$ and $P(A_2 = 1 \mid A_1, R)$ are non-zero.

I2 *Consistency* (Robins 1997). A participant's observed responder status is consistent with the participant's corresponding potential responder status under the assigned first-stage treatment; i.e.,

$$R_i = \mathbb{1}_{\{A_{1,i}=1\}} R^{(1)} + \mathbb{1}_{\{A_{1,i}=-1\}} R^{(-1)}.$$

And a participant's observed repeated measures outcomes are consistent with the participant's corresponding potential repeated measures outcomes under the assigned treatment sequence. For observations at measurement occasion j in stage k, we write $Y_{j,i} = Z_k(\bar{A}_k, R_i, \boldsymbol{Y}_i^{(d)})$ see table A.1. Here, "stage k" is defined such that measurement occasions $j = 1, \ldots, T_1$ are in stage $k = 1$; occasions $j = T_1 + 1, \ldots, T$ are in stage $k = 2$.

I3 *Sequential randomization.* At each stage in the SMART, observed treatments A_1 and A_2 are assigned independently of future potential outcomes, given the participant's history up to that point. That is,

$$\{\boldsymbol{Y}_{t \le t^*}^{(d)}, R(a_1)\} \perp\!\!\!\perp A_1$$
$$\{\boldsymbol{Y}_{t > t^*}^{(d)}\} \perp\!\!\!\perp A_2 \mid A_1, R$$

Identifiability assumptions I1 and I3 are satisfied by design in a SMART (see, e.g., Lavori

Table A.1: Design-specific consistency assumptions. $d \in \mathcal{D}$ indexes embedded DTRs (a_1, a_{2R}, a_{2NR}).

Design	Time t	$Z_k(\bar{A}_k, R_i, \boldsymbol{Y}_i^{(d)})$
I	t_0	$Y_{t,i}^{(d)}$
	$t_0 < t \leq t^*$	$\frac{1}{4} \sum_{d \in \mathcal{D}} \mathbb{1}_{\left\{A_{1,i}=a_1^{(d)}\right\}} Y_{t,i}^{(d)}$
	$t > t^*$	$\sum_{d \in \mathcal{D}} \frac{1}{2} \mathbb{1}_{\left\{A_{1,i}=a_1^{(d)}\right\}} \left(R_i \mathbb{1}_{\left\{A_{2,i}=a_{2R}^{(d)}\right\}} + (1-R_i) \mathbb{1}_{\left\{A_{2,i}=a_{2NR}^{(d)}\right\}} \right) Y_{t,i}^{(d)}$
II	t_0	$Y_{t,i}^{(d)}$
	$t_0 < t \leq t^*$	$\frac{1}{2} \sum_{d \in \mathcal{D}} \mathbb{1}_{\left\{A_{1,i}=a_1^{(d)}\right\}} Y_{t,i}^{(d)}$
	$t > t^*$	$\sum_{d \in \mathcal{D}} \mathbb{1}_{\left\{A_{1,i}=a_1^{(d)}\right\}} \left(\frac{1}{2} R_i + (1-R_i) \mathbb{1}_{\{A_{2,i}=a_2\}} \right) Y_{t,i}^{(d)}$
III	t_0	$Y_{t,i}^{(d)}$
	$t_0 < t \leq t^*$	$\sum_{d \in \mathcal{D}} \mathbb{1}_{\left\{A_{1,i}=a_1^{(d)}\right\}} \left(\frac{1}{2} \mathbb{1}_{\left\{a_1^{(d)}=1\right\}} \right) Y_{t,i}^{(d)}$
	$t > t^*$	$\sum_{d \in \mathcal{D}} \mathbb{1}_{\left\{A_{1,i}=a_1^{(d)}\right\}} \left(\mathbb{1}_{\left\{a_1^{(d)}=-1\right\}} + \mathbb{1}_{\left\{a_1^{(d)}=1\right\}} \left(\frac{1}{2} R_i + (1-R_i) \mathbb{1}_{\left\{A_{2,i}=a_2^{(d)}\right\}} \right) \right) Y_{t,i}^{(d)}$

The factor of $1/2$ applied to some (or all) participants when $t > t*$ accounts for the fact that these participants are consistent with two DTRs. In design I, all participants are consistent with two DTRs. In design II, only responders are consistent with two DTRs, so, if $R_i = 1$ for some i, $Y_{t>t^*,i}^{(a_1,0,1)} = Y_{t>t^*,i}^{(a_1,0,-1)} := Y_{t>t^*,i}^{(a_1,0,0)}$. Similarly for responders to $a_1 = 1$ in design III.

and Dawson (2014)); identifiability assumption I2 is connects the potential outcomes and observed data, and is typically accepted in the analysis of randomized trials.

APPENDIX B
Proofs and Derivations

B.1 Proofs of Propositions 2.1 and 2.2

We first prove proposition 2.1, that $\hat{\boldsymbol{\theta}}$, the solution to equation (2.6) over $\boldsymbol{\theta}$, is asymptotically consistent for $\boldsymbol{\theta}^*$, the true regression parameter.

Define $\hat{\boldsymbol{\theta}}_n$ to be the solution of the estimating equations

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n}\sum_{d\in\mathcal{D}}\left[W^{(d)}\left(A_{1,i}, R_i, A_{2,i}\right)\cdot \boldsymbol{D}^{(d)}(\boldsymbol{X}_i)^\top \boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})\right)\right] \tag{(2.6) revisited}$$

using data from n individuals. Let $\boldsymbol{Z}_i$ contain the ith individual's observed covariates (including outcome, treatment assignments, etc.). We can re-write equation (2.6) as

$$\mathbf{0} = \Psi_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\psi_{\boldsymbol{\theta}}(\boldsymbol{Z}_i), \tag{B.1}$$

where

$$\psi_{\boldsymbol{\theta}}(\boldsymbol{Z}_i) = \sum_{d\in\mathcal{D}} W^{(d)}\left(A_{1,i}, R_i, A_{2,i}\right)\cdot \boldsymbol{D}^{(d)}(\boldsymbol{X}_i)^\top \boldsymbol{V}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\tau})^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})\right). \tag{B.2}$$

Let $\hat{\boldsymbol{\theta}}_n$ be a solution to equation (B.2) for given n, and define $\boldsymbol{\theta}^*$ as the true parameter value, such that $\boldsymbol{\theta}^*$ is a zero of $\Psi(\boldsymbol{\theta}) = \mathrm{E}[\psi_{\boldsymbol{\theta}}(\boldsymbol{Z})]$.

Assuming the parameter space Θ is compact, $\sup_{\boldsymbol{\theta}\in\Theta}\|\Psi_n(\boldsymbol{\theta}) - \Psi(\boldsymbol{\theta})\| \xrightarrow{p} 0$ by the weak law of large numbers for random functions. If the model $\boldsymbol{\mu}^{(d)}(\boldsymbol{X}_i;\boldsymbol{\theta})$ is correctly specified and $\boldsymbol{\theta}^*$

is the unique solution of $\Psi(\boldsymbol{\theta}) = \mathrm{E}[\psi_{\boldsymbol{\theta}}(\boldsymbol{Z})]$, then consistency follows from standard results for M-estimation of a location parameter (see, e.g., Keener (2010) Theorems 9.2, 9.4, and 9.33).

To prove proposition 2.2, consider a first-order Taylor expansion of the estimating equations (2.6) about $\boldsymbol{\theta}^*$, assuming continuous differentiability of $\psi_{\boldsymbol{\theta}}$:

$$\mathbf{0} = \Psi_n(\hat{\boldsymbol{\theta}}_n) = \Psi_n\left(\boldsymbol{\theta}^*\right) + \Psi_n'\left(\tilde{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right), \tag{B.3}$$

where $\tilde{\boldsymbol{\theta}}$ is some intermediate value between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$. Note that $\Psi_n'\left(\tilde{\boldsymbol{\theta}}\right)$ is a $p \times p$ matrix, where p is the dimension of $\boldsymbol{\theta}$. If $\Psi_n'\left(\tilde{\boldsymbol{\theta}}\right)$ is non-singular, equation (B.3) can be re-written as

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right) = -\sqrt{n}\Psi_n'\left(\tilde{\boldsymbol{\theta}}\right)^{-1}\Psi_n\left(\boldsymbol{\theta}^*\right). \tag{B.4}$$

By the central limit theorem, $\sqrt{n}\Psi_n\left(\boldsymbol{\theta}^*\right) \Rightarrow \mathcal{N}\left(\mathbf{0}, \mathrm{Var}(\psi_{\boldsymbol{\theta}}(\boldsymbol{Z}))\right)$.

Under sufficient regularity conditions (see, e.g., van der Vaart (1998) theorem 5.41), and because $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$, we have $-\Psi_n'(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} -\mathrm{E}\left[\psi_{\boldsymbol{\theta}}'(\boldsymbol{Z})\right]$.

Define $\boldsymbol{B} = \mathrm{E}\left[\psi_{\boldsymbol{\theta}}'(\boldsymbol{Z})\right]$ and $\boldsymbol{M} = \mathrm{Var}(\psi_{\boldsymbol{\theta}}(\boldsymbol{Z})) = \mathrm{E}\left[\psi_{\boldsymbol{\theta}}(\boldsymbol{Z})^{\otimes 2}\right]$. By Slutsky's theorem and the delta method, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right) \Rightarrow \mathcal{N}\left(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\right). \tag{B.5}$$

This completes the proof.

