Title

Sample size considerations for micro-randomized trials with binary proximal outcomes

Journal

Statistics in Medicine, 42(16)

ISSN

0277-6715

Authors

Cohn, Eric R

Qian, Tianchen

Murphy, Susan A

Publication Date

2023-07-20

DOI

10.1002/sim.9748

HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2023 July 20.

Published in final edited form as:

Stat Med. 2023 July 20; 42(16): 2777–2796. doi:10.1002/sim.9748.

Sample Size Considerations for Micro-Randomized Trials with Binary Proximal Outcomes

Eric R. Cohn,
Department of Biostatistics, Harvard University.

Tianchen Qian*,
Department of Statistics, University of California, Irvine.

Susan A. Murphy
Department of Statistics, Harvard University.

Abstract

Micro-randomized Trials (MRTs) are a novel experimental design for developing mobile health interventions. Participants are repeatedly randomized in an MRT, resulting in longitudinal data with time-varying treatments. Causal excursion effects are the main quantities of interest in MRT primary and secondary analyses. We consider MRTs where the proximal outcome is binary and the randomization probability is constant or time-varying but not data-dependent. We develop a sample size formula for detecting a nonzero marginal excursion effect. We prove that the formula guarantees power under a set of working assumptions. We demonstrate via simulation that violations of certain working assumptions do not affect the power, and for those that do, we point out the direction in which the power changes. We then propose practical guidelines for using the sample size formula. As an illustration, the formula is used to size an MRT on interventions for excessive drinking. The sample size calculator is implemented in R package `MRTSampleSizeBinary` and an interactive R Shiny app. This work can be used in trial planning for a wide range of MRTs with binary proximal outcomes.

Keywords

Causal excursion effect; Causal inference; Longitudinal data analysis; Mobile health; Sample size calculation

1 Introduction

Mobile health interventions target healthy behavior change and are delivered through mobile devices such as smartphones and wearable trackers. They are usually delivered in the form of push notifications, text messages, or audible pings. They have the potential to be delivered

*To whom correspondence should be addressed. t.qian@uci.edu.

9Supplementary Materials

The following are included in the Supplementary Materials. Section A derives the large sample distribution of $\hat{\beta}$ in Section 3.1. Section B lists a set of weaker working assumptions under which Theorem 1 holds and proves Theorem 1. Section C presents details of the generative models in the simulation. Section D includes additional simulation results.

to each individual at the time and in the context they are most likely to benefit. To realize this potential, it is important to gather empirical evidence to inform when and under what context the interventions are the most beneficial in order to improve and optimize them.

The micro-randomized trial (MRT) is an optimization trial design that provides data to answer such questions[1–4]. In an MRT, each participant is repeatedly randomized among multiple options of an intervention, usually hundreds or thousands of times throughout the trial. Each of such times is referred to as a decision point. After each decision point, a near-term, proximal outcome is measured, which is typically an outcome that the intervention is directly targeting. Data from MRTs allow researchers to investigate whether the intervention is effective on average and whether/how the intervention effect is moderated by contextual information[5,6]. This can lead to further optimization of the intervention through techniques such as reinforcement learning[7].

We consider MRTs where the proximal outcome following each decision point is binary. This is common because a natural proximal outcome for many mobile health interventions measures whether the participant adheres to the notification or message. For example, in the Substance Abuse Research Assistant (SARA) MRT for developing non-monetary incentives to improve self-report completion rate[8], one intervention is an inspirational quote whose delivery is micro-randomized every day at 4pm, and the proximal outcome is whether the participant completes the self-report that evening. In the BariFit MRT for developing smartphone-based support for weight maintenance post-bariatric surgery[9], one intervention is a message reminder for completing the daily food log, and the proximal outcome is whether the participant completes the food log on that day.

An MRT is typically sized to ensure adequate power for detecting a clinically meaningful average effect of the intervention (i.e., a marginal causal effect). Due to the lack of readily-available methodology and software to determine the sample size for MRTs with binary proximal outcomes, researchers currently rely on simulation-based sample size calculation or the sample size formula for MRTs with continuous outcomes[1]. The former can be time-consuming and the latter is inappropriate for binary outcomes.

We develop a sample size formula for MRTs with binary proximal outcomes where the randomization probability is constant or time-varying but not data-dependent. We prove that under a set of working assumptions, the sample size formula guarantees desired power to detect a pre-specified marginal causal effect with type I error control. We demonstrate via simulation that violations of certain working assumptions do not affect the power, and for those that do, we point out the direction in which the power changes. We provide practical guidelines for using the sample size formula. A sample size calculator is implemented in R package `MRTSampleSizeBinary`[10] and an interactive web app (https://tqian.shinyapps.io/mrt_ss_binary/). This work can be used in trial planning for a wide range of MRTs with binary proximal outcomes.

The rest of the paper is organized as follows. In Section 2 we present notation and review the marginal excursion effect, a key quantity in MRT primary analysis. In Section 3 we list the working assumptions and derive the sample size formula. In Section 4 we provide a

summary of the simulation findings and a list of practical guidelines for using the sample size formula. In Section 5 we illustrate the use of the sample size formula by applying it to sizing a real trial. Section 6 presents details of the simulation study. Section 7 concludes with a discussion.

2 Preliminaries

2.1 Notation

We focus on settings where (i) each participant is in the MRT for the same number of decision points, (ii) the treatment option at each decision point is binary (e.g., delivering or not delivering a message), (iii) the timing of decision points are pre-determined (e.g., at fixed calendar times determined before the study), and (iv) the randomization probability at each decision point is constant, or dependent on the decision point index solely and not dependent on other time-varying information such as the outcomes at previous decision points.

Let n denote the number of participants in an MRT and m the total number of decision points for each participant. For the i-th participant, denote by $A_{it} \in \{0, 1\}$ their randomized treatment assignment at decision point t. For example, $A_{it} = 1$ if a treatment is delivered to participant i at decision point t, and 0 if not. Let p_t denote the randomization probability at t. Let $Y_{i,t+1} \in \{0, 1\}$ denote the binary proximal outcome measured following decision point t. Without loss of generality, suppose $Y_{i,t+1} = 1$ is the desired outcome. For example, $Y_{i,t+1} = 1$ if participant i adheres to the push notification at decision point t, and 0 if not.

In an MRT, there may be decision points when it is inappropriate or unethical to deliver a treatment. For example, for safety reasons a mobile health app is usually designed so that a treatment is never delivered when the participant is detected to be driving. At such decision points, a participant is considered "unavailable" or "ineligible for randomization," randomization does not occur, and the only possible treatment option is "no treatment." Let I_{it} denote the availability indicator: $I_{it} = 1$ if participant i is available at decision point t, and $I_{it} = 0$ if not. A_{it} is always 0 if $I_{it} = 0$. Taking availability into account, the randomization probability is defined as $p_t = P(A_{it} = 1 \mid I_{it} = 1)$.

MRT data usually also include participants' baseline and time-varying covariates. Appropriately adjusting for them can increase estimation precision in the primary and secondary analyses[4]. We do not consider covariates in the sample size calculation to avoid the need to specify the covariate-outcome relationship during trial planning; this often makes the sample size formula conservative. We consider the following longitudinal observations for participant i: $O_i = (I_{i1}, A_{i1}, Y_{i2}, \ldots, I_{im}, A_{im}, Y_{i,m+1})$. We assume that $\{O_i : i = 1, \ldots, n\}$ are independent and identically distributed (i.i.d.) draws from an unknown distribution P^*. We use letters without subscript i to denote variables from a generic participant. We use $\mathbb{R}^p$ to denote the p-dimensional Euclidean space.

2.2 Marginal Excursion Effect

The marginal excursion effect (MEE)[6] is the main quantity of interest in MRT primary analysis and one of the parameters in the sample size formula. We use potential outcomes notation[11,12] to define the effect. Let $\bar{A}_t = (A_1, A_2, \ldots, A_t)$ denote the vector of treatment assignments up to t and $\bar{a}_t = (a_1, a_2, \ldots, a_t)$ a realization of $\bar{A}_t$. Let $Y_{t+1}(\bar{a}_t)$ denote the potential proximal outcome that would have been observed if the participant were assigned $\bar{a}_t$, a treatment sequence up to and including decision point t. Similarly, $I_t(\bar{a}_{t-1})$ denotes the potential availability at time t under treatment sequence $\bar{a}_{t-1}$. MEE for binary proximal outcomes is defined as[6]

$$\text{MEE}(t) = \log \frac{P\left\{Y_{t+1}\left(\bar{A}_{t-1}, 1\right) = 1 \mid I_t\left(\bar{A}_{t-1}\right) = 1\right\}}{P\left\{Y_{t+1}\left(\bar{A}_{t-1}, 0\right) = 1 \mid I_t\left(\bar{A}_{t-1}\right) = 1\right\}}, \text{ for } t = 1, \ldots, m\,. \tag{1}$$

An MEE(t) value greater than 0 indicates that the treatment is effective at decision point t. The two probabilities in (1) are conditional on the decision point being available $\left(I_t\left(\bar{A}_{t-1}\right) = 1\right)$, because scientifically the treatment effect is of interest only at available decision points, and statistically the treatment effect at unavailable decision points cannot be identified without further assumptions.

Each treatment $A_t (1 \leq t \leq m)$ is randomly assigned according to the randomization probability p_t. We refer to this assignment mechanism as the treatment protocol. MEE(t) is called an *excursion* effect because it is a contrast between what would happen to the proximal outcome under two excursions from the treatment protocol: following the treatment protocol until $t-1$ then always assigning treatment at t (the corresponding potential outcome being $Y_{t+1}\left(\bar{A}_{t-1}, 1\right)$), and following the treatment protocol until $t-1$ then always assigning no treatment at t (the corresponding potential outcome being $Y_{t+1}\left(\bar{A}_{t-1}, 0\right)$).

MEE(t) is a marginal effect because it is not conditional on any history information (such as past treatment assignments and covariate values). It is possible that treatments assigned at earlier decision points, such as A_{t-1} and A_{t-2}, have an impact on the current proximal outcome, Y_{t+1}. Such delayed effects may be attributed to habit formation (a positive delayed effect) or user-burden/habituation (a negative delayed effect). It is also likely that the effect of the treatment would vary according to certain baseline or time-varying covariate values (i.e., the existence of effect modification). We are not ruling out such possibilities by focusing on MEE(t). Rather, as a marginal quantity MEE(t) averages over any such delayed effects or effect modification. The focus on marginal quantities in the primary analysis is consistent with other optimization trials such as the factorial design[13]. Delayed effects and effect modification may be further explored in secondary and exploratory analyses.

Because the randomization probability p_t may only depend on the decision point index but not other history information, MEE(t) can be expressed in terms of observed data distribution under standard causal assumptions (consistency, positivity, ignorability):

$$\mathrm{MEE}(t) = \log \frac{P(Y_{t+1} = 1 \mid A_t = 1, I_t = 1)}{P(Y_{t+1} = 1 \mid A_t = 0, I_t = 1)}. \quad (2)$$

The exact formulation of the causal assumptions and the justification of this result is in Section 3 of Qian et al.[6].

3 Methods

3.1 A Test Statistic That Guarantees Type I Error Control

We develop the sample size formula based on a test statistic for testing the null hypothesis

$$H_0: \mathrm{MEE}(t) = 0 \text{ for } 1 \le t \le m$$

against the alternative hypothesis

$$H_1: \mathrm{MEE}(t) \neq 0 \text{ for some } t \in \{1,2,\ldots,m\}.$$

An omnibus test that aims to detect every possible MEE(t) under H_1 will have low power to detect nonzero MEE(t) in specific trends in t[14]. Instead, we propose a test statistic that will have high power against some target alternative MEE(t) ($t = 1, \ldots, m$) with a small number of degrees of freedom. The rationale is to trade off bias and variance in order to achieve high power to detect alternatives close to the target alternative. Here variance is roughly characterized by the degrees of freedom in a t-statistic, and bias is how different the true MEE(t) function is from the target alternative.