B.2 Derivation of Sample Size Formulae for Three Measurements

We derive the sample size formulae for comparing two DTRs which recommend different first-stage treatments that are embedded in a SMART in which a continuous repeated-measures outcome is collected throughout the study. These formulae are based on the regression analyses described in section 2.2 and a Wald test.

We consider a SMART in which the outcome is collected three timepoints: at baseline ($t = 0$), immediately before assessing response/non-response ($t = 1$), and at the end of the study ($t = 2$). We ignore the presence of baseline covariates $\boldsymbol{X}$ and assume $\boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})$ is piecewise-linear in $\boldsymbol{\theta}$ (see, for example, model (2.1)).

Recall from section 2.3 that we wish to the null hypothesis $H_0 : \boldsymbol{c}^\top \boldsymbol{\theta} = 0$. In particular, we are interested in contrasts $\boldsymbol{c}$ which yield an end-of-study comparison between two embedded DTRs which recommend different first-stage treatments. Since a comparison of two embedded DTRs will yield a 1-degree of freedom Wald test, we use a Z statistic:

$$Z = \frac{\sqrt{n}\boldsymbol{c}^\top \hat{\boldsymbol{\theta}}}{\sigma_c},$$

where $\sigma_c = \sqrt{\boldsymbol{c}^\top \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}}$. Under H_0, by asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, the test statistic follows an asymptotic standard normal distribution. Suppose we wish to size the SMART to detect the alternative hypothesis $\boldsymbol{c}^\top \boldsymbol{\theta} = \Delta$. By the definition of type-II error, we have

$$\begin{aligned}
\beta &= P\left(\left|\frac{\sqrt{n}\boldsymbol{c}^\top \hat{\boldsymbol{\theta}}}{\sigma_c}\right| \leq z_{1-\alpha/2} \mid \boldsymbol{c}^\top \boldsymbol{\theta} = \Delta\right) \\
&= P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}}{\sigma_c}\boldsymbol{c}^\top \hat{\boldsymbol{\theta}} \leq z_{1-\alpha/2} \mid \boldsymbol{c}^\top \boldsymbol{\theta} = \Delta\right) \\
&= P\left(-z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\boldsymbol{c}^\top \boldsymbol{\theta} \leq \frac{\sqrt{n}}{\sigma_c}\boldsymbol{c}^\top \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \leq z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\boldsymbol{c}^\top \boldsymbol{\theta} \mid \boldsymbol{c}^\top \boldsymbol{\theta} = \Delta\right) \\
&= P\left(-z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\Delta \leq \frac{\sqrt{n}}{\sigma_c}\boldsymbol{c}^\top \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \leq z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\Delta\right) \\
&= \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\Delta\right) - \Phi\left(-z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\Delta\right) \\
&\leq \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{n}}{\sigma_c}\Delta\right),
\end{aligned}$$

we arrive at the following form for the minimum-required sample size:

$$n \geq \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \frac{\sigma_c^2}{\Delta^2}, \tag{B.6}$$

where z_p is the pth quantile of the standard normal distribution. Formula (B.6) is a fairly standard result in the clinical trials literature (Lachin 1981; Friedman, Furberg, and DeMets 2010); however, because of the dependence on σ_c, the formula is not useful as written. The goal of this appendix is to derive a closed-form upper bound on σ_c so as to obtain a sample size formula in terms of marginal quantities which can be more easily elicited from clinicians, or estimated from the literature.

Recall the definitions of $\boldsymbol{B}$ and $\boldsymbol{M}$ in equations (2.7) and (2.8), respectively. These quantities depend on $\boldsymbol{D}^{(d)}$, the partial derivative matrix of $\boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})$ and $\boldsymbol{V}^{(d)}(\boldsymbol{\tau})$, the working covariance matrix for $\boldsymbol{Y}$. By assumed linearity of $\boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})$, $\boldsymbol{D}^{(d)}$ is a fixed, constant matrix for all d. Furthermore, we assume that the working covariance matrix $\boldsymbol{V}^{(d)}(\boldsymbol{\tau})$ is correctly specified and satisfies working assumption A2.2 so that $\boldsymbol{V}^{(d)}(\boldsymbol{\tau}) = \boldsymbol{\Sigma}$ for all $d \in \mathcal{D}$. Note that $\boldsymbol{\Sigma}$ is non-random.

The estimand in equation (2.11) is a function of potential outcomes; as written in equations (2.7) and (2.8), $\boldsymbol{B}$ and $\boldsymbol{M}$ are functions of observed data. We begin by expressing $\boldsymbol{B}$ in terms of potential outcomes. Under the positivity, consistency, and sequential ignorability conditions (identifiability assumptions I1 to I3) and assuming that $\boldsymbol{V}^{(d)}(\boldsymbol{\tau})$ is correctly specified and equal to $\boldsymbol{\Sigma}$, we can apply lemma 4.1 of Murphy et al. (2001) so that

$$\begin{aligned}
\boldsymbol{B} &= \sum_{d \in \mathcal{D}} \mathop{\mathrm{E}}_{A_1,R,A_2}\left[W^{(d)}(A_1, R, A_2)\boldsymbol{D}^{(d)}\left(\boldsymbol{V}^{(d)}(\boldsymbol{\tau})\right)^{-1}\left(\boldsymbol{D}^{(d)}\right)^{\top}\right] \\
&= \sum_{d \in \mathcal{D}} \boldsymbol{D}^{(d)}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{D}^{(d)}\right)^{\top},
\end{aligned} \tag{B.7}$$

since $\boldsymbol{D}^{(d)}$ and $\boldsymbol{\Sigma}$ are non-random and $\mathrm{E}[W^{(d)}(A_1, R, A_2)] = 1$.

We now turn our attention to $\boldsymbol{M}$. Expanding the outer product inside the expectation, we have

$$\begin{aligned}
\boldsymbol{M} = \mathop{\mathrm{E}}_{A_1,R,A_2,\boldsymbol{Y}} & \left[\left(\sum_{d\in\mathcal{D}} W^{(d)}(A_1,R,A_2)\boldsymbol{D}^{(d)} \left(\boldsymbol{V}^{(d)}(\boldsymbol{\tau})\right)^{-1} \left(\boldsymbol{Y} - \boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})\right) \right)^{\otimes 2} \right] \\
&= \sum_{d\in\mathcal{D}} \mathop{\mathrm{E}}_{A_1,R,A_2,\boldsymbol{Y}} \left[\left(W^{(d)}(A_1,R,A_2)\right)^2 \left(\boldsymbol{D}^{(d)}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})\right)\right)^{\otimes 2} \right] \\
&\quad + \sum_{d\neq d'} \mathop{\mathrm{E}}_{A_1,R,A_2,\boldsymbol{Y}} \Big[W^{(d)}(A_1,R,A_2)W^{(d')}(A_1,R,A_2)\boldsymbol{D}^{(d)}\boldsymbol{\Sigma}^{-1} \\
&\qquad\qquad \left(\boldsymbol{Y} - \boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})\right)\left(\boldsymbol{Y} - \boldsymbol{\mu}^{(d')}(\boldsymbol{\theta})\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{D}^{(d')}\right)^\top \Big].
\end{aligned} \tag{B.8}$$

Consider a single summand of the first term in equation (B.8). We can write this as

$$\boldsymbol{D}^{(d)}\boldsymbol{\Sigma}^{-1} \mathop{\mathrm{E}}_{A_1,R,A_2,\boldsymbol{Y}} \left[W^{(d)}(A_1,R,A_2)^2 \left(\boldsymbol{Y} - \boldsymbol{\mu}^{(d)}(\boldsymbol{\theta})\right)^{\otimes 2} \right] \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{D}^{(d)}\right)^\top$$

The inner expectation is a $T \times T$ matrix, the (i, j)th element of which is

$$\mathop{\mathrm{E}}_{A_1,R,A_2,\boldsymbol{Y}} \left[W^{(d)}(A_1,R,A_2)^2 \left(Y_{t_i} - \mu^{(d)}_{t_i}(\boldsymbol{\theta})\right)\left(Y_{t_j} - \mu^{(d)}_{t_j}(\boldsymbol{\theta})\right) \right]. \tag{B.9}$$

Notice that the work above is design-independent: $\boldsymbol{B}$ and $\boldsymbol{M}$ have the same form as equations (B.7) and (B.8), respectively, for all designs. Below, we proceed only for design II, but derivations for designs I and III are analogous, substituting appropriate definitions of $W^{(d)}(A_1, R, A_2)$. Recall that, for design II, when all randomization probabilities are 0.5, $W^{(d)}(A_1, R, A_2) = 2\mathbb{1}_{\left\{A_1 = a_1^{(d)}\right\}}(R + 2(1 - R)\mathbb{1}_{\left\{A_2 = a_2^{(d)}\right\}})$. Further, we restrict our focus to three timepoints, denoted t_0 (baseline), $t_1 = t^*$, and $t_2 > t^*$.