We consider the setting where MEE(t) in the target alternative is a linear function of a vector parameter $\beta_0 \neq 0$:

$$\mathrm{MEE}(t) = f(t)^T \beta_0 \text{ for } 1 \le t \le m.$$

Here $f(t)$ is a pre-specified p-dimensional vector-valued function of t and $\beta_0 \in \mathbb{R}^p$. The choice of the target alternative is usually determined in conversation with the scientific team. For example, if the scientific team does not expect the effect of the intervention to vary greatly over time, a constant-in-time MEE(t) with $f(t) = 1$ would be a good choice for the target alternative. If, however, the scientific team conjectures that the effect of the intervention might be close to zero and gradually increase with time early in the study and possibly decrease later in the study, a quadratic target alternative with $f(t) = \left(1, t, t^2\right)^T$ might be a good choice.

We construct the test statistic as follows. Let $g(t)^T \alpha$ be a working model for $\log E(Y_{t+1} \mid A_t = 0, I_t = 1)$, where $g(t)$ is a q-dimensional feature vector and $\alpha \in \mathbb{R}^q$. We require that $p_t f(t)$ is a subset of $g(t)$[1]. Let $(\hat{\alpha}, \hat{\beta})$ denote the solution to the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m} I_{it}\left\{e^{-(A_{it}-p_t)f(t)^T\beta}Y_{i,t+1}-e^{g(t)^T\alpha}\right\}\begin{bmatrix} g(t) \\ (A_{it}-p_t)f(t)\end{bmatrix}=0. \tag{3}$$

Equation (3) is a modified version of equation (10) in Qian et al.[6], in that we replaced their $\exp\{-A_{it}f(t)^T\beta\}$ by $\exp\{-(A_{it}-p_t)f(t)^T\beta\}$; this modification is to enable the derivation of an analytic sample size formula in Section 3.2. Suppose the data generating distribution P^* satisfies $\mathrm{MEE}(t)=f(t)^T\beta (t=1,\ldots,m)$ for some $\beta\in\mathbb{R}^p$. We show in Section A of the Supplementary Materials that with large n, $\hat{\beta}$ approximately follows a normal distribution:

$$\hat{\beta}\approx N\left(\beta,\frac{1}{n}\hat{M}^{-1}\hat{\Sigma}\hat{M}^{-1,T}\right), \tag{4}$$

where

$$\begin{aligned}\widehat{M}&=\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m} I_{it}e^{-(A_{it}-p_t)f(t)^T\hat{\beta}}Y_{i,t+1}(A_{it}-p_t)^2f(t)f(t)^T,\\ \widehat{\Sigma}&=\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}\sum_{s=1}^{m} I_{it}I_{is}r_{it}(\hat{\alpha},\hat{\beta})r_{is}(\hat{\alpha},\hat{\beta})(A_{it}-p_t)(A_{is}-p_s)f(t)f(s)^T,\\ r_{it}(\hat{\alpha},\hat{\beta})&=e^{-(A_{it}-p_t)f(t)^T\hat{\beta}}Y_{i,t+1}-e^{g(t)^T\hat{\alpha}}.\end{aligned} \tag{5}$$

We consider the Wald-type test statistic

$$T=n\hat{\beta}^T(\hat{M}^{-1}\hat{\Sigma}\hat{M}^{-1,T})^{-1}\hat{\beta}. \tag{6}$$

Setting $\beta=0$ in (4) implies that under H_0 the large sample distribution of T is χ^2_p, a chi-squared distribution with p degrees of freedom. Thus, a hypothesis test that uses the critical value of the chi-squared distribution will have nominal type I error control asymptotically. To correct the downward bias of the sandwich estimator $\hat{M}^{-1}\hat{\Sigma}\hat{M}^{-1,T}$ when the sample size n is small, we use the critical value from a scaled F-distribution because $\frac{n-q-p}{p(n-q-1)}T$ approximately follows $F_{p,n-q-p}$, an F-distribution with degrees of freedom $(p,n-q-p)$[1,15]. In particular, the rejection region of a test with significance level η is

$$\left\{T:\frac{n-q-p}{p(n-q-1)}T>F^{-1}_{p,n-q-p}(1-\eta)\right\}, \tag{7}$$

where $F^{-1}_{p,n-q-p}$ is the quantile function of $F_{p,n-q-p}$. We further incorporated the small sample correction in Mancl and DeRouen[16] by replacing $\hat{\Sigma}$ in (6) with an adjusted version using the "hat" matrix, which improves type I error control for small sample sizes.

[1]In other words, we require that for each t, the linear span of $g(t)$ contains $p_tf(t)$. For instance, if $f(t)=(1,t)$ so that $f(t)^T\beta_0=\beta_{00}+\beta_{01}t$, and p_t is a constant (say $p_t=0.6$ for all t), then $g(t)$ must also contain at least $(1,t)$. If $f(t)=(1,t)$ and p_t is a linear function in t, then $g(t)$ must contain at least $\left(1,t,t^2\right)$.

3.2 A Sample Size Formula That Guarantees Power Under Working Assumptions

Equation (4) implies that when the data generating distribution P^* satisfies the target alternative $\text{MEE}(t) = f(t)^T\beta_0$ for all t, the test statistic T approximately follows $\chi^2_p(\lambda(n))$, a non-central chi-squared distribution with p degrees of freedom and non-centrality parameter $\lambda(n) = n\beta_0^T\left(M^{-1}\Sigma M^{-1,T}\right)^{-1}\beta_0$, where M and Σ are the probability limits of $\hat{M}$ and $\hat{\Sigma}$ as $n \to \infty$, respectively. To improve small sample performance, we use the F-distribution approximation instead. In particular, the scaled test statistic $\frac{n-q-p}{p(n-q-1)}T$ approximately follows $F_{p,n-q-p;\lambda(n)}$, a non-central F-distribution with degrees of freedom $(p, n-q-p)$ and non-centrality parameter $\lambda(n)$[1,15]. In order to have at least $1-b$ power to detect the target alternative, the sample size n must satisfy

$$P\left\{\frac{n-q-p}{p(n-q-1)}T > F^{-1}_{p,n-q-p}(1-\eta)\right\} \geq 1-b, \quad \text{where } \frac{n-q-p}{p(n-q-1)}T \sim F_{p,n-q-p;\lambda(n)}.$$

Therefore, the required sample size is the smallest integer n such that

$$1 - F_{p,n-q-p;\lambda(n)}\left\{F^{-1}_{p,n-q-p}(1-\eta)\right\} \geq 1-b. \tag{8}$$

The sample size formula (8) relies on $\lambda(n)$, which depends on the data generating distribution P^* that is typically unknown during trial planning. To make it feasible to compute n from (8), we make the following working assumptions about P^*.

(WA-a) (Known MEE.) Suppose $\text{MEE}(t) = f(t)^T\beta_0$ for $1 \leq t \leq m$, where both $f(t)$ and $\beta_0 \in \mathbb{R}^p$ are known.

(WA-b) (Known success probability under no treatment.) Suppose $E(Y_{t+1} \mid A_t = 0, I_t = 1) = e^{g(t)^T\alpha_0}$ for $1 \leq t \leq m$, where both $g(t)$ and $\alpha_0 \in \mathbb{R}^q$ are known.

(WA-c) (Known availability probability.) Suppose $E(I_t) = \tau(t)$ for $1 \leq t \leq m$, where $\tau(t)$ is known.

(WA-d) (No serial correlation in the outcome.) Suppose that, for every pair (t, s) with $1 \leq s < t \leq m$ $E(Y_{t+1} \mid I_t = 1, I_s = 1, A_t, A_s, Y_{s+1})$ does not depend on Y_{s+1}.

(WA-e) (Exogenous availability process.) Suppose I_t is independent of prior treatments or prior outcomes; i.e., $I_t \perp \{A_s, Y_{s+1} : 1 \leq s < t\}$ for $1 \leq t \leq m$.

Under the working assumptions, the sample size formula is summarized in the following theorem, and all the inputs to the formula are listed in Table 1.

Theorem 1.

(Sample size formula under working assumptions.) Suppose the data generating distribution P^* satisfies (WA-a)-(WA-e), and $\beta_0 \neq 0$. Suppose $p_t f(t)$ is a subset of $g(t)$. Then the probability limits of $\hat{M}$ and $\hat{\Sigma}$ as $n \to \infty$ are

$$M = \sum_{t=1}^{m} \tau(t) e^{p_t f(t)^T \beta_0 + g(t)^T \alpha_0} (1 - p_t) p_t f(t) f(t)^T,$$
$$\Sigma = \sum_{t=1}^{m} \tau(t) e^{2 p_t f(t)^T \beta_0 + g(t)^T \alpha_0} (1 - p_t) p_t \left[(1 - p_t) e^{-f(t)^T \beta_0} + p_t - e^{g(t)^T \alpha_0} \right] f(t) f(t)^T. \quad (9)$$

Furthermore, let $\lambda(n) = n \beta_0^T \left(M^{-1} \Sigma M^{-1,T} \right)^{-1} \beta_0$, and let n_0 be the smallest integer n that satisfies the sample size formula (8). Then for n_0 i.i.d. samples from P^*, the testing procedure (7) has at least $1 - b$ power.

We prove Theorem 1 under a set of weaker but less interpretable working assumptions in Section B of the Supplementary Materials.

(WA-a)-(WA-c) assume the knowledge of certain parameter values about P^*, and (WA-d)-(WA-e) assume specific independence properties about P^*. In particular, (WA-a) specifies $\text{MEE}(t)$, $1 \leq t \leq m$, for the target alternative. We treat it as a working assumption to assess the performance of the sample size formula when it is violated, i.e., when the true $\text{MEE}(t)$ under P^* is different than the target alternative.

(WA-b) states that the researcher knows $E(Y_{t+1} \mid A_t = 0, I_t = 1)$, the success probability of the binary outcome under no treatment at t, and how it changes over time. We use "success probability null curve" to refer to

$$\text{SPNC}(t) := E(Y_{t+1} \mid A_t = 0, I_t = 1). \quad (10)$$

One caveat is that $E(Y_{t+1} \mid A_t = 0, I_t = 1)$ averages over the past treatments $(A_1, A_2, \ldots, A_{t-1})$ and past outcomes $(Y_2, Y_3, \ldots, Y_t)$, and thus $\text{SPNC}(t)$ depends on the magnitude of potential delayed effect and serial correlation.

(WA-c) states that for each decision point, the researcher knows the probability of a participant being available. Note that if (WA-e) is violated, $\tau(t)$ would incorporate how I_t depends on previous A_s and $Y_{s+1} (1 \leq s < t)$.

(WA-d) states that the outcome does not depend on previous outcomes given previous treatments. This assumption is made to facilitate an analytic sample size formula. This assumption is implausible in most mobile health applications because for the same participant, outcomes measured closer in time are likely correlated. In the simulation studies we will see that the sample size formula still performs well when (WA-d) is violated.

(WA-e) states that the availability indicator, I_t, does not depend on prior treatments and prior outcomes. This assumption may or may not be plausible depending on the particular study. For example, (WA-e) is plausible if a decision point is unavailable due to a technical glitch

unrelated to the participant's behavior or treatment. (WA-e) will be violated if availability reflects burden considerations in that a decision point is unavailable when treatment is delivered at a recent decision point. Note that (WA-e) holds for MRTs with no availability considerations because I_t will always be 1.

4 Performance of the Sample Size Formula under Working Assumption Violations and Practical Guidelines

We conducted extensive simulation studies to evaluate the performance of the sample size formula when all working assumptions hold and when certain working assumptions are violated. Here we summarize the performance of the formula and provide practical guidelines. Detailed simulation results are in Section 6.

We start with definitions necessary for summarizing the simulation findings. Define the average treatment effect (ATE) as

$$\text{ATE} = \frac{\sum_{t=1}^{m} E(Y_{t+1} \mid A_t = 1, I_t = 1)E(I_t)}{\sum_{t=1}^{m} E(Y_{t+1} \mid A_t = 0, I_t = 1)E(I_t)}, \tag{11}$$

the average success probability under the null (ASPN) as

$$\text{ASPN} = \frac{\sum_{t=1}^{m} E(Y_{t+1} \mid A_t = 0, I_t = 1)E(I_t)}{\sum_{t=1}^{m} E(I_t)}, \tag{12}$$

and the average availability (AA) as

$$\text{AA} = \frac{1}{m}\sum_{t=1}^{m} E(I_t)\,. \tag{13}$$

ATE is the multiplicative causal excursion effect averaged over time and weighted by availability (Remark 5 of Qian et al.[6]). ASPN is $\text{SPNC}(t) := E(Y_{t+1} \mid A_t = 0, I_t = 1)$ averaged over time and weighted by availability. AA is $\tau(t) := E(I_t)$ averaged over time. ATE, ASPN, and AA summarize the magnitude of $\text{MEE}(t)$, $\text{SPNC}(t)$, and $\tau(t)$, respectively. In addition, we will use "pattern" to refer to how $\text{MEE}(t)$, $\text{SPNC}(t)$, and $\tau(t)$ vary over time *independently of their magnitude*. As we will see, the magnitude and the pattern of $\text{MEE}(t)$, $\text{SPNC}(t)$, and $\tau(t)$ impact the performance of the sample size formula in different ways.

We distinguish two versions of each of the quantities: one corresponding to the true data generating distribution (denoted by superscript $*$) and the other corresponding to the input to the sample size formula (denoted by superscript "w", meaning "working"). For example, ATE^* is (11) with the expectations calculated according to the true data generating distribution P^* (which is unknown unless in simulations), and ATE^{w} is (11) with the expectations calculated using the input to the sample size formula assuming all working assumptions hold. Using this notation, (WA-a) is equivalent to $\text{MEE}^{\text{w}}(t) = \text{MEE}^*(t)$, (WA-b) is equivalent to $\text{SPNC}^{\text{w}}(t) = \text{SPNC}^*(t)$, and (WA-c) is equivalent to $\tau^{\text{w}}(t) = \tau^*(t)$.

The simulation findings are as follows. Regardless of whether the working assumptions hold, the type I error rate is always controlled at the desired 0.05 level. So we will focus on power next. When all the working assumptions hold, the MRT is adequately powered. When certain working assumptions are violated, the performance of the sample size formula depends on the discrepancy between the ∗-quantities and the w-quantities (Table 2). In general, correctly specifying the magnitudes of $\text{MEE}(t)$, $\text{SPNC}(t)$, and $\tau(t)$ is critical for adequate power. The performance is robust to the patterns of $\text{MEE}(t)$, $\text{SPNC}(t)$, and $\tau(t)$ in that as long as one uses constant $\text{MEE}^{\text{W}}(t)$ and $\text{SPNC}^{\text{W}}(t)$ the power will be adequate. When there is a delayed effect or when the outcome is serially correlated, the MRT is adequately powered if ASPN^{W} accounts for the delayed effect / serial correlation so that $\text{ASPN}^{\text{W}} = \text{ASPN}^{*}$; otherwise, the power will depend on the sign of the delayed effect / serial correlation. When availability depends on past treatments and outcomes, the MRT can be slightly under-powered for some generative models.

The practical guidelines for specifying the inputs to the sample size formula are listed in Table 3.

5 Application

We illustrate the use of the sample size formula by determining the sample size of the Drink Less MRT. The Drink Less MRT was conducted in 2021, aimed at optimizing the Drink Less smartphone app to help people reduce harmful alcohol consumption[17]. Participants were randomized every day at 8 pm for 30 days, each time with probability 0.6 to receive an engagement push notification and with probability 0.4 to receive nothing. The protocol of the study can be found in Bell et al.[18]. The considerations presented here are simplified for the purpose of illustrating the use of the sample size formula. Participants were always considered available during the Drink Less MRT, but we will later consider a hypothetical situation where participants can sometimes be unavailable to illustrate this aspect of the sample size formula.

According to the MRT design we set $m = 30,\ p_t = 0.6$ and $\tau(t) = 1,\ 1 \le t \le m$, and we set the desired power at 0.8 and the type I error level at 0.05. Following the recommendations in Section 4, for the remaining inputs to the sample size formula (Table 1) we set both $\text{MEE}(t) = f(t)^T \beta_0$ and $\text{SPNC}(t) = \exp\{g(t)^T \alpha_0\}$ as constant. Thus, $f(t) = g(t) = 1$, and β_0 and α_0 are completely determined by ATE and ASPN, respectively. Figure 1 shows how n varies with ATE and ASPN under this set of inputs. The sample size of the MRT should then be determined based on the conjectured values of ATE and ASPN through conversations with domain experts. For this illustration, suppose we choose $\text{ATE} = 1.15$ and $\text{ASPN} = 0.3$. Then the resulting sample size is $n = 123$.

Based on the performance of the sample size formula when working assumptions are violated (Section 4), as long as one is conservative in specifying ATE and ASPN for constant $\text{MEE}(t)$ and $\text{SPNC}(t)$, the output sample size ($n = 123$) should guarantee the desired power even if the true $\text{MEE}(t)$ and $\text{SPNC}(t)$ are not constant. Nonetheless, it is always helpful to explore how sensitive n is to the inputs. Suppose we keep $\text{ATE} = 1.15$ and $\text{ASPN} = 0.3$, and

consider as inputs a variety of linear or quadratic MEE(t) and SPNC(t) parameterized by θ_f and θ_g (see Figure 4 and Section 6.1 for the parameterization). Figure 2 shows how n depends the pattern of the input MEE(t) and SPNC(t). When the input MEE(t) is linear or quadratic, n is mostly larger than when MEE(t) is constant (Figure 2a). It is true even when $\theta_f = 0$ so that the linear or quadratic MEE(t) becomes a flat line. This is because the extra degrees of freedom used by β_0 when MEE(t) is linear or quadratic lead to a larger n (see (8)). When the input SPNC(t) is linear or quadratic, n is smaller than when SPNC(t) is constant (Figure 2b). This implies that one may choose a nonlinear SPNC(t) to achieve a smaller n if they are confident in the prior knowledge about how the success probability of the outcome may change over time.

Last, we consider the hypothetical situation where participants can be unavailable so that $\tau(t) < 1$ at least for some t, and examine how n should be adjusted accordingly. According to the simulation results, one should consider a constant $\tau(t)$ as input to the sample size formula if the true $\tau(t)$ is not known, and the average availability (AA) determines the input $\tau(t)$. In this case, the required n increases as AA decreases, and n approximately doubles when AA decreases from 1 to 0.5 (Figure 3a). Now suppose we fix AA $= 0.7$ but vary the pattern of the input $\tau(t)$ (parameterized by θ_τ; see Figure 4 and Section 6.1 for the parameterization). Figure 3b shows that different patterns of $\tau(t)$ have minimal impact on n. Therefore, it is safe to set a constant $\tau(t)$ and choose n based on a conjectured AA value.

6 Detailed Simulation Results

6.1 Generative Models

Throughout, we set the desired type I error to be $\eta = 0.05$ and the desired power to be $1 - b = 0.8$. We set the total number of decision points per individual as $m = 30$. We used 2000 repetitions for each simulation. We consider a simple generative model (GM-0) and three generative models with more complicated features: one where the treatments have a delayed effect on the outcomes (GM-DE), one where the outcomes are serially correlated (GM-SC), and one with an endogenous availability process where I_t depends on previous outcomes and previous treatments (GM-EA). We first describe GM-0, then describe how the other three GMs differ from GM-0. Details about the generative models are in Section C of the Supplementary Materials.

Recall we use superscript * to denote quantities about the true data generating distribution and use superscript "w" to denote inputs to the sample size formula. GM-0 is characterized by the following parameters: p_t, α^*, β^*, and $g^*(t)$, $f^*(t)$, $\tau^*(t)$ for $1 \le t \le m$. p_t is the randomization probability, α^* and $g^*(t)$ dictate the success probability null curve, β^* and $f^*(t)$ dictate the marginal excursion effect, and $\tau^*(t)$ is the probability of being available. Each individual's data is generated independently as follows: for individual i, for $1 \le t \le m$, I_{it}~Bernoulli($\tau^*(t)$), A_{it}~Bernoulli(p_t) if $I_{it} = 1$ and $A_{it} = 0$ if $I_{it} = 0$, and $Y_{i,t+1}$~Bernoulli($\exp\{g^*(t)^T\alpha^* + A_{it}f^*(t)^T\beta^*\}$). This implies that MEE$^*(t) = f^*(t)^T\beta^*$ and SPNC$^*(t) \equiv E(Y_{t+1} \mid A_t = 0, I_t = 1) = \exp\{g^*(t)^T\alpha^*\}$

For GM-DE, the generation of I_{it} and A_{it} is the same as GM-0, and the success probability for $Y_{i,t+1}$ is $\exp\{g^*(t)^T\alpha^* + A_{it}f^*(t)^T\beta^* + \gamma_1^* A_{i,t-1}\}$. We generated A_{i0} from Bernoulli(p_0) with $p_0 := p_1$ to make the above display well-defined for $t = 1$. Under GM-DE, we have $\text{MEE}^*(t) = f^*(t)^T\beta^*$ and $\text{SPNC}^*(t) = \{\exp(\gamma_1^*)p_{t-1} + 1 - p_{t-1}\}\exp\{g^*(t)^T\alpha^*\}$.

For GM-SC, the generation of I_{it} and A_{it} is the same as GM-0, and the success probability for $Y_{i,t+1}$ is $\exp\{g(t)^T\alpha + A_{it}f(t)^T\beta + \gamma_2 Y_{it}\}$. We only consider the case where individuals are always available ($\tau^*(t) = 1$ for all t) and the causal effect $\text{MEE}^*(t)$ is a constant. Under GM-SC, we have $\text{MEE}^*(t) = \beta_0$, and the analytic form of $\text{SPNC}^*(t)$ is in Section C.4.1 of the Supplementary Materials.

For GM-EA, the generation of A_{it} and Y_{it} is the same as GM-0, and the success probability for I_{it} is $0.5 + \gamma_3 A_{i,t-1} + \gamma_4 Y_{it} I_{i,t-1}$. We only consider the case where both $\text{SPNC}^*(t)$ and $\text{MEE}^*(t)$ are constant. Under GM-EA, we have $\text{MEE}^*(t) = \beta_0$, and the analytic form of $\tau^*(t)$ is in Section C.5.1 of the Supplementary Materials.

To clearly present the simulation results especially when certain working assumptions are violated, we parameterized $\text{MEE}^*(t)$, $\text{SPNC}^*(t)$ and $\tau^*(t)$ (as well as $\text{MEE}^{\text{W}}(t)$, $\text{SPNC}^{\text{W}}(t)$ and $\tau^{\text{W}}(t)$) as follows. $\text{MEE}(t)$ (either $\text{MEE}^*(t)$ or $\text{MEE}^{\text{W}}(t)$) can be constant, linear, or quadratic in t. If linear or quadratic, $\text{MEE}(t)$ is parameterized in a way that it is determined by ATE and a $\theta_f \in [-1, 1]$. For example, if linear, $\theta_f = -1/0/1$ correspond to an increasing/constant/decreasing $\text{MEE}(t)$, respectively (Figure 4b). If quadratic, different θ_f values correspond to different quadratic patterns (Figure 4c). If constant, ATE alone determines $\text{MEE}(t)$ and no θ_f is needed (Figure 4a). For $\text{SPNC}(t)$, we used a similar parameterization with parameter θ_g to characterize various patterns of its log-transformation, $g(t)^T\alpha$ (Figure 4d–4f). The range of θ_g is narrower than [−1, 1] to ensure that the success probability is always within [0, 1]. For $\tau(t)$, we considered a linear and a periodic pattern, both parameterized by θ_τ, along with a constant pattern (Figure 4g). Details about the parameterization are presented in Sections C.2.1–C.2.3 of the Supplementary Materials.