Consider, for example, $t = t_1$. By repeated use of iterated expectation and application of

identifiability assumptions I2 and I3, equation (B.9) becomes

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{Y_{t_0},A_1,Y_{t_1},R,A_2,Y_{t_2}}\left[W^{(d)}(A_1,R,A_2)^2\left(Y_{t_1}-\mu_{t_1}^{(d)}(\boldsymbol{\theta})\right)^2\right]\\
&=\mathop{\mathrm{E}}_{Y_{t_0},A_1,Y_{t_1},R,A_2}\left[4\mathbb{1}_{\left\{A_1=a_1^{(d)}\right\}}\left(R+4(1-R)\mathbb{1}_{\left\{A_2=a_2^{(d)}\right\}}\right)\left(Y_{t_1}-\mu_{t_1}^{(d)}(\boldsymbol{\theta})\right)^2\right]\\
&=\mathop{\mathrm{E}}_{Y_{t_0}^{(d)},A_1,Y_{t_1},R^{(a_1)},A_2^{(d)}}\left[4\mathbb{1}_{\left\{A_1=a_1^{(d)}\right\}}\left(R^{(a_1)}+4(1-R^{(a_1)})\mathbb{1}_{\left\{A_2=a_2^{(d)}\right\}}\right)\left(Y_{t_1}^{(d)}-\mu_{t_1}^{(d)}(\boldsymbol{\theta})\right)^2\right]\\
&=\mathop{\mathrm{E}}_{S_2(\bar{A}_1)}\left[4\mathbb{1}_{\left\{A_1=a_1^{(d)}\right\}}\left(R^{(a_1)}+4(1-R^{(a_1)})\mathop{\mathrm{E}}_{A_2|S_2(\bar{A}_1)}\left[\mathbb{1}_{\left\{A_2=a_2^{(d)}\right\}}\right]\right)\left(Y_{t_1}^{(d)}-\mu_{t_1}^{(d)}(\boldsymbol{\theta})\right)^2\right]\\
&=\mathop{\mathrm{E}}_{Y_{t_0}^{(d)},A_1,Y_{t_1}^{(d)},R^{(a_1)}}\left[4\mathbb{1}_{\left\{A_1=a_1^{(d)}\right\}}\left(2-R^{(a_1)}\right)\left(Y_{t_1}^{(d)}-\mu_{t_1}^{(d)}(\boldsymbol{\theta})\right)^2\right]. \qquad \text{(B.10)}\\
&=4\mathop{\mathrm{E}}_{Y_1^{(d)}}\left[\left(Y_1-\mu_1^{(d)}\right)^2\right]-2\mathop{\mathrm{E}}_{Y_1^{(d)},R^{(a_1)}}\left[\left(Y_1-\mu_1^{(d)}\right)^2R^{(a_1)}\right] \qquad \text{(B.11)}\\
&=4\sigma^2-2\operatorname{Cov}\left(\left(Y_1-\mu_1^{(d)}\right)^2,R^{(a_1)}\right)-2\operatorname{E}\left[R^{(a_1)}\right]\operatorname{E}\left[\left(Y_1-\mu_1^{(d)}\right)^2\right] \qquad \text{(B.12)}\\
&=2(2-r_{a_1})\sigma^2. \qquad \text{(B.13)}
\end{aligned}
$$

Equation (B.11) follows from equation (B.10) by identifiability assumption I3 and smoothing over $Y_{t_0}^{(d)}$, equation (B.12) arises from the definition of covariance, and equation (B.13) is a consequence of working assumption A2.1(b).

Similar derivations and applications of the remaining working assumptions allow us to bound $\boldsymbol{c}^\top\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}$ above by

$$
\begin{aligned}
\boldsymbol{c}^\top\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c} &\leq 2\cdot\frac{1}{2}\left((2-r_1)+(2-r_{-1})\right)\boldsymbol{c}^\top\boldsymbol{B}^{-1}\left(\sum_{d\in\mathcal{D}}\boldsymbol{D}^{(d)}\boldsymbol{\Sigma}^{-1}\right)^{\otimes 2}\boldsymbol{B}^{-1}\boldsymbol{c}\\
&=\frac{4\sigma^2(1-\rho)\left(\rho^2+4\rho-\frac{1}{2}(r_1+r_{-1})(2\rho+1)+2\right)}{1+\rho}. \qquad \text{(B.14)}
\end{aligned}
$$

Plugging equation (B.14) into formula (B.6) leads to the aforementioned "sharp" sample size

formula for design II. Some algebra shows that

$$\sigma_c^2 \leq 4\sigma^2 \cdot \left(1-\rho^2\right) \cdot \frac{1}{2}\left((2-r_1)+(2-r_{-1})\right), \tag{B.15}$$

which allows for an easy-to-understand sample size formula. Plugging this result into formula (B.6), we arrive at formula (2.13).

B.3 Derivation of Sample Size Formulae for Arbitrary Measurements

We first establish two definitions and a lemma which will be useful for constructing an upper bound on the variance of estimand of interest for sample size calculations.

Definition B.1 (Positive semi-definite)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ be a $p \times p$ symmetric real-valued matrix. We say $\boldsymbol{A}$ is positive semi-definite if for any vector $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$.*

Definition B.2 (Loewner partial order)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two symmetric matrices. We say that $\boldsymbol{A} \overset{L}{\geq} \boldsymbol{B}$ if $\boldsymbol{A}-\boldsymbol{B}$ is positive semi-definite.*

Lemma B.1. *Let $\boldsymbol{A}$, $\boldsymbol{B}$ be symmetric matrices. If $\boldsymbol{B} \overset{L}{\geq} \boldsymbol{A}$, then $\boldsymbol{C}^\top \boldsymbol{B}\boldsymbol{C} \overset{L}{\geq} \boldsymbol{C}^\top \boldsymbol{A}\boldsymbol{C}$ for any matrix $\boldsymbol{C}$ of suitable dimension.*

Proof. By definition B.2, $\boldsymbol{B}-\boldsymbol{A}$ is positive semi-definite; i.e., for any vector $\boldsymbol{x}$ of appropriate length, $\boldsymbol{x}^\top(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{x} \geq 0$. Let $\boldsymbol{z} = \boldsymbol{C}\boldsymbol{x}$ for a suitably-sized $\boldsymbol{C}$. Then

$$\boldsymbol{z}^\top(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{z} \geq 0$$
$$\boldsymbol{x}^\top\boldsymbol{C}^\top(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{C}\boldsymbol{x} \geq 0.$$

Since x is arbitrary, $\boldsymbol{C}^\top(\boldsymbol{B}-\boldsymbol{A})\boldsymbol{C}$ is positive semi-definite by definition, completing the proof. □

Consider a SMART in which the outcome is measured at T occasions, labeled $\{t_1, \ldots, t_{T-T_2}, \ldots, t_T\}$, where $T_2 \in \{1, \ldots, T-2\}$ is the number of measurements in the second stage. As above, we can express the estimand of interest as a contrast of regression parameters in the marginal mean models described in Section 3.1, omitting baseline covariates $\boldsymbol{X}_i$. As above, we write the estimand as a contrast of regression parameters.

Consider model (2.1), a simplified version of which is reproduced below:

$$\begin{aligned}\mu_j^{(a_1,a_{2R},a_{2NR})}(\boldsymbol{\beta}) = \beta_0 &+ \mathbb{1}_{\{t_j \le t^*\}}\left(\beta_1 t_j + \beta_2 a_1 t_j\right) \\ &+ \mathbb{1}_{\{t_j > t^*\}}\left(t^*\beta_1 + t^*\beta_2 a_1 + \beta_3(t_j - t^*) + \beta_4(t_j - t^*)a_1\right. \\ &\left. + \beta_5(t_j - t^*)a_{2NR} + \beta_6(t_j - t^*)a_1 a_{2NR}\right).\end{aligned} \tag{2.1 revisited}$$

We can write the end-of-study estimand as

$$\mathrm{E}\left[Y_{i,T}^{(1,a_{2\mathrm{R}},a_{2\mathrm{NR}})} - Y_{i,T}^{(-1,a'_{2\mathrm{R}},a'_{2\mathrm{NR}})}\right] = \boldsymbol{c}^\top \boldsymbol{\beta},$$

where

$$\boldsymbol{c}^\top = \left(0, 0, 2t^*, 0, 2(t_T - t^*), (t_T - t^*)\left(a_{2\mathrm{NR}} - a'_{2\mathrm{NR}}\right), (t_T - t^*)\left(a_{2\mathrm{NR}} + a'_{2\mathrm{NR}}\right)\right).$$

As before, we wish to size the study for the hypothesis test

$$\mathrm{H}_0 : \boldsymbol{c}^\top \boldsymbol{\beta} = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \boldsymbol{c}^\top \boldsymbol{\beta} = \Delta,$$

where Δ is a fixed alternative value, with power $1 - \gamma$. The test statistic is

$$Z = \frac{\sqrt{n}\boldsymbol{c}^\top \hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}^\top \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}}},$$

which follows an asymptotic standard normal distribution under H_0 by the results in Appendix B.1.