6.2 Simulation Results When All Working Assumptions Hold

Using GM-0, we conducted simulations under a variety of settings with various patterns and magnitudes of $\text{MEE}^*(t)$, $\text{SPNC}^*(t)$, and $\tau^*(t)$ when all working assumptions hold, that is, when $\text{MEE}^{\text{W}}(t) = \text{MEE}^*(t)$, $\text{SPNC}^{\text{W}}(t) = \text{SPNC}^*(t)$, and $\tau^{\text{W}}(t) = \tau^*(t)$. ((WA-d) and (WA-e) automatically hold under GM-0.) We considered 1,372 simulation settings consisting of a combination of constant/linear p_t, constant/linear/quadratic $\text{MEE}(t)$ with various θ_f values, constant/log-linear/log-quadratic $\text{SPNC}(t)$ with various θ_g values (subject to the constraint that $p_t f(t)$ is a subset of $g(t)$), constant/linear/periodic $\tau(t)$ with various θ_τ values, and a variety of values for ATE, ASPN, and AA. Details for the simulation settings are listed in Section E of the Supplementary Materials. In all settings, the power is always about the desired level (Figure 5). The power is slightly lower than the desired 0.8 (around 0.78) for settings where the output sample size n is small (close to 20), likely due to the hat-matrix-based small sample correction[16] employed in the estimator.

6.3 Simulation Results When (WA-a) Is Violated

Suppose that (WA-a) is violated and the remaining working assumptions are satisfied.

Under GM-0, we considered two ways (WA-a) can be violated: the magnitude of MEE(t) is misspecified ($\text{ATE}^{\text{W}} \neq \text{ATE}^*$), or the pattern of MEE(t) is incorrect. The pattern of MEE(t) being incorrect means that either the polynomial degree of MEE*(t) is different from $\text{MEE}^{\text{W}}(t)$ (for example, one is constant and the other is non-constant linear), or that the two are of the same polynomial degree but $\theta_f^* \neq \theta_f^w$.

When the pattern of MEE(t) is correct but its magnitude is misspecified, if $\text{ATE}^{\text{W}} > \text{ATE}^*$ then the MRT is under-powered; if $\text{ATE}^{\text{W}} < \text{ATE}^*$ then the MRT is over-powered (Figure 6). This is expected because, with a fixed power, a larger effect size would correspond to a smaller sample size. In addition, the power curves against $\text{ATE}^*/\text{ATE}^{\text{W}}$ are all clustered together for a variety of SPNC*(t) and a variety of patterns of MEE*(t). This indicates that the impact of misspecifying ATE is so substantial that it overwhelms the other aspects of the generative model.

When the magnitude of MEE(t) is correct ($\text{ATE}^{\text{W}} = \text{ATE}^*$) but its pattern is misspecified, whether the MRT is adequately powered depends on the type of misspecification. If the input $\text{MEE}^{\text{W}}(t)$ is constant but the truth MEE*(t) is linear or quadratic, the MRT is adequately powered (Figure 7a). On the contrary, if the input $\text{MEE}^{\text{W}}(t)$ is linear or quadratic but the truth MEE*(t) is constant, the MRT is under-powered unless θ_f^w is chosen close to 0 so that $\text{MEE}^{\text{W}}(t)$ is nearly constant (Figure 7b). The same result was observed over a range of SPNC(t) patterns and ATE and ASPN values, shown by the clustered curves. Section D.1 of the Supplementary Materials includes additional simulation results where one of $\text{MEE}^{\text{W}}(t)$ and MEE*(t) is linear and the other is quadratic, or the two are of the same polynomial shape but $\theta_f^w \neq \theta_f^*$. The conclusion is that when the pattern of MEE(t) is misspecified, the desired power is guaranteed only if the input $\text{MEE}^{\text{W}}(t)$ is constant; when $\text{MEE}^{\text{W}}(t)$ is linear or quadratic and is different from MEE*(t), the MRT can be severely under-powered.

6.4 Simulation Results When (WA-b) Is Violated

Suppose that (WA-b) is violated and the remaining working assumptions are satisfied.

Under GM-0, we considered two ways (WA-b) can be violated: the magnitude of SPNC(t) is misspecified ($\text{ASPN}^{\text{W}} \neq \text{ASPN}^*$), or the pattern of SPNC(t) is incorrect. The pattern of SPNC(t) being incorrect means that either the polynomial degree of SPNC*(t) is different from $\text{SPNC}^{\text{W}}(t)$, or that the two are of the same polynomial degree but $\theta_g^* \neq \theta_g^w$.

When the pattern of SPNC(t) is correct and its magnitude is misspecified, if $\text{ASPN}^{\text{w}} > \text{ASPN}^*$ then the MRT is under-powered; if $\text{ASPN}^{\text{W}} < \text{ASPN}^*$ then the MRT is over-powered (Figure 8). The amount of over-/under-power is moderated by the value of ASPN*, shown by the three clusters of curves. Therefore, if one is uncertain about the magnitude of the success probability under no treatment, one should specify a smaller ASPN^{W}.

When the magnitude of SPNC(t) is correct (ASPN^{W} = ASPN*) but the pattern of SPNC(t) is misspecified, whether the MRT is adequately powered depends on the type of misspecification. If the input SPNC^{W}(t) is constant but the truth SPNC*(t) is linear or quadratic, the MRT is adequately powered (Figure 9a). On the contrary, if the input SPNC^{W}(t) is linear or quadratic but the truth SPNC*(t) is constant, the MRT is under-powered unless θ_g^w is chosen close to 0 so that SPNC^{W}(t) is nearly constant (Figure 9b). Section D.2 of the Supplementary Materials includes additional simulation results where one of SPNC^{W}(t) and SPNC*(t) is linear and the other is quadratic, or the two are of the same polynomial shape but $\theta_g^w \neq \theta_g^*$. The conclusion is that when the pattern of SPNC(t) is misspecified, the desired power is guaranteed only if the input SPNC^{W}(t) is constant; when SPNC^{W}(t) is linear or quadratic and is different from SPNC*(t), the MRT can be severely under-powered.

Under GM-DE, we simulated under a variety of γ_1^* values (recall γ_1^* captures the magnitude of delayed effect of A_{t-1} on Y_{t+1}). The violation of (WA-b) is such that the input SPNC^{W}(t) takes a constant, log-linear, or log-quadratic form but the true SPNC*(t) = $\{\exp(\gamma_1^*)p_{t-1} + 1 - p_{t-1}\}\exp\{g^*(t)^T\alpha^*\}$ does not take any of these three forms. When we choose the input α^W so that ASPN^{W} = ASPN*, the MRT is adequately powered regardless of the magnitude and direction of the delayed effect (Figure 10). However, when we ignored the delayed effect by setting $\alpha^W = \alpha^*$ and $g^W(t) = g^*(t)$, a positive/negative γ_1^* would result in an over-/under-powered MRT. This is because a positive/negative γ_1^* would then correspond to a positive/negative delayed effect, which in turn would correspond to SPNC^{W}(t) being less/greater than SPNC*(t) for all t and thus ASPN^{W} will be less/greater than ASPN*. The same result is observed over a range of SPNC(t) patterns and MEE(t) patterns, shown by the clustered curves.

6.5 Simulation Results When (WA-c) Is Violated

Suppose that (WA-c) is violated and the remaining working assumptions are satisfied.

Under GM-0, we considered two ways (WA-c) can be violated: the magnitude of AA is misspecified, or the pattern of $\tau(t)$ is incorrect. The pattern of $\tau(t)$ being incorrect means that either one of $\tau^*(t)$ and $\tau^W(t)$ is linear and the other is periodic, or they are both linear or periodic but $\theta_\tau^* \neq \theta_\tau^w$.

When the pattern of $\tau(t)$ is correct and its magnitude is misspecified, if AA^{W} > AA* then the MRT is under-powered; if AA^{W} < AA* then the MRT is over-powered (Figure 11). This is expected as a larger availability probability corresponds to a larger proportion of decision points that are used in the analysis. Therefore, if one is uncertain about the magnitude of the availability probability, one should be conservative and specify a smaller one. The same result is observed over a range of SPNC(t) patterns and MEE(t) patterns, shown by the clustered curves.

When the magnitude of $\tau(t)$ is correct ($\text{AA}^{\text{W}} = \text{AA}^*$) but the pattern of $\tau(t)$ is misspecified, the MRT is adequately powered as long as the input $\tau^{\text{W}}(t)$ is constant, regardless of whether the true $\tau^*(t)$ is periodic or linear (Figure 12). Therefore, one can use a constant $\tau^{\text{W}}(t)$ and focus on getting a good estimate of the average availability AA^*. The same result was observed over a range of $\text{SPNC}(t)$ patterns and $\text{MEE}(t)$ patterns, shown by the clustered curves.

6.6 Simulation Results When (WA-d) Is Violated

To simulate when (WA-d) is violated, we used GM-SC with a variety of γ_2^* values, where γ_2^* encodes the magnitude of serial correlation. $\text{SPNC}^*(t)$ depends on γ_2^* in a rather complicated way (see Section C.4.1 of the Supplementary Materials). We considered three ways to specify $\text{SPNC}^{\text{W}}(t)$: $\text{SPNC}^{\text{W}}(t) = \text{SPNC}^*(t)$ (thus $\text{ASPN}^{\text{W}} = \text{ASPN}^*$), $\text{SPNC}^{\text{W}}(t) \neq \text{SPNC}^*(t)$ but $\text{ASPN}^{\text{W}} = \text{ASPN}^*$, and $\text{SPNC}^{\text{W}}(t) \neq \text{SPNC}^*(t)$ and $\text{ASPN}^{\text{W}} \neq \text{ASPN}^*$. Figure 13 shows that as long as $\text{ASPN}^{\text{W}} = \text{ASPN}^*$ (the first two cases), the MRT will be adequately powered even if the pattern of $\text{SPNC}^{\text{W}}(t)$ is incorrect (thus (WA-b) is violated). When both $\text{SPNC}^{\text{W}}(t)$ and ASPN^{W} are incorrect, a larger γ_2^* results in a larger ASPN^*, which in turn results in a higher power (solid curves in Figure 13). In summary, when the outcome has serial correlation, as long as $\text{ASPN}^{\text{W}} = \text{ASPN}^*$, the MRT will be adequately powered.

6.7 Simulation Result When (WA-e) Is Violated

To simulate when (WA-e) is violated, we used GM-EA with a variety of γ_3^* and γ_4^* values, where γ_3^* encodes the impact of A_{t-1} on I_t and γ_4^* encodes the impact of Y_t on I_t. $\tau^*(t)$ depends on γ_3^* and γ_4^* in a rather complicated way (see Section C.5.1 of the Supplementary Materials). We considered two ways to specify $\tau^{\text{W}}(t)$: $\tau^{\text{W}}(t) = \tau^*(t)$ (thus $\text{AA}^{\text{W}} = \text{AA}^*$), and $\tau^{\text{W}}(t) \neq \tau^*(t)$ and $\text{AA}^{\text{W}} \neq \text{AA}^*$. Figure 14 shows that even when $\tau^{\text{W}}(t) = \tau^*(t)$, the MRT can be slightly over- or under-powered depending on the direction of γ_3^* and γ_4^*, even though the power at most deviates from 0.8 by about 0.05. Thus, if (WA-e) might be violated in practice, the researcher can increase the output sample size slightly to be conservative.

6.8 Simulation Results When Multiple Working Assumptions Are Violated

We consider simulation settings where multiple working assumptions are simultaneously violated in the following way, which results in a factorial design of 96 generative models:

- (WA-a) is violated in that the pattern of $\text{MEE}(t)$ is misspecified but its magnitude is correct. In particular, $\text{MEE}^{\text{W}}(t)$ is constant and $\text{MEE}^*(t)$ is linear or quadratic with $\theta_f^* \in \{-0.3, 0.3\}$, and $\text{ATE}^{\text{W}} = \text{ATE}^* = 1.2$. And
- (WA-b) is violated in that the pattern of $\text{SPNC}(t)$ is misspecified but its magnitude is correct. In particular, $\text{SPNC}^{\text{W}}(t)$ is constant and $\text{SPNC}^*(t)$ is linear or quadratic with $\theta_g^* \in \{-0.3, 0.3\}$, and $\text{ASPN}^{\text{W}} = \text{ASPN}^* \in \{0.2, 0.4\}$. And
- (WA-c) is violated in that the pattern of $\tau(t)$ is misspecified but its magnitude is correct. In particular, $\tau^{\text{W}}(t)$ is constant and $\tau^*(t)$ is linear or periodic with $\theta_\tau^* \in \{0.1, 0.2\}$, and $\text{AA}^{\text{W}} = \text{AA}^* = 0.6$.

Sections 6.3–6.5 showed that each individual violations of the above does not affect the performance of the sample size formula. Figure 15 showed that when (WA-a), (WA-b), and (WA-c) are simultaneously violated in this way, the MRT is still adequately powered. This provides a stronger evidence for the practical guidelines in Table 3, that one should set MEE(t), SPNC(t), and $\tau(t)$ to constant values unless there is compelling evidence to the contrary.

7 Discussion

We considered the primary analysis of micro-randomized trials (MRT) with binary proximal outcomes and derived a sample size formula to guarantee desired power and type I error control under a set of working assumptions. The sample size formula is user-friendly with intuitive inputs. Extensive simulations showed that in order for the sample size formula to have good performance, one needs to have good knowledge about the following quantities (or be conservative when specifying them): the multiplicative average treatment effect (ATE), the average success probability of the outcome under no treatment (ASPN), and the average probability of availability (AA). The sample size formula is robust to the pattern of the marginal excursion effect over time (MEE(t)), the pattern of the success probability null curve (SPNC(t)), and the expected availability over time $\tau(t)$, in that they can be incorrectly specified as constant and the desired power will still be achieved. We provided practical guidelines on how to use the formula and illustrated the formula by sizing the Drink Less MRT.

The sample size formula is applicable to MRTs where the treatment A_t is binary. It is common for mobile health interventions to have more than two options. For example, an app may choose from multiple types of push notifications if randomized to receive treatment. To size such MRTs, one can combine all active treatments into one bucket and denote it by $A_t = 1$, and use the "any treatment vs. no treatment" comparison as the primary analysis. This way the sample size formula can still be used. Exploratory analysis can further assess the differential effect of different treatment types. If, however, one wishes to size the study for comparison across active treatments, the current method cannot be used, and we leave the sample size calculator for such settings for future research.

The sample size formula relies on knowledge of the magnitude of the treatment effect, the success probability of the outcome under no treatment, and the availability probability. When such knowledge is unavailable, an alternative is a two-stage adaptive design approach, where a pilot study is first conducted to obtain preliminary estimates of the inputs to the sample size formula, and then the sample size is re-estimated using the pilot information and additional subjects are enrolled to obtain the desired power. Such a strategy may be more efficient than a fixed-sample design, especially if uncertainty during the planning stage is high. It remains an open question of how to precisely operationalize this approach.

Another open question, as pointed out by a reviewer, is how to define a standardized effect size that combines the treatment effect and the success probability null curve. By employing such a standardized effect size, the sample size calculator may no longer depend on the the success probability null curve upon fixing the standardized effect size. However, given

the potential need for new causal effect definitions and corresponding estimators, further research is necessary to fully explore this approach.

We assumed that the number of decision points, m, is the same for all participants. This is usually a reasonable assumption when sizing the MRT. For MRTs where the actual number of decision points may differ by participant due to drop out, as a rule of thumb one may set a decreasing time trend for expected availability (i.e., treating those who dropped out as unavailable later in the study).

In deriving the sample size formula we did not consider time-varying covariates other than the decision point index. Therefore, in the sample size formula, the impact of the time-varying covariates on the outcome is marginalized and incorporated into how the causal effect and the success probability under the null change over time. In analyzing the MRT data, one should still adjust for time-fixed and time-varying covariates to improve precision, which can further increase power.

R code to reproduce the results in the paper can be downloaded at https://github.com/tqian/paper_MRTSampleSizeBinary.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors would like to thank Dr. Lauren Bell, Dr. Elizabeth Williamson, and Dr. Claire Garnett, as this project is partly motivated by collaborations with them. The authors would like to acknowledge grant support from National Institutes of Health (P50DA054039, P41EB028242, UH3DE028723).

References

[1]. Liao P, Klasnja P, Tewari A, Murphy SA. Sample size calculations for micro-randomized trials in mHealth. Statistics in medicine 2016; 35(12): 1944–1971. [PubMed: 26707831]

[2]. Dempsey W, Liao P, Klasnja P, Nahum-Shani I, Murphy SA. Randomised trials for the Fitbit generation. Significance 2015; 12(6): 20–23. [PubMed: 26807137]

[3]. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions.. Health Psychology 2015; 34(S): 1220.

[4]. Qian T, Walton AE, Collins LM, et al. The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations.. Psychological Methods 2022.

[5]. Boruvka A, Almirall D, Witkiewitz K, Murphy SA. Assessing time-varying causal effect moderation in mobile health. Journal of the American Statistical Association 2018; 113(523): 1112–1121. [PubMed: 30467446]

[6]. Qian T, Yoo H, Klasnja P, Almirall D, Murphy SA. Estimating time-varying causal excursion effects in mobile health with binary outcomes. Biometrika 2021; 108(3): 507–527. [PubMed: 34629476]

[7]. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2020; 4(1): 1–22. [PubMed: 35846237]

[8]. Rabbi M, Kotov MP, Cunningham R, et al. Toward increasing engagement in substance use data collection: development of the Substance Abuse Research Assistant app and protocol for

a microrandomized trial using adolescents and emerging adults. JMIR research protocols 2018; 7(7).

[9]. Klasnja P, Rosenberg DE, Zhou J, Anau J, Gupta A, Arterburn DE. A quality-improvement optimization pilot of BariFit, a mobile health intervention to promote physical activity after bariatric surgery. Translational Behavioral Medicine 2020.

[10]. Wong-Toi E, Dahdoul T, Qian T. MRTSampleSizeBinary: Sample Size Calculator for MRT with Binary Outcomes. 2021. R package version 0.1.0.

[11]. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies.. Journal of educational Psychology 1974; 66(5): 688.

[12]. Robins J A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. Mathematical modelling 1986; 7(9–12): 1393–1512.

[13]. Collins LM, Dziak JJ, Kugler KC, Trail JB. Factorial experiments: efficient tools for evaluation of intervention components. American journal of preventive medicine 2014; 47(4): 498–504. [PubMed: 25092122]

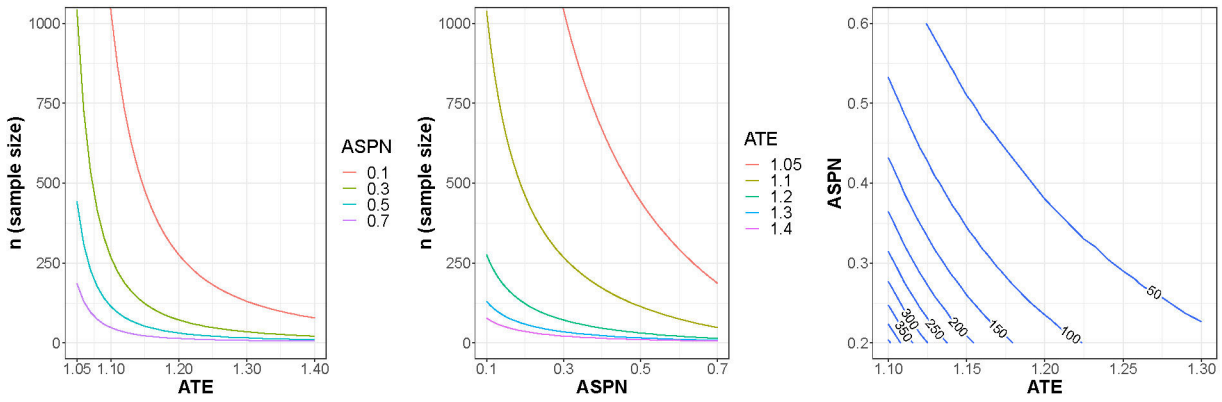
[14]. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. John Wiley & Sons. 2012.

[15]. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Statistics in medicine 2002; 21(10): 1429–1441. [PubMed: 12185894]

[16]. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001; 57(1): 126–134. [PubMed: 11252587]

[17]. Garnett C, Crane D, West R, Brown J, Michie S. The development of Drink Less: an alcohol reduction smartphone app for excessive drinkers. Translational behavioral medicine 2019; 9(2): 296–307. [PubMed: 29733406]

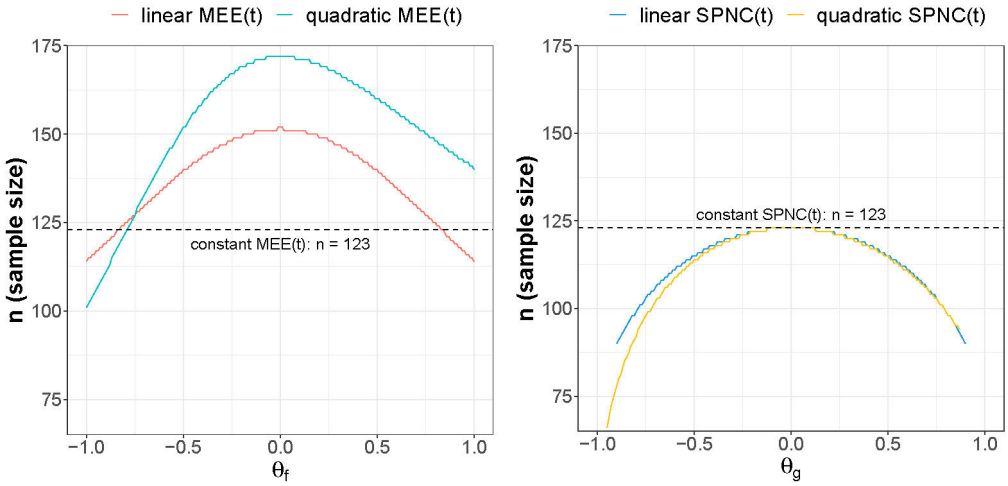
[18]. Bell L, Garnett C, Qian T, Perski O, Potts HW, Williamson E. Notifications to Improve Engagement With an Alcohol Reduction App: Protocol for a Micro-Randomized Trial. JMIR research protocols 2020; 9(8): e18690. [PubMed: 32763878]



(a) n vs. ATE (b) n vs. ASPN (c) Contour plot for n

Figure 1:

The dependence of the sample size formula output (n) on ATE and ASPN, where $m = 30,\ p_t = 0.6,\ \tau(t) = 1$, and $f(t) = g(t) = 1$ so that both MEE(t) and SPNC(t) are constant.