As in appendix B.2, developing a useful sample size formula depends on obtaining a tractable

expression for $\boldsymbol{c}^\top \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}$. With a general number of measurement occasions, the expression for $\boldsymbol{B}$ given in equation (B.7) holds. We now consider a more general expansion of $\boldsymbol{M}$ which will accommodate more than three measurement occasions. For a given $d \in \mathcal{D}$, define $C(d) = \left\{d' \in \mathcal{D} : d' \neq d, a_1^{(d')} = a_1^{(d)}\right\}$ to be the set of DTRs which "share a path" with d. We can write $\boldsymbol{M}$ as the sum of DTR-specific components $\boldsymbol{M}_d$ and "cross-DTR" products $\boldsymbol{M}_{d,d'}$ for $d' \in C(d)$:

$$\begin{aligned}
\boldsymbol{M} :&= \mathrm{E}\left[\left(\sum_{d\in\mathcal{D}} W_i^{(d)} \left(\boldsymbol{D}^{(d)}\right)^\top \left(\boldsymbol{V}^{(d)}\right)^{-1} \left(\boldsymbol{Y}_i - \mu^{(d)}\right)\right)^{\otimes 2}\right] \\
&= \sum_{d\in\mathcal{D}} \mathrm{E}\left[\left(W_i^{(d)}\right)^2 \left(\boldsymbol{D}^{(d)}\right)^\top \left(V^{(d)}\right)^{-1} \left(\boldsymbol{Y}_i - \mu^{(d)}\right)^{\otimes 2} \left(V^{(d)}\right)^{-1} \boldsymbol{D}^{(d)}\right] \\
&\quad + \sum_{d\in\mathcal{D}} \sum_{d'\in C(d)} \mathrm{E}\left[W_i^{(d)} W_i^{(d')} \left(\boldsymbol{D}^{(d)}\right)^\top \left(\boldsymbol{V}^{(d)}\right)^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}\right)\right. \\
&\qquad \left.\left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d')}\right)^\top \left(\boldsymbol{V}^{-1}\right)^{(d')} \boldsymbol{D}^{(d')}\right] \\
&= \sum_{d\in\mathcal{D}} \boldsymbol{M}_d + \sum_{d\in\mathcal{D}} \sum_{d'\in C(d)} \boldsymbol{M}_{d,d'}
\end{aligned}$$

Individuals in a SMART cannot experience treatments consistent with DTRs which do not share a path; we will see that the definition of the weights in equation (2.5) and the identifiability indicators in table A.1 imply that we do not need to consider cross-DTR products $\boldsymbol{M}_{d,d'}$ for $d' \notin C(d)$ as in such situations the products $W_i^{(d)} W_i^{(d')} = 0$.

We now consider Design II SMARTs and examine $\boldsymbol{M}_d$ for any $d \in \mathcal{D}$:

$$\begin{aligned}
\boldsymbol{M}_d = \mathrm{E}&\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}\right\}} \left(R_i^{(a_1)} + 4\left(1 - R_i^{(a_1)}\right)\mathbb{1}_{\left\{A_{2i}=a_{2\mathrm{NR}}^{(d)}\right\}}\right) \left(\boldsymbol{D}^{(d)}\right)^\top \left(\boldsymbol{V}^{(d)}\right)^{-1}\right. \\
&\left.\left(\boldsymbol{Y}_i - \mu^{(d)}\right)^{\otimes 2} \left(\boldsymbol{V}^{(d)}\right)^{-1} \boldsymbol{D}^{(d)}\right]
\end{aligned} \tag{B.16}$$

$$\begin{aligned}
= \left(\boldsymbol{D}^{(d)}\right)^\top &\left(\boldsymbol{V}^{(d)}\right)^{-1} \mathrm{E}\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}\right\}} \left(R_i^{(a_1)} + 4\left(1 - R_i^{(a_1)}\right)\mathbb{1}_{\left\{A_{2i}=a_2^{(d)}\right\}}\right)\right. \\
&\left.\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] \left(\boldsymbol{V}^{(d)}\right)^{-1} \boldsymbol{D}^{(d)},
\end{aligned} \tag{B.17}$$

where equation (B.17) follows from equation (B.16) by identifiability assumption I2 and the fact that $\boldsymbol{D}^{(d)}$ is fixed when there are no baseline covariates in the model.

We now consider just the inner expectation. Recall from section 1.1 that we use $S_j(a_{j-1})$ to denote information collected in the period after providing treatment a_{j-1} until immediately prior to providing subsequent treatment a_j, and $\bar{S}_j(\bar{a}_{j-1}) = \{S_1, S_2(a_1), \ldots, S_{j-1}(\bar{a}_{j-2}), S_j(\bar{a}_j - 1)\}$ represents the "history" of observed data until the time at which a_j is recommended. Under identifiability assumption I3, the inner expectation in equation (B.17) becomes

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{\bar{S}_3(\bar{A}_2)}\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}\right\}}\left(R_i^{(a_1)}+4\left(1-R^{(a_1^{(d)})}\right)\mathbb{1}_{\left\{A_{2i}=a_2^{(d)}\right\}}\right)\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] \\
&=\mathop{\mathrm{E}}_{\bar{S}_2(\bar{A}_1)}\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}\right\}}\mathop{\mathrm{E}}_{S_3(\bar{A}_2)}\left[\left(R^{(a_1^{(d)})}+4\left(1-R^{(a_1^{(d)})}\right)\mathbb{1}_{\left\{A_{2i}=a_2^{(d)}\right\}}\right)\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\mid \bar{S}_2(\bar{A}_1)\right]\right] \\
&=\mathop{\mathrm{E}}_{\bar{S}_2(\bar{A}_1),S_3(a_2^{(d)})}\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}\right\}}\left(R^{(a_1^{(d)})}+2\left(1-R^{(a_1^{(d)})}\right)\right)\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] \\
&=\mathop{\mathrm{E}}_{\bar{S}_3(\bar{a}_2^{(d)})}\left[2\left(2-R^{(a_1^{(d)})}\right)\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] \\
&=4\mathop{\mathrm{E}}_{\bar{S}_3(\bar{a}_2^{(d)})}\left[\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right]-2\mathop{\mathrm{E}}_{\bar{S}_3(\bar{a}_2^{(d)})}\left[R^{(a_1^{(d)})}\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\right] \\
&=4\boldsymbol{\Sigma}^{(d)}-2P\left(R_i^{(a_1)}=1\right)\mathop{\mathrm{E}}_{\bar{S}_3(\bar{a}_2^{(d)})}\left[\left(\boldsymbol{Y}^{(d)}-\boldsymbol{\mu}^{(d)}\right)^{\otimes 2}\mid R_i^{(a_1)}=1\right] as
\end{aligned}
\tag{B.18}
$$

We would like to construct a simple upper bound on $\boldsymbol{c}^\top\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$. Under working assumption A3.1, equation (B.18) is bounded above (in the Loewner sense; see definition B.2) by $2\left(2-P(R^{(a_1^{(d)})}=1)\right)\boldsymbol{\Sigma}^d$. By constructing an upper bound on $\boldsymbol{M}$, we will arrive at an upper bound on $\boldsymbol{c}^\top\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c}$ by lemma B.1.

Now, assuming that $\boldsymbol{V}^{(d)}$ is correctly specified (i.e., $\boldsymbol{V}^{(d)}=\boldsymbol{\Sigma}^{(d)}=\boldsymbol{\Sigma}$ under working assumption A2.2), we have

$$
\boldsymbol{M}_d \overset{L}{\leq} 2\left(2-P(R^{(a_1^{(d)})}=1)\right)\left(\boldsymbol{D}^{(d)}\right)^\top\left(\boldsymbol{\Sigma}^{(d)}\right)^{-1}\boldsymbol{D}^{(d)} \tag{B.19}
$$

We now construct an upper bound in the Loewner sense on $\boldsymbol{M}_{d,d'}$ for $d' \in C(d)$. For a design II SMART, therefore,

$$\begin{aligned} W_i^{(d)} W_i^{(d')} &= 2\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}\right\}} \left(R_i + 2\left(1 - R_i\right) \mathbb{1}_{\left\{A_{2i}=a_2^{(d)}\right\}} \right) \\ &\quad \times 2\mathbb{1}_{\left\{A_{1i}=a_1^{(d')}\right\}} \left(R_i + 2\left(1 - R_i\right) \mathbb{1}_{\left\{A_{2i}=a_2^{(d')}\right\}} \right) \\ &= 4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}=a_1^{(d')}\right\}} R_i^{(a_1)}. \end{aligned}$$

Thus, we have

$$\begin{aligned} \boldsymbol{M}_{d,d'} &= \mathrm{E}\left[W_i^{(d)} W_i^{(d')} \left(\boldsymbol{D}^{(d)}\right)^\top \left(\boldsymbol{V}^{(d)}\right)^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}\right) \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d')}\right)^\top \left(\boldsymbol{V}^{-1}\right)^{(d')} \boldsymbol{D}^{(d')} \right] \\ &= \left(\boldsymbol{D}^{(d)}\right)^\top \left(\boldsymbol{V}^{(d)}\right)^{-1} \mathrm{E}\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}=a_1^{(d')}\right\}} R_i \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}\right) \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d')}\right)^\top \right] \\ &\quad \left(\boldsymbol{V}^{-1}\right)^{(d')} \boldsymbol{D}^{(d')}. \end{aligned} \tag{B.20}$$