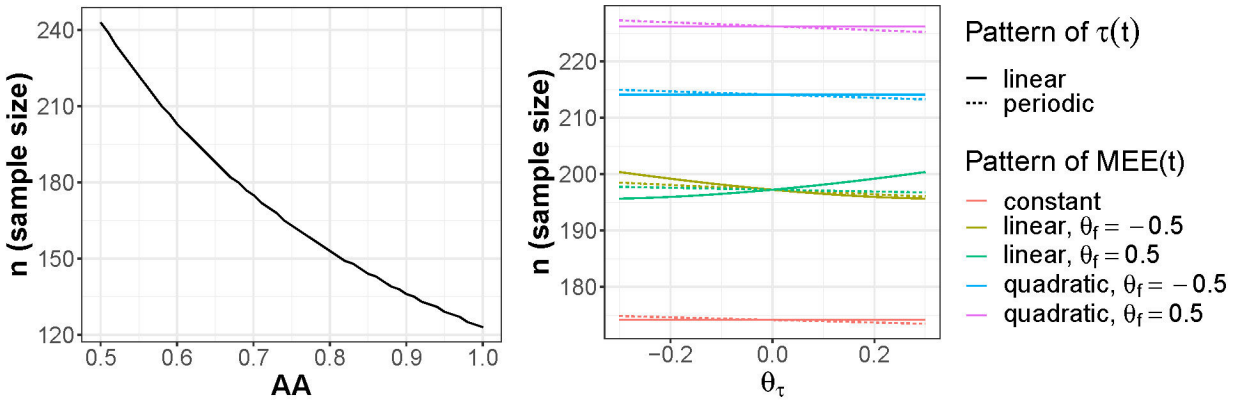


(a) n vs. pattern of MEE(t), with a constant SPNC(t)

(b) n vs. pattern of SPNC(t), with a constant MEE(t)

Figure 2:

The dependence of the sample size formula output (n) on the pattern of MEE(t) (parametrized by θ_f) and the pattern of SPNC(t) (parameterized by θ_g), where $m = 30,\ p_t = 0.6,\ \tau(t) = 1,$ ATE = 1.15 and ASPN = 0.3. The n corresponding to constant MEE(t) and constant SPNC(t) is 123, which is marked with a dashed line.



(a) n vs. AA, assuming a constant $\tau(t)$ (b) n vs. pattern of $\tau(t)$, under a few MEE(t) patterns

Figure 3:

The dependence of the sample size formula output (n) on the average availability (AA) and the pattern of $\tau(t)$ (parametrized by θ_τ), where for both panels $m = 30,\ p_t = 0.6,\ \tau(t) = 1,\ \text{ATE} = 1.15,\ \text{ASPN} = 0.3$. For the left panel, $f(t) = g(t) = 1$ so that both MEE(t) and SPNC(t) are constant. For the right panel, MEE(t) takes a few different patterns and SPNC(t) is still constant.

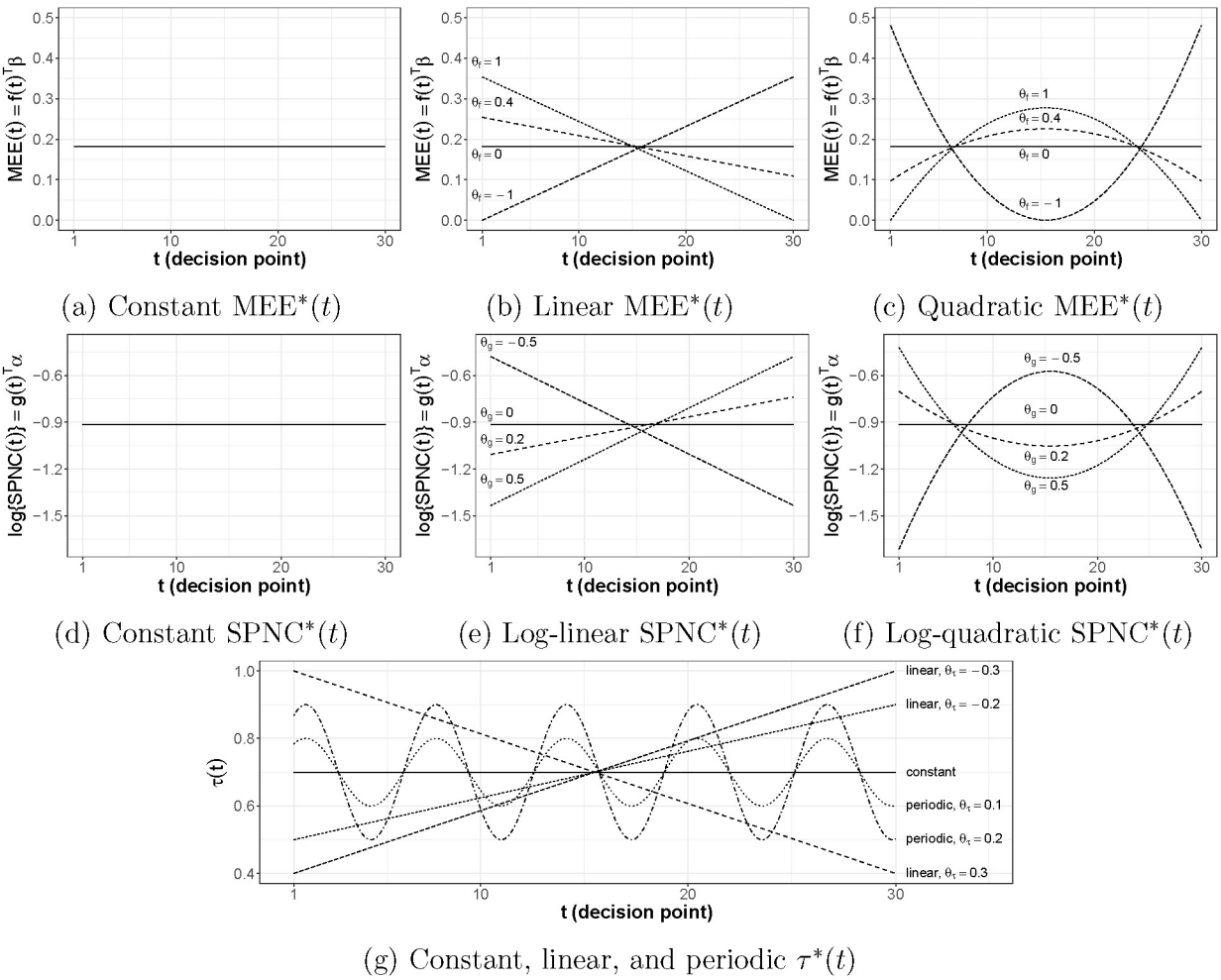


Figure 4:

(a-c) Illustration of MEE*(t) = $f^*(t)^T\beta^*$ parameterized by θ_f^* for ATE* = 1.2. (d-f) Illustration of log{SPNC*(t)} = $g^*(t)^T\alpha^*$ parameterized by θ_g^* for ASPN* = 0.4. (g) Illustration of $\tau^*(t)$ parameterized by θ_τ^* for AA* = 0.7. The figures serve to illustrate the various patterns considered. In the simulation, we also vary ATE*, ASPN*, and AA* (not shown here).

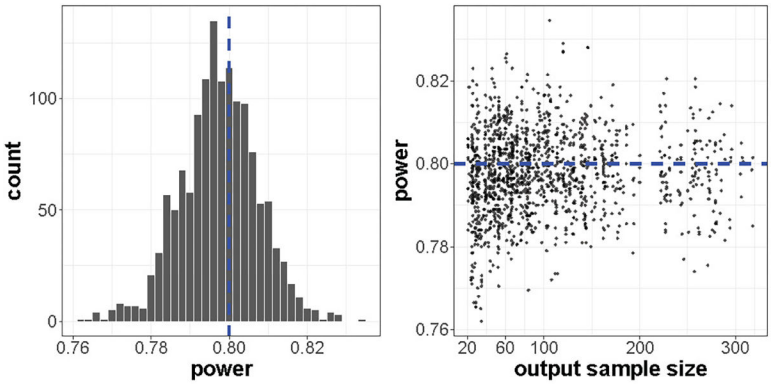


Figure 5:

Power when all working assumptions hold. The left panel shows the histogram of the power under 1,372 settings. The right panel shows the power under each setting against the output n from the sample size formula.

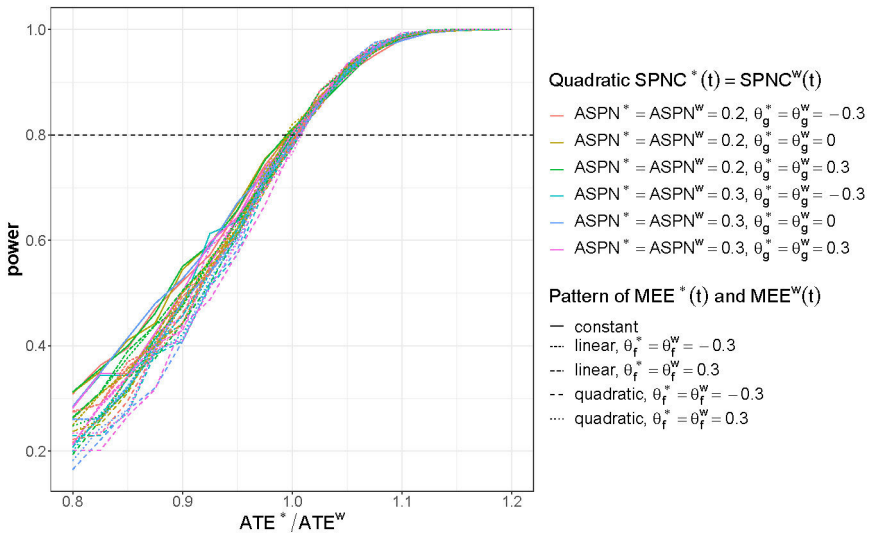
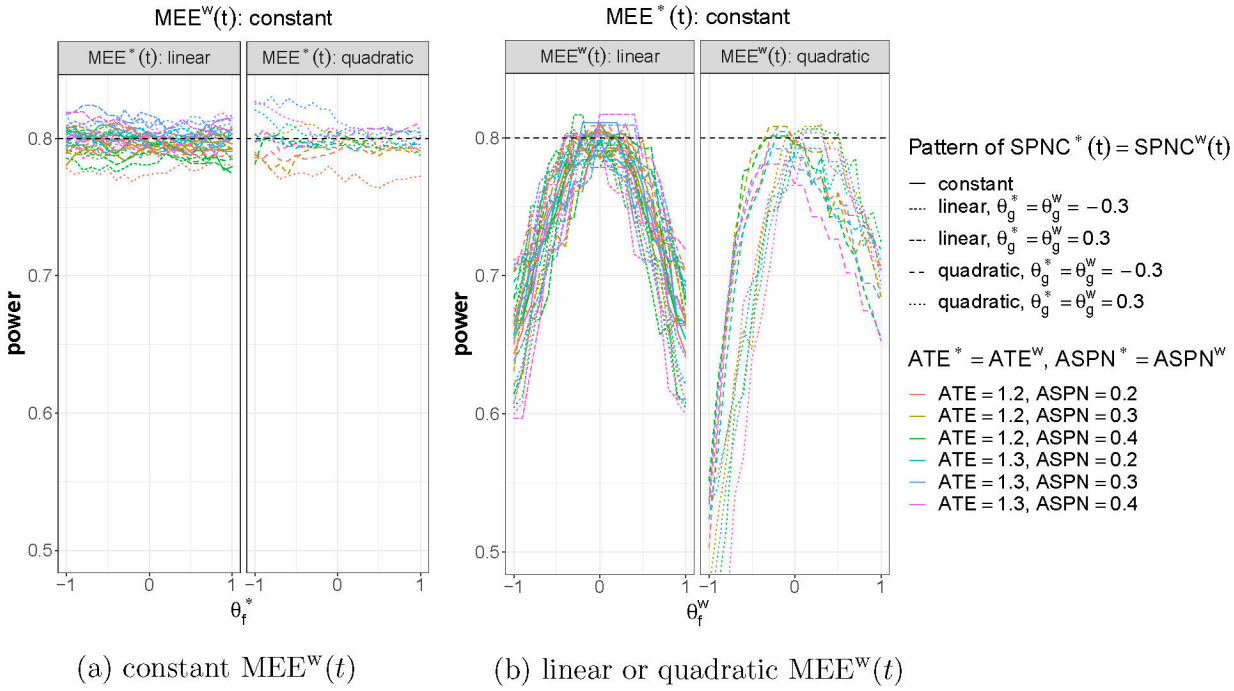


Figure 6:

Power when (WA-a) is violated in that $\text{ATE}^{\text{W}} \neq \text{ATE}^*$. Here, $\text{ATE}^* = 1.4$, $\text{SPNC}^*(t) = \text{SPNC}^{\text{W}}(t)$ are both quadratic, the patterns of $\text{MEE}^*(t)$ and $\text{MEE}^{\text{W}}(t)$ are the same, and $\tau^*(t) = \tau^{\text{W}}(t) \equiv 1$.



(a) constant $\mathrm{MEE}^{\mathrm{w}}(t)$ (b) linear or quadratic $\mathrm{MEE}^{\mathrm{w}}(t)$

Figure 7:

Power when (WA-a) is violated in that the pattern of $\mathrm{MEE}(t)$ is incorrect: one of $\mathrm{MEE}^*(t)$ and $\mathrm{MEE}^{\mathrm{w}}(t)$ is constant and the other is linear or quadratic. Here, $\mathrm{ATE}^* = \mathrm{ATE}^{\mathrm{w}}$, $\mathrm{ASPN}^* = \mathrm{ASPN}^{\mathrm{w}}$, $\mathrm{SPNC}^*(t) = \mathrm{SPNC}^{\mathrm{w}}(t)$, and $\tau^*(t) = \tau^{\mathrm{w}}(t) \equiv 1$.

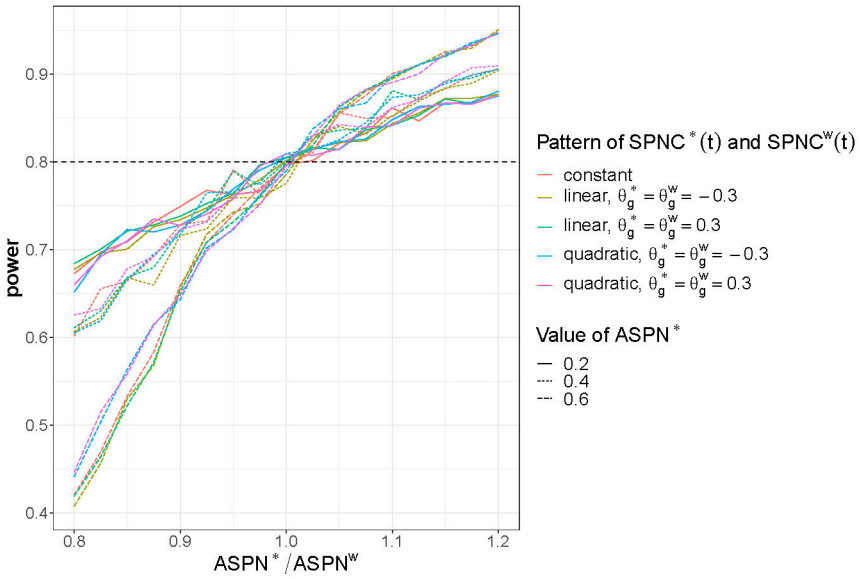
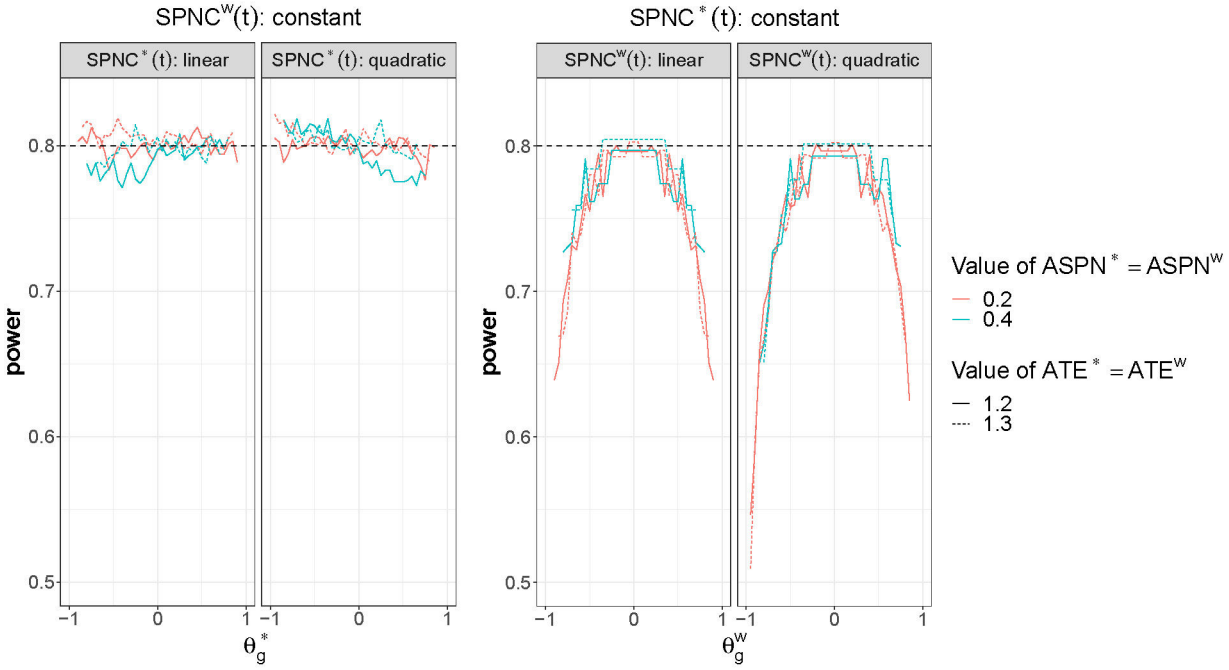


Figure 8:

Power when (WA-b) is violated in that $\text{ASPN}^{\text{W}} \neq \text{ASPN}^{*}$. Here, $\text{MEE}^{*}(t) = \text{MEE}^{\text{W}}(t)$ are both constant with $\text{ATE}^{*} = \text{ATE}^{\text{W}} = 1.2$, $\text{SPNC}^{*}(t)$ and $\text{SPNC}^{\text{W}}(t)$ are of the same pattern, and $\tau^{*}(t) = \tau^{\text{W}}(t) \equiv 1$.



(a) constant $\text{SPNC}^{\text{W}}(t)$ (b) linear or quadratic $\text{SPNC}^{\text{W}}(t)$

Figure 9:

Power when (WA-b) is violated in that the pattern of $\text{SPNC}(t)$ is incorrect: one of $\text{SPNC}^*(t)$ and $\text{SPNC}^{\text{W}}(t)$ is constant and the other is linear or quadratic. Here, $\text{MEE}^*(t) = \text{MEE}^{\text{W}}(t)$ are both constant, $\text{ASPN}^* = \text{ASPN}^{\text{W}}$, and $\tau^*(t) = \tau^{\text{W}}(t) \equiv 1$.

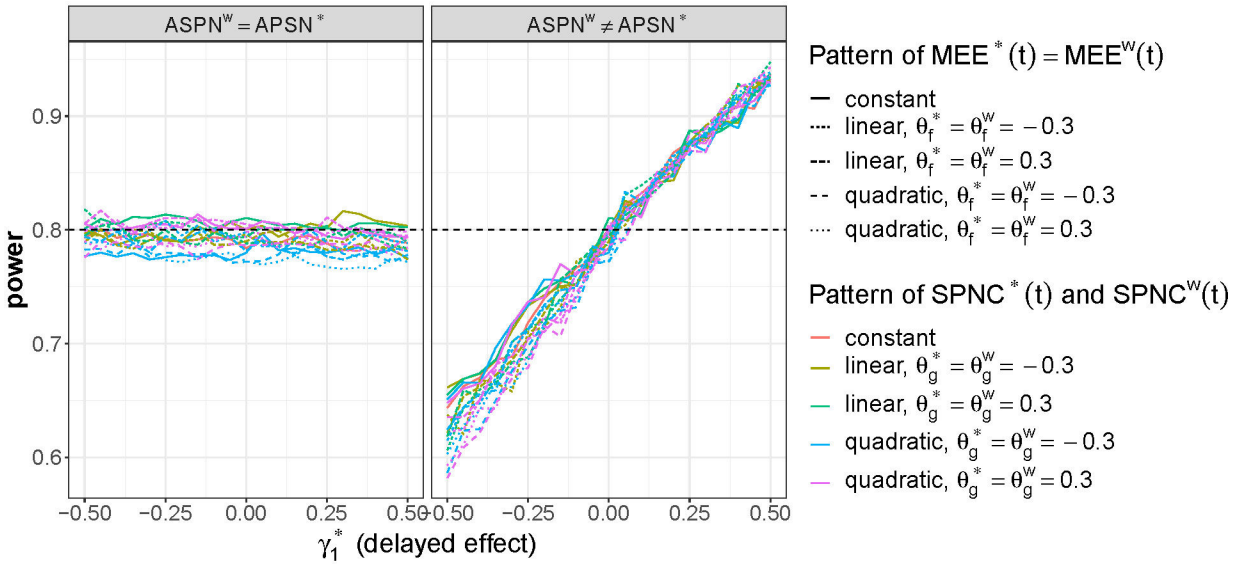


Figure 10:

Power when (WA-b) is violated in that $A_{i,t-1}$ has a delayed effect on $Y_{i,t+1}$ so that $\text{SPNC}^{\text{W}}(t) \neq \text{SPNC}^{*}(t)$. Here, $\text{MEE}^{*}(t) = \text{MEE}^{\text{W}}(t)$ with $\text{ATE}^{*} = \text{ATE}^{\text{W}} = 1.3$, $\text{ASPN}^{*} = 0.3$, and $\tau^{*}(t) = \tau^{\text{W}}(t) \equiv 1$.

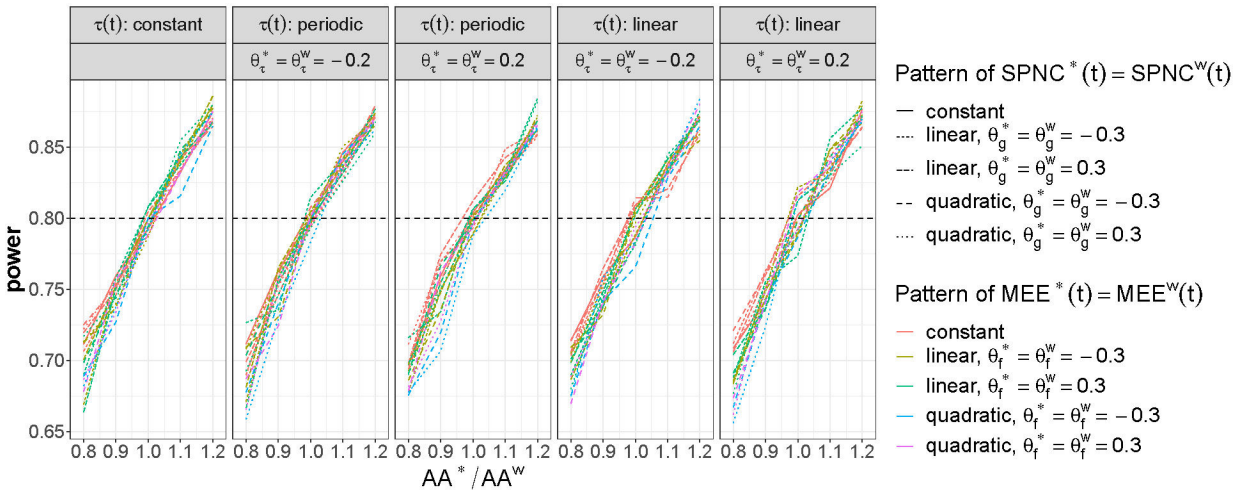


Figure 11:

Power when (WA-c) is violated in that $AA^W \neq AA^*$ but $\tau^*(t)$ and $\tau^W(t)$ are of the same pattern. Here, $MEE^*(t) = MEE^W(t)$ with $ATE^* = ATE^W = 1.3$, $SPNC^*(t) = SPNC^W(t)$ with $ASPN^* = ASPN^W = 0.3$, and $AA^* = 0.6$.