We focus, as above, on the inner expectation in equation (B.20). Following the above applica-

tion of identifiability assumptions I2 and I3, we have

$$
\begin{aligned}
&\mathrm{E}\left[4\mathbb{1}_{\left\{A_{1i}=a_1^{(d)}=a_1^{(d')}\right\}} R_i \left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{Y}_i - \boldsymbol{\mu}^{(d')}\right)^\top\right] \\
&= 2\mathbb{1}_{\left\{a_1^{(d)}=a_1^{(d')}\right\}} \mathrm{E}\left[R_i^{(a_1)}\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{Y}_i^{(d')} - \boldsymbol{\mu}^{(d')}\right)^\top\right] \\
&= 2\mathbb{1}_{\left\{a_1^{(d)}=a_1^{(d')}\right\}} P\left(R_i^{(a_1)}=1\right) \mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{Y}_i^{(d')} - \boldsymbol{\mu}^{(d)} + \boldsymbol{\mu}^{(d)} - \boldsymbol{\mu}^{(d')}\right)^\top \mid R_i^{(a_1)}=1\right] \\
&= 2\mathbb{1}_{\left\{a_1^{(d)}=a_1^{(d')}\right\}} P\left(R_i^{(a_1)}=1\right)\Bigg(\mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{Y}_i^{(d')} - \boldsymbol{\mu}^{(d)}\right)^\top \mid R_i^{(a_1)}=1\right] \\
&\qquad + \mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{\mu}^{(d)} - \boldsymbol{\mu}^{(d')}\right)^\top \mid R_i^{(a_1)}=1\right]\Bigg)
\end{aligned}
\tag{B.21}
$$

$$
\begin{aligned}
&= 2P\left(R_i^{(a_1)}=1\right)\Bigg(\mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)^{\otimes 2} \mid R_i^{(a_1)}=1\right] \\
&\qquad + \mathrm{E}\left[\left(\boldsymbol{Y}_i^{(d)} - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{\mu}^{(d)} - \boldsymbol{\mu}^{(d')}\right)^\top \mid R_i^{(a_1)}=1\right]\Bigg),
\end{aligned}
\tag{B.22}
$$

where equation (B.22) follows from equation (B.21) by recognizing the fact that, for responders, DTRs d and $d' \in C(d)$ make identical treatment recommendations a_1 and $a_2 R$. Therefore, responders experience these DTRs in the same way, so the potential outcomes under both should be identical. We also drop the indicator since it evaluates to one for d and d' by definition of d'.

Again, assuming $\boldsymbol{V}^{(d)} = \boldsymbol{\Sigma}^{(d)} = \boldsymbol{\Sigma}$, we have, by working assumption A3.1,

$$
\begin{aligned}
\boldsymbol{M}_{d,d'} &\overset{L}{\leq} 2\left(\boldsymbol{D}^{(d)}\right)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{D}^{(d')} + 2\left(\boldsymbol{D}^{(d)}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\nu}^{(d)} - \boldsymbol{\mu}^{(d)}\right)\left(\boldsymbol{\mu}^{(d)} - \boldsymbol{\mu}^{(d')}\right)^\top \boldsymbol{D}^{(d')} \\
&\approx 2\left(\boldsymbol{D}^{(d)}\right)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{D}^{(d')}.
\end{aligned}
\tag{B.23}
$$

Combining equations (B.23) and (B.19), we have

$$
\boldsymbol{c}^\top \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{c} \leq 2\boldsymbol{c}^\top \boldsymbol{B}^{-1}\left(\sum_{d\in\mathcal{D}} \boldsymbol{\Sigma}^{-1}(2 - P\left(R^{(a_1^{(d)})}=1\right))\boldsymbol{D}^{(d)} + \boldsymbol{D}^{(d')}\right)\boldsymbol{B}^{-1}\boldsymbol{c}. \tag{B.24}
$$

Recall the "clock-time" parametrization of measurement times described in section 3.1, so

that we may refer to measurement times $\boldsymbol{u}$. After tedious algebra, we can write the right-hand side of equation (B.24) as

$$\begin{aligned}
&\sigma^2(1-\rho)(1+(T-1)\rho)\left(\frac{4(2-\bar{r})u_{2T}^2\left(g_1(\rho)g_2(\rho)-s_2^2h_1^2(\rho)\right)}{g_2(\rho)\left(g_1(\rho)g_2(\rho)-s_2^2h_1^2(\rho)\right)}\right.\\
&\qquad\left.+\frac{(6-r_1-r_{-1})\left(u_{1T}g_2(\rho)-u_{2T}s_2h_1(\rho)\right)^2}{g_2(\rho)\left(g_1(\rho)g_2(\rho)-s_2^2h_1^2(\rho)\right)}\right)\\
&\leq 4(2-\bar{r})\sigma^2\omega\left(\rho,\boldsymbol{u},T_2\right), \qquad \text{(B.25)}
\end{aligned}$$

where $r_{a_1}=P(R^{(a_1)}=1)$, $\bar{r}=(r_1+r_{-1})/2$, $s_k=\sum_{j=1}^{T}u_{kj}$, $h_k(\rho)=(1+(T-1)\rho)u_{kT}-\rho\sum_{j=1}^{T}u_{kT}$,

$$g_k(\rho)=\left(1+(T-1)\rho\right)\sum_{j=1}^{T}u_{kj}^2-\rho\left(\sum_{j=1}^{T}u_{kj}\right)^2,$$

and

$$\omega\left(\rho,\boldsymbol{u},T_2\right)=(1-\rho)(1+(T-1)\rho)\cdot\frac{u_{2T}^2g_1(\rho)+u_{1T}^2g_2(\rho)-2u_{1T}u_{2T}s_2h_1(\rho)}{g_1(\rho)g_2(\rho)-s_2^2h_1^2(\rho)}. \qquad \text{(B.26)}$$

Equation (B.26) simplifies to equation (3.7) under working assumption A3.3. We arrive at formula (3.6) by plugging equation (B.25) into formula (B.6).

APPENDIX C

Further Exploration of the Within-Person Deflation Factor

Recall the within-person deflation factor

$$\omega\left(\rho, \boldsymbol{u}, T_2\right) = (1-\rho)(1+(T-1)\rho) \times \frac{u_{2T}^2 g_1(\rho) + u_{1T}^2 g_2(\rho) - 2u_{1T}u_{2T}s_2 h_1(\rho)}{g_1(\rho)g_2(\rho) - s_2^2 h_1^2(\rho)}. \quad \text{((B.26) revisited)}$$

where $r_{a_1} = P(R^{(a_1)} = 1)$, $\bar{r} = (r_1 + r_{-1})/2$, $s_k = \sum_{j=1}^{T} u_{kj}$, $h_k(\rho) = (1+(T-1)\rho)u_{kT} - \rho \sum_{j=1}^{T} u_{kT}$,

$$g_k(\rho) = \left(1 + (T-1)\rho\right) \sum_{j=1}^{T} u_{kj}^2 - \rho \left(\sum_{j=1}^{T} u_{kj} \right)^2. \quad \text{(B.3 revisited)}$$

We show that $\omega\left(0, \boldsymbol{u}, 1\right) = 1$ for any measurement times $\boldsymbol{t}$. Here, note that $h_1(0) = u_{1T}$, $s_2 = \sum_{j=1}^{T} u_{2j} = u_{2T}$, $g_1(0) = \sum_{j=1}^{T} u_{1j}^2$, and $g_2(0) = \sum_{j=1}^{T} u_{2j}^2 = u_{2T}^2$. Plugging in to equation (B.26), we have

$$\omega\left(0, \boldsymbol{u}, 1\right) = \frac{u_{2T}^2 \sum_{j=1}^{T} u_{1j}^2 + u_{1T}^2 u_{2T}^2 - 2u_{1T}^2 u_{2T}^2}{u_{2T}^2 \sum_{j=1}^{T} u_{1j}^2 - u_{2T}^2 u_{1T}^2} = 1.$$

We next explore the behavior of $\omega\left(\rho, \boldsymbol{u}, T_2\right)$ when we do not make working assumption A3.3, i.e., when measurement times are not equally spaced in each stage. To do this, we recreate figure 3.2, this time adding noise to the measurement times $\boldsymbol{t}$. We keep t_1, $t_{T_1} = t^*$, and t_T fixed. For other measurements, we add uniformly-distributed noise to the equally-spaced times so that the t_j take values in non-overlapping windows. The resulting values of $\omega\left(\rho, \boldsymbol{u}, T_2\right)$ are depicted in figure C.1.

We see that the deflation factor is noticeably less "well-behaved" when measurement times

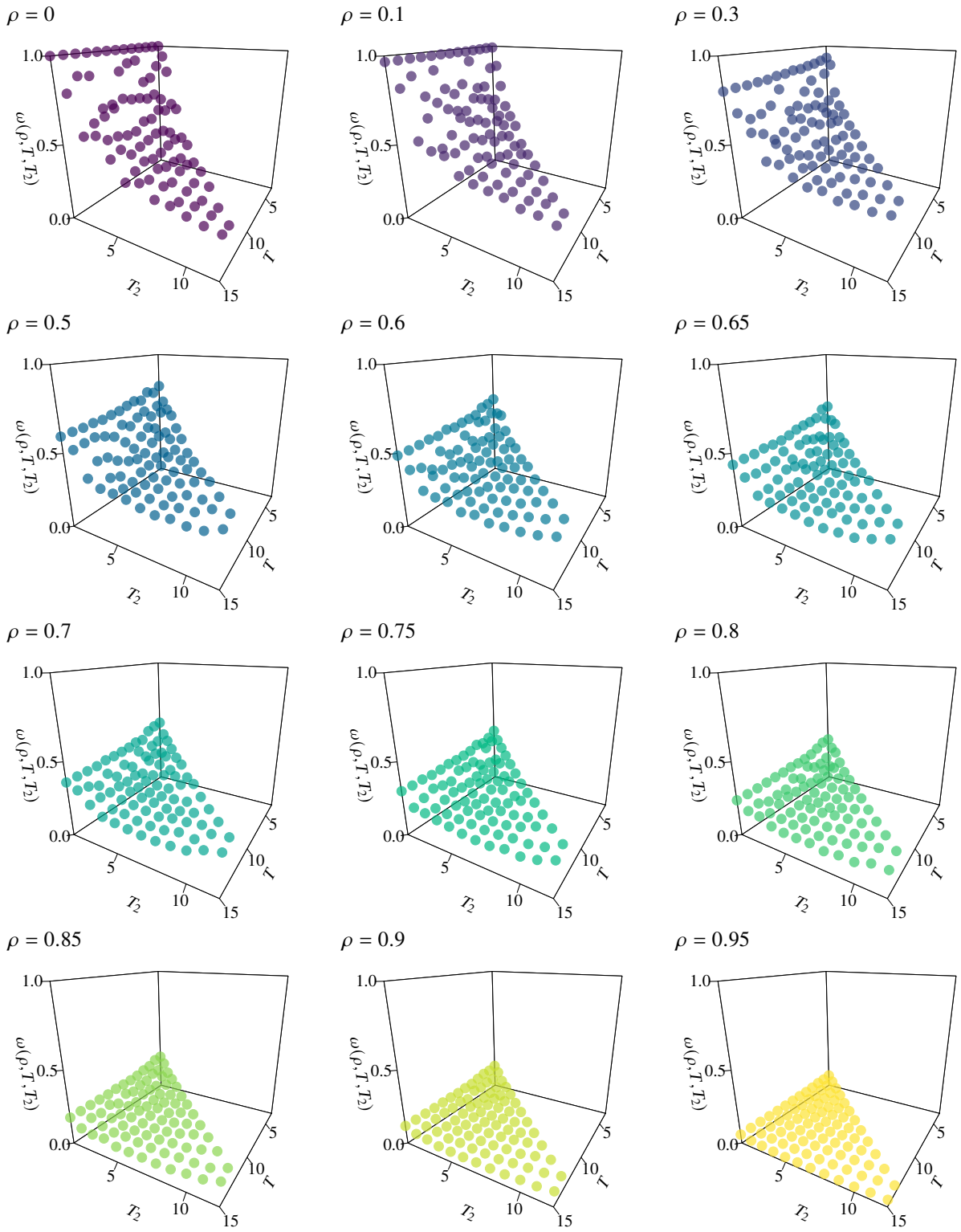


Figure C.1: Within-person deflation factor $\omega\left(\rho, \boldsymbol{u}, T_2\right)$ when working assumption A3.3 is violated. The function is bounded above by 1 for all ρ, T, and T_2, demonstrating that it is in fact a deflation factor. The function tends to decrease with T, but is quite jagged when T is large relative to T_2.

are not equally-spaced. Generally, we still see the trends discussed in section 3.2; namely, as ρ, T and T_2 increase, the deflation factor tends to decrease, with ω still obtaining a minimum on the interior of the domain of T_2 for large values of T. We conjecture that some or all of the jaggedness in figure C.1 arises from the fact that we are not "adding" measurement occasions to the SMART: each point in each plot is for a different set of measurement times $\boldsymbol{t}$.

BIBLIOGRAPHY

Almirall, D., C. DiStefano, Y.-C. Chang, S. Shire, A. Kaiser, X. Lu, I. Nahum-Shani, R. Landa, P. Mathy, and C. Kasari. 2016. "Longitudinal Effects of Adaptive Interventions With a Speech-Generating Device in Minimally Verbal Children With ASD." *Journal of Clinical Child & Adolescent Psychology* 45, no. 4 (2016): 442–456. ISSN: 1537-4416, 1537-4424. https://doi.org/10.1080/15374416.2016.1138407.

Almirall, D., I. Nahum-Shani, N. E. Sherwood, and S. A. Murphy. 2014. "Introduction to SMART Designs for the Development of Adaptive Interventions: With Application to Weight Loss Research." *Translational Behavioral Medicine* 4 (3): 260–274. ISSN: 1869-6716; EN :1613-9860. https://doi.org/10.1007/s13142-014-0265-0.

Almirall, D., I. Nahum-Shani, L. Wang, and C. Kasari. 2018. "Experimental Designs for Research on Adaptive Interventions: Singly and Sequentially Randomized Trials." In *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics,* edited by L. M. Collins and K. C. Kugler, 89–120. Statistics for Social and Behavioral Sciences. Cham: Springer International Publishing. ISBN: 978-3-319-91776-4. https://doi.org/10.1007/978-3-319-91776-4_4.

Arndt, S., R. Jorge, C. Turvey, and R. G. Robinson. 2000. "Adding Subjects or Adding Measurements: Which Increases the Precision of Longitudinal Research?" *Journal of Psychiatric Research* 34, no. 6 (2000): 449–455. ISSN: 0022-3956. https://doi.org/10/bjgwtn.

Auyeung, S. F., Q. Long, E. B. Royster, S. Murthy, M. D. McNutt, D. Lawson, A. Miller, A. Manatunga, and D. L. Musselman. 2009. "Sequential Multiple-Assignment Randomized Trial Design of Neurobehavioral Treatment for Patients with Metastatic Malignant Melanoma Undergoing High-Dose Interferon-Alpha Therapy." *Clinical Trials: Journal of the Society for Clinical Trials* 6 (5): 480–490. ISSN: 1740-7745, 1740-7753. https://doi.org/10.1177/1740774509344633.

Bloch, D. A. 1986. "Sample Size Requirements and the Cost of a Randomized Clinical Trial with Repeated Measurements." *Statistics in Medicine* 5 (6): 663–667. ISSN: 1097-0258. https://doi.org/10/bqvr7z.

Boen, J. R., and D. J. Zahn. 1982. *The Human Side of Statistical Consulting*. Belmont, California: Lifetime Learning Publications. ISBN: 0-534-97949-1. http://hdl.handle.net/2027/mdp.39015016210638.

Budney, A. J. 2014. "Behavioral Treatment of Adolescent Substance Use." Accessed December 8, 2018. https://clinicaltrials.gov/ct2/show/NCT02063984.

Chakraborty, B., and E. E. M. Moodie. 2013. *Statistical Methods for Dynamic Treatment Regimes.* Statistics for Biology and Health. New York, NY: Springer New York. ISBN: 978-1-4614-7427-2. https://doi.org/10.1007/978-1-4614-7428-9.

Cheung, Y. K., B. Chakraborty, and K. W. Davidson. 2015. "Sequential Multiple Assignment Randomized Trial (SMART) with Adaptive Randomization for Quality Improvement in Depression Treatment Program: SMART with Adaptive Randomization." *Biometrics* 71 (2): 450–459. ISSN: 0006341X. https://doi.org/10.1111/biom.12258.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, N.J: L. Erlbaum Associates. ISBN: 978-0-8058-0283-2.

Cole, S. R., and M. A. Hernán. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168 (6): 656–664. ISSN: 00029262. https://doi.org/10.1093/aje/kwn164. pmid: 18682488.

Collins, L. M., I. Nahum-Shani, and D. Almirall. 2014. "Optimization of Behavioral Dynamic Treatment Regimens Based on the Sequential, Multiple Assignment, Randomized Trial (SMART)." *Clinical Trials* 11 (4): 426–434. ISSN: 1740-7745, 1740-7753. https://doi.org/10/f6cjxm.

Cook, N. R., and J. H. Ware. 1983. "Design and Analysis Methods for Longitudinal Research." *Annual Review of Public Health* 4 (1): 1–23. ISSN: 0163-7525, 1545-2093. https://doi.org/10/br5tbh.

Diegidio, P., S. Hermiz, J. Hibbard, M. Kosorok, and C. S. Hultman. 2017. "Hypertrophic Burn Scar Research: From Quantitative Assessment to Designing Clinical Sequential Multiple Assignment Randomized Trials." *Clinics in Plastic Surgery* 44 (4): 917–924. ISSN: 00941298. https://doi.org/10.1016/j.cps.2017.05.024.

Dragalin, V. 2006. "Adaptive Designs: Terminology and Classification." *Drug Information J* 40 (4): 425–435. ISSN: 0092-8615, 2164-9200. https://doi.org/10/ghpbrt.

Eckshtain, D. 2013. "Using SMART Experimental Design to Personalize Treatment for Child Depression." Accessed December 8, 2018. https://clinicaltrials.gov/ct2/show/NCT01880814.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware. 2011. *Applied Longitudinal Analysis.* 2nd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley. ISBN: 978-0-470-38027-7.

Fitzsimons, H., M. Tuten, K. O'Grady, M. S. Chisolm, and H. E. Jones. 2015. "A Smart Design: Response to Reinforcement-Based Treatment Intensity among Pregnant, Drug-Dependent Women." *Drug and Alcohol Dependence* 156 (2015): e69. ISSN: 0376-8716. https://doi.org/10/gfn9pt.