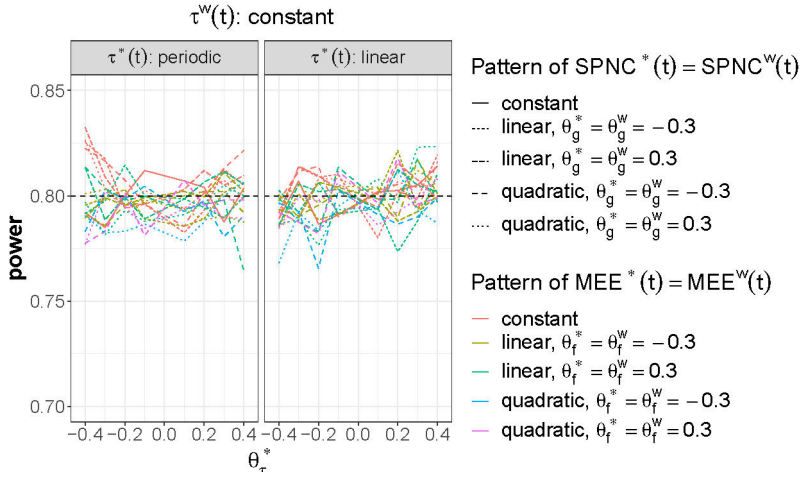


Figure 12:

Power when (WA-c) is violated in that the pattern of $\tau(t)$ is incorrect.

Here, $\text{MEE}^*(t) = \text{MEE}^W(t)$ with $\text{ATE}^* = \text{ATE}^W = 1.3$, $\text{SPNC}^*(t) = \text{SPNC}^W(t)$ with ASPN $\text{ASPN}^* = \text{ASPN}^W = 0.3$, and $\text{AA}^* = \text{AA}^W = 0.6$.

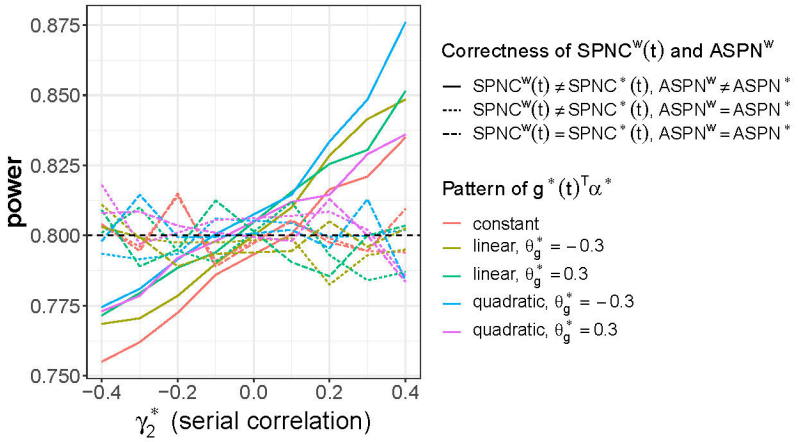


Figure 13:

Power when (WA-d) is violated in that there is serial correlation in the outcome. Here, $\text{MEE}^*(t) = \text{MEE}^W(t) = \log(1.2)$ with $\text{ATE}^* = \text{ATE}^W = 1.2$, $\text{ASPN}^* = 0.2$, and $\tau^*(t) = \tau^W(t) = 1$.

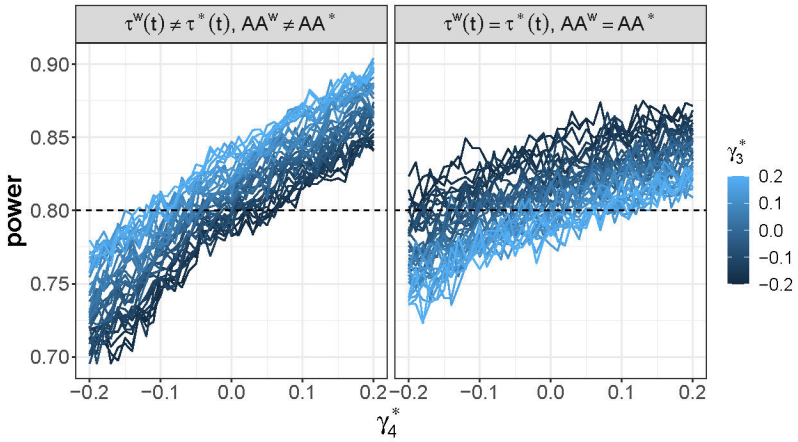


Figure 14:

Power when (WA-e) is violated in that I_t depends on A_{t-1} and Y_t. Here, MEE*(t) = MEE^W(t) = log(1.4) with ATE* = ATE^W = 1.4, and SPNC*(t) = SPNC^W(t) = 0.5 with ASPN* = ASPN^W = 0.5.

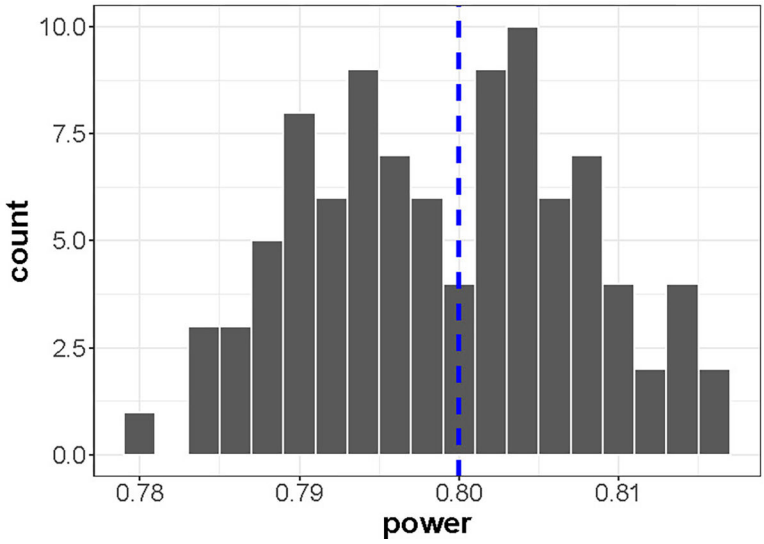


Figure 15:

Power when (WA-a), (WA-b), and (WA-c) are simultaneously violated in that the patterns of MEE(t), SPNC(t) and $\tau(t)$ are all misspecified but their magnitudes are all correct. The histogram shows power under 96 generative models, as detailed in Section 6.8.

Table 1:

Input to the sample size formula.

Input	Interpretation
$1-b$	desired power
η	desired type I error
m	total number of decision points per participant
$p_t(1 \le t \le m)$	randomization probability at a decision point
$\tau(t)\ (1 \le t \le m)$	probability of a participant being availability at a decision point
$\beta_0,\ f(t)(1 \le t \le m)$	marginal excursion effect for the target alternative: $\mathrm{MEE}(t) = f(t)^T\beta_0$
$\alpha_0,\ g(t)\ (1 \le t \le m)$	success probability null curve: $E(Y_{t+1} \mid A_t = 0, I_t = 1) = \exp\{g(t)^T\alpha_0\}$

Table 2:

Performance of the sample size formula when working assumptions are violated.

WA violated	Details about the setting		Power	Setting
(WA-a)	MEE(t) magnitude incorrect	$ATE^w > ATE^*$	↓	1
		$ATE^w < ATE^*$	↑	2
	MEE(t) pattern incorrect	MEE^w (t) constant	→	3
		MEE^w (t) non-constant	↓*	4
(WA-b)	SPNC(t) magnitude incorrect	$ASPN^w > ASPN^*$	↓	5
		$ASPN^w < ASPN^*$	↑	6
	SPNC(t) pattern incorrect	$SPNC^w$ (t) constant	→	7
		$SPNC^w$ (t) non-constant	↓*	8
	delayed effect but is accounted for ($ASPN^w = ASPN^*$)		→	9
	delayed effect not accounted for ($ASPN^w \neq ASPN^*$)	positive delayed effect	↑	10
		negative delayed effect	↓	11
(WA-c)	$\tau(t)$ magnitude incorrect	$AA^w > AA^*$	↓	12
		$AA^w < AA^*$	↑	13
	$\tau(t)$ pattern incorrect	$\tau^w(t)$ constant	→	14
		$\tau^w(t)$ non-constant	→	15
(WA-d) & (WA-b)	serial corr. but is accounted for ($ASPN^w = ASPN^*$)		→	16
	serial corr. not accounted for ($ASPN^w \neq ASPN^*$)	positive serial corr.	↑	17
		negative serial corr.	↓	18
(WA-e)	endogenous availability process		↓*	19

↓ under-powered; ↑ over-powered; → adequately powered (precisely at the desired level); ↓* under-powered for some generative models. In settings 1–8 and 12–15, we considered two ways to violate each of (WA-a), (WA-b), and (WA-c): MEE(t), SPNC(t), or $\tau(t)$ has an incorrect magnitude or has a correct magnitude but an incorrect pattern. In settings 9–11, we considered a delayed effect of past treatments on the outcome, which violates (WA-b) because SPNC*(t) would depend on the delayed effect in a complicated way and we do not expect $SPNC^W(t) = SPNC^*(t)$. In settings 16–18, we considered a serial correlation in the outcome, which violates (WA-b) in addition to (WA-d) because SPNC*(t) would depend on the serial correlation in a complicated way and we do not expect $SPNC^W(t) = SPNC^*(t)$. In setting 19, we allow the availability to depend on past treatments and outcomes.

Table 3:

Practical guidelines for specifying the inputs to the sample size formula. MEE(t): marginal excursion effect at time t, defined in (2). ATE: treatment effect averaged over time on the multiplicative scale, defined in (11). SPNC(t): success probability under the null at time t, defined in (10). ASPN: success probability under the null averaged over time, defined in (12). $\tau(t) := E(I_t)$: expected availability at time t. AA: expected availability averaged over time, defined in (13).

For $1-b,\ \eta,\ m,\ p_t$
• Specify according to the MRT design.
For $\text{MEE}(t) = f(t)^T\beta_0$
• Use a constant pattern unless strong prior knowledge about the specific non-linear form.
• Calculate ATE ($= \beta_0$ if constant $\text{MEE}(t)$).
• Be conservative in ATE; e.g., use the lower end of the conjectured range.
For $\text{SPNC}(t) = \exp\{g(t)^T\alpha_0\}$
• Use a constant pattern unless strong prior knowledge about the specific non-linear form.
• Calculate ASPN ($= \exp(\alpha_0)$ if constant $\text{SPNC}(t)$).
• Be conservative in ASPN; e.g., use the lower end of the conjectured range.
• Use an even lower ASPN if a negative delayed effect or negative serial correlation is expected.
For $\tau(t)$ (for MRT with availability considerations)
• Use a constant pattern unless strong prior knowledge about the specific non-linear form.
• Calculate AA.
• Be conservative in AA; e.g., use the lower end of the conjectured range.
• If availability may depend on prior treatments or outcomes, slightly increase the sample size.
• Try out multiple sets of inputs (formed based on domain knowledge and prior data) to the sample size formula to understand the sensitivity of the output sample size to the inputs. Be conservative and use the larger sample size among the calculated results if budget permits.