Friedman, L. M., C. Furberg, and D. L. DeMets. 2010. *Fundamentals of Clinical Trials.* 4th ed. New York: Springer. ISBN: 978-1-4419-1585-6.

Fu, S. S., A. J. Rothman, D. M. Vock, B. Lindgren, D. Almirall, A. Begnaud, A. Melzer, et al. 2017. "Program for Lung Cancer Screening and Tobacco Cessation: Study Protocol of a Sequential, Multiple Assignment, Randomized Trial." *Contemporary Clinical Trials* 60:86–95. ISSN: 15517144. https://doi.org/10.1016/j.cct.2017.07.002.

Gail, M., and R. Simon. 1985. "Testing for Qualitative Interactions between Treatment Effects and Patient Subsets." *Biometrics* 41 (2): 361–372. ISSN: 0006-341X (Print) 0006-341X (Linking). https://doi.org/10.2307/2530862.

Hamaker, E. L., and M. Wichers. 2017. "No Time Like the Present: Discovering the Hidden Dynamics in Intensive Longitudinal Data." *Current Directions in Psychological Science* 26 (1): 10–15. ISSN: 0963-7214, 1467-8721. https://doi.org/10/f9r4kc.

Hedeker, D., R. D. Gibbons, and C. Waternaux. 1999. "Sample Size Estimation for Longitudinal Designs with Attrition: Comparing Time-Related Contrasts Between Two Groups." *J. Educ. Behav. Stat.* 24 (1): 70–93.

Hibbard, J. C., J. S. Friedstat, S. M. Thomas, R. E. Edkins, C. S. Hultman, and M. R. Kosorok. 2018. "LIBERTI: A SMART Study in Plastic Surgery." *Clinical Trials* 15 (3): 286–293. ISSN: 1740-7745, 1740-7753. https://doi.org/10.1177/1740774518762435.

Kahan, B. C., V. Jairath, C. J. Doré, and T. P. Morris. 2014. "The Risks and Rewards of Covariate Adjustment in Randomized Trials: An Assessment of 12 Outcomes from 8 Studies." *Trials* 15 (1): 139. ISSN: 1745-6215. https://doi.org/10.1186/1745-6215-15-139.

Kasari, C., A. Kaiser, K. Goods, J. Nietfeld, P. Mathy, R. Landa, S. A. Murphy, and D. Almirall. 2014. "Communication Interventions for Minimally Verbal Children with Autism: A Sequential Multiple Assignment Randomized Trial." *Journal of the American Academy of Child and Adolescent Psychiatry* 53 (6): 635–646. ISSN: 15275418. https://doi.org/10.1016/j.jaac.2014.01.019. pmid: 24839882.

Keener, R. W. 2010. *Theoretical Statistics: Topics for a Core Course.* Springer Texts in Statistics. New York: Springer. ISBN: 978-0-387-93838-7.

Kidwell, K. M. 2014. "SMART Designs in Cancer Research: Past, Present, and Future." *Clinical Trials: Journal of the Society for Clinical Trials* 11 (4): 445–456. ISSN: 1740-7745, 1740-7753. https://doi.org/10.1177/1740774514525691.

Kidwell, K. M., N. J. Seewald, Q. Tran, C. Kasari, and D. Almirall. 2018. "Design and Analysis Considerations for Comparing Dynamic Treatment Regimens with Binary Outcomes from Sequential Multiple Assignment Randomized Trials." *Journal of Applied Statistics* 45, no. 9 (2018): 1628–1651. ISSN: 0266-4763, 1360-0532. https://doi.org/10.1080/02664763.2017.1386773.

Kidwell, K. M., and A. S. Wahed. 2013. "Weighted Log-Rank Statistic to Compare Shared-Path Adaptive Treatment Strategies." *Biostatistics* 14, no. 2 (2013): 299–312. ISSN: 1465-4644. https://doi.org/10/gfppsw.

Kilbourne, A. M., K. M. Abraham, D. E. Goodrich, N. W. Bowersox, D. Almirall, Z. Lai, and K. M. Nord. 2013. "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." *Implementation Science* 8 (1). ISSN: 1748-5908. https://doi.org/10.1186/1748-5908-8-136.

Kilbourne, A. M., S. N. Smith, S. Y. Choi, E. Koschmann, C. Liebrecht, A. Rusch, J. L. Abelson, et al. 2018. "Adaptive School-Based Implementation of CBT (ASIC): Clustered-SMART for Building an Optimized Adaptive Implementation Intervention to Improve Uptake of Mental Health Interventions in Schools." *Implementation Sci* 13 (1): 119. ISSN: 1748-5908. https://doi.org/10/gd7jt2.

Kosorok, M. R., and E. E. M. Moodie, eds. 2015. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2015. ISBN: 978-1-61197-417-1. https://doi.org/10.1137/1.9781611974188.

Lachin, J. M. 1981. "Introduction to Sample Size Determination and Power Analysis for Clinical Trials." *Controlled Clinical Trials* 2 (2): 93–113. ISSN: 01972456. https://doi.org/10.1016/0197-2456(81)90001-5.

Lavori, P. W., and R. Dawson. 2004. "Dynamic Treatment Regimes: Practical Design Considerations." *Clin Trials* 1, no. 1 (2004): 9–20. ISSN: 17407745, 17407753. https://doi.org/10/cqtvnn. pmid: 16281458.

———. 2000. "A Design for Testing Clinical Strategies: Biased Adaptive within-Subject Randomization." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 163 (1): 29–38. https://doi.org/10.1111/1467-985X.00154. JSTOR: 2680506.

———. 2014. "Introduction to Dynamic Treatment Strategies and Sequential Multiple Assignment Randomization." *Clinical Trials: Journal of the Society for Clinical Trials* 11 (4): 393–399. ISSN: 1740-7745, 1740-7753. https://doi.org/10.1177/1740774514527651.

Lenth, R. V. 2001. "Some Practical Guidelines for Effective Sample Size Determination." *The American Statistician* 55 (3): 187–193. ISSN: 0003-1305, 1537-2731. https://doi.org/10/b4523s.

Li, Z. 2017. "Comparison of Adaptive Treatment Strategies Based on Longitudinal Outcomes in Sequential Multiple Assignment Randomized Trials." *Statistics in Medicine* 36, no. 3 (2017): 403–415. ISSN: 02776715. https://doi.org/10.1002/sim.7136.

Li, Z., and S. A. Murphy. 2011. "Sample Size Formulae for Two-Stage Randomized Trials with Survival Outcomes." *Biometrika* 98 (3): 503–518. ISSN: 00063444. https://doi.org/10.1093/biomet/asr019. pmid: 22363091.

Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22. https://doi.org/10.1093/BIOMET/73.1.13.

Lipsitz, S., G. Fitzmaurice, D. Sinha, N. Hevelone, J. Hu, and L. L. Nguyen. 2017. "One-Step Generalized Estimating Equations With Large Cluster Sizes." *Journal of Computational and Graphical Statistics* 26, no. 3 (2017): 734–737. ISSN: 1061-8600, 1537-2715. https://doi.org/10/gfn5fd.

Liu, J., and G. A. Colditz. 2017. "Optimal Design of Longitudinal Data Analysis Using Generalized Estimating Equation Models." *Biometrical Journal* 59 (2): 315–330. ISSN: 1521-4036. https://doi.org/10/f3s8zk.

Longford, N. T. 1999. "Selection Bias and Treatment Heterogeneity in Clinical Trials." *Statistics in Medicine* 18, no. 12 (1999): 1467–1474. ISSN: 0277-6715, 1097-0258. https://doi.org/10.1002/(SICI)1097-0258(19990630)18:12<1467::AID-SIM149>3.0.CO;2-H.

Lu, X., I. Nahum-Shani, C. Kasari, K. G. Lynch, D. W. Oslin, W. E. Pelham, G. Fabiano, and D. Almirall. 2016. "Comparing Dynamic Treatment Regimes Using Repeated-Measures Outcomes: Modeling Considerations in SMART Studies." *Statistics in Medicine* 35 (10): 1595–1615. ISSN: 1097-0258. https://doi.org/10/gg2gxc.

Markowitz, J. C., and B. L. Milrod. 2015. "What to Do When a Psychotherapy Fails." *The Lancet Psychiatry* 2 (2): 186–190. ISSN: 22150366. https://doi.org/10/gfj39f.

Martin, L., M. Hutchens, C. Hawkins, and A. Radnov. 2017. "How Much Do Clinical Trials Cost?" *Nat Rev Drug Discov* 16 (6): 381–382. ISSN: 1474-1776, 1474-1784. https://doi.org/10/gjhm5x.

Maxwell, S. E. 1998. "Longitudinal Designs in Randomized Group Comparisons: When Will Intermediate Observations Increase Statistical Power?" *Psychological Methods* 3 (3): 275–290. ISSN: 1082-989X. https://doi.org/10/dw3ktb.

McKay, J. R., M. L. Drapkin, D. H. A. Van Horn, K. G. Lynch, D. W. Oslin, D. DePhilippis, M. Ivey, and J. S. Cacciola. 2015. "Effect of Patient Choice in an Adaptive Sequential Randomization Trial of Treatment for Alcohol and Cocaine Dependence." *Journal of Consulting and Clinical Psychology* 83 (6): 1021–1032. ISSN: 1939-2117, 0022-006X. https://doi.org/10.1037/a0039534.

Meurer, W. J., R. J. Lewis, and D. A. Berry. 2012. "Adaptive Clinical Trials: A Partial Remedy for the Therapeutic Misconception?" *JAMA* 307, no. 22 (2012): 2377–2378. ISSN: 0098-7484. https://doi.org/10/gf3pmm.

Murphy, S. A., M. J. van der Laan, J. M. Robins, Conduct Problems Prevention Research Group, and Group. 2001. "Marginal Mean Models for Dynamic Regimes." *Journal of the American Statistical Association* 96 (456): 1410–1423. ISSN: 0162-1459. https://doi.org/10.1198/016214501753382327. pmid: 20019887.

Murphy, S. A. 2005. "An Experimental Design for the Development of Adaptive Treatment Strategies." *Statistics in Medicine* 24 (10): 1455–1481. ISSN: 0277-6715. https://doi.org/10.1002/sim.2022.

Murphy, S. A., and D. Almirall. 2009. "Dynamic Treatment Regimens." In *Encyclopedia of Medical Decision Making,* 1:419–422. Thousand Oaks, CA: SAGE Publications. ISBN: 978-1-4129-5372-6.

Murphy, S. A., K. G. Lynch, D. Oslin, J. R. McKay, and T. TenHave. 2007. "Developing Adaptive Treatment Strategies in Substance Abuse Research." *Drug and Alcohol Dependence* 88:S24–S30. ISSN: 03768716. https://doi.org/10.1016/j.drugalcdep.2006.09.008.

Myers, B. A., Y. Pillay, W. Guyton Hornsby, J. Shubrook, C. Saha, K. J. Mather, K. Fitzpatrick, and M. de Groot. 2019. "Recruitment Effort and Costs from a Multi-Center Randomized Controlled Trial for Treating Depression in Type 2 Diabetes." *Trials* 20, no. 1 (2019): 621. ISSN: 1745-6215. https://doi.org/10/gg5d49.

Naar-King, S., D. A. Ellis, A. Idalski Carcone, T. Templin, A. J. Jacques-Tiura, K. Brogan Hartlieb, P. Cunningham, and K.-L. C. Jen. 2016. "Sequential Multiple Assignment Randomized Trial (SMART) to Construct Weight Loss Interventions for African American Adolescents." *Journal of Clinical Child & Adolescent Psychology* 45, no. 4 (2016): 428–441. ISSN: 1537-4416, 1537-4424. https://doi.org/10/gf4ks4.

Nahum-Shani, I., D. Almirall, J. R. T. Yap, J. R. McKay, K. G. Lynch, E. A. Freiheit, and J. J. Dziak. 2020. "SMART Longitudinal Analysis: A Tutorial for Using Repeated Outcome Measures from SMART Studies to Compare Adaptive Interventions." *Psychological Methods* 25 (1): 1–29. ISSN: 1082-989X. https://doi.org/10/ggttht.

Nahum-Shani, I., M. Qian, D. Almirall, W. E. Pelham, B. Gnagy, G. A. Fabiano, J. G. Waxmonsky, J. Yu, and S. A. Murphy. 2012a. "Experimental Design and Primary Data Analysis Methods for Comparing Adaptive Interventions." *Psychological Methods* 17 (4): 457–477. ISSN: 1939-1463, 1082-989X. https://doi.org/10.1037/a0029372.

———. 2012b. "Q-Learning: A Data Analysis Method for Constructing Adaptive Interventions." *Psychological Methods* 17 (4): 478–494. ISSN: 1082-989X. https://doi.org/10.1037/a0029373. pmid: 23025434.

NeCamp, T., A. Kilbourne, and D. Almirall. 2017. "Comparing Cluster-Level Dynamic Treatment Regimens Using Sequential, Multiple Assignment, Randomized Trials: Regression Estimation and Sample Size Considerations." *Statistical Methods in Medical Research* 26 (4): 1572–1589. ISSN: 0962-2802, 1477-0334. https://doi.org/10.1177/0962280217708654.

Oetting, A. I., J. A. Levy, R. D. Weiss, and S. A. Murphy. 2011. "Statistical Methodology for a SMART Design in the Development of Adaptive Treatment Strategies." In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures,* edited by P. E. Shrout, K. M. Keyes, and K. Ornstein, 179–205. New York: Oxford University Press. ISBN: 978-0-19-975464-9.

Ogbagaber, S. B., J. Karp, and A. S. Wahed. 2016. "Design of Sequentially Randomized Trials for Testing Adaptive Treatment Strategies." *Statistics in Medicine* 35 (6): 840–858. https://doi.org/10.1002/sim.6747.

Oslin, D. 2005. "Managing Alcoholism in People Who Do Not Respond to Naltrexone." ClinicalTrials.gov. Accessed December 8, 2018. https://clinicaltrials.gov/ct2/show/NCT00115037.

Overall, J. E., and S. R. Doyle. 1994. "Estimating Sample Sizes for Repeated Measurement Designs." *Controlled Clinical Trials* 15 (2): 100–123. ISSN: 01972456. https://doi.org/10.1016/0197-2456(94)90015-9.

Pelham, W. E., Jr., G. A. Fabiano, J. G. Waxmonsky, A. R. Greiner, E. M. Gnagy, W. E. P. III, S. Coxe, et al. 2016. "Treatment Sequencing for Childhood ADHD: A Multiple-Randomization Study of Adaptive Medication and Behavioral Interventions." *Journal of Clinical Child & Adolescent Psychology* 45, no. 4 (2016): 396–415. ISSN: 1537-4416. https://doi.org/10/gfn9xr. pmid: 26882332.

Raudenbush, S. W., and L. Xiao-Feng. 2001. "Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Group Differences in Polynomial Change." *Psychological Methods* 6 (4): 387–401. ISSN: 1082-989X. https://doi.org/10.1037//1082-989X.6.4.387.

Robins, J. M. 1997. "Causal Inference from Complex Longitudinal Data." In *Latent Variable Modeling and Applications to Causality,* edited by M. Berkane, redacted by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, and S. Zeger, 120:69–117. New York, NY: Springer New York. ISBN: 978-0-387-94917-8. https://doi.org/10.1007/978-1-4612-1842-5_4.

Seewald, N. J., K. M. Kidwell, I. Nahum-Shani, T. Wu, J. R. McKay, and D. Almirall. 2020. "Sample Size Considerations for Comparing Dynamic Treatment Regimens in a Sequential Multiple-Assignment Randomized Trial with a Continuous Longitudinal Outcome." *Stat Methods Med Res* 29, no. 7 (2020): 1891–1912. ISSN: 0962-2802. https://doi.org/10/gf85ss.

Sertkaya, A., H.-H. Wong, A. Jessup, and T. Beleche. 2016. "Key Cost Drivers of Pharmaceutical Clinical Trials in the United States." *Clin Trials* 13 (2): 117–126. ISSN: 1740-7745, 1740-7753. https://doi.org/10.1177/1740774515625964.

Thall, P. F. 2015. "SMART Design, Conduct, and Analysis in Oncology." In *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine,* edited by M. R. Kosorok and E. E. M. Moodie, 41–54. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2015. ISBN: 978-1-61197-417-1. http://epubs.siam.org/doi/book/10.1137/1.9781611974188.

Van der Vaart, A. W. 1998. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK ; New York, NY, USA: Cambridge University Press. ISBN: 978-0-521-49603-2.

Vock, D. M., and D. Almirall. 2018. "Sequential Multiple Assignment Randomized Trial (SMART)." In *Wiley StatsRef: Statistics Reference Online,* edited by N. Balakrishnan, T. Colton, W. Everitt, F. Piegorsch, and J. L. Teugels. ISBN: 978-1-118-44511-2. https://doi.org/10.1002/9781118445112.stat08073.

Wallace, M. P., and E. E. M. Moodie. 2014. "Personalizing Medicine: A Review of Adaptive Treatment Strategies." *Pharmacoepidemiology and Drug Safety* 23 (6): 580–585. ISSN: 1099-1557 (Electronic)\r1053-8569 (Linking). https://doi.org/10.1002/pds.3606.

Walls, T. A., and J. L. Schafer, eds. 2006. *Models for Intensive Longitudinal Data.* Oxford; New York: Oxford University Press. ISBN: 978-0-19-517344-4.

Watkins, C. J. C. H. 1989. "Learning from Delayed Rewards," King's College.

Zhang, S., and C. Ahn. 2011a. "Adding Subjects or Adding Measurements in Repeated Measurement Studies Under Financial Constraints." *Statistics in Biopharmaceutical Research* 3, no. 1 (2011): 54–64. ISSN: null. https://doi.org/10/bftphq.

———. 2011b. "How Many Measurements for Time-Averaged Differences in Repeated Measurement Studies?" *Contemporary Clinical Trials* 32, no. 3 (2011): 412–417. ISSN: 1551-7144. https://doi.org/10/dwv3qn.

Zhang, S., J. Cao, and C. Ahn. 2014. “A GEE Approach to Determine Sample Size for Pre- and Post-Intervention Experiments with Dropout.” *Computational Statistics & Data Analysis* 69 (2014): 114–121. ISSN: 0167-9473. https://doi.org/10/ggh65b.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian. 2015. “Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes.” *Biometrics* 71 (4): 895–904. ISSN: 0006341X. https://doi.org/10.1111/biom.12354.