The Annals of Applied Statistics
2020, Vol. 14, No. 2, 661–684
https://doi.org/10.1214/19-AOAS1293

THE STRATIFIED MICRO-RANDOMIZED TRIAL DESIGN: SAMPLE SIZE CONSIDERATIONS FOR TESTING NESTED CAUSAL EFFECTS OF TIME-VARYING TREATMENTS

BY WALTER DEMPSEY[1,*], PENG LIAO[1,**], SANTOSH KUMAR[2] AND SUSAN A. MURPHY[3]

[1]*Department of Biostatistics, University of Michigan,* [*]*wdem@umich.edu;* [**]*pengliao@umich.edu*
[2]*Department of Computer Science, University of Memphis, santosh.kumar@memphis.edu*
[3]*Department of Statistics, Harvard University, samurphy@fas.harvard.edu*

Technological advancements in the field of mobile devices and wearable sensors have helped overcome obstacles in the delivery of care, making it possible to deliver behavioral treatments anytime and anywhere. Here, we discuss our work on the design of a mobile health smoking cessation intervention study with the goal of assessing whether reminders, delivered at times of stress, result in a reduction/prevention of stress in the near-term and whether this effect changes with time in study. Multiple statistical challenges arose in this effort, leading to the development of the *stratified micro-randomized trial* design. In these designs each individual is randomized to treatment repeatedly at times determined by predictions of risk. These *risk times* may be impacted by prior treatment. We describe the statistical challenges and detail how they can be met.

1. Introduction. The rise of wearable technologies has generated increased scientific interest in the use and development of mobile interventions. Such mobile technology holds promise in providing accessible support to individuals in need. Mobile interventions to maintain adherence to HIV medication and smoking cessation, for example, have shown sufficient effectiveness to be recommended for inclusion in health services (Free et al. (2013)). Scientists are increasingly interested in understanding whether it is useful to trigger delivery of treatments at risk times, such as when the individual is stressed (Hovsepian et al. (2015)), anxious or disengaging. Because treatments delivered by phone or wearable can be perceived as intrusive and burdensome, a further goal is to assess if treatment effects change through time.

This paper focuses on applied experimental trial design in the new area of mobile health. In particular, we discuss and illustrate the *stratified micro-randomized trial* (sMRT) design. This is motivated by our work on the design of multiple sMRTs. This paper's main focus is *Sense2Stop*, a mobile health smoking cessation study that is currently underway. In this study participants are trained in stress reduction exercises prior to their smoking quit date. Apps that can be used to guide the participant through the exercises are installed on study-provided phone. These apps can be accessed at any time by a participant. However, a common problem is that, at the very times at which practicing these exercises might be most useful, participants do not do so. The scientific team is most interested in understanding whether reminders to practice stress-reduction exercises will be useful in reducing/preventing future stress if the reminders are delivered at times the participant is classified as stressed. Thus, some reminders are to occur at these *stress times* and the remaining at times the participant is not classified as stressed. A primary goal of this study is to assess whether the reminders,

Received June 2018; revised May 2019.
Key words and phrases. Sequential randomization, nested causal effects, stratified micro-randomized trials, mobile health, weighted-centered least-squares method.

delivered at stress times, result in a reduction/prevention of stress over the subsequent hour and whether this effect changes with time.

The design of this sMRT as well as others present a number of challenges:

1. Expressing the primary scientific hypothesis in terms of a causal effect is nontrivial.
2. The primary hypothesis test procedure (e.g., test statistic and rejection region) should balance small sample bias and power when the alternative hypothesis is true.
3. We aim to construct a primary hypothesis test procedure (e.g., test statistic and rejection region) that avoids introducing causal bias.
4. Sometimes the primary hypothesis concerns the distribution of a response that should accrue over a time period in which there is no subsequent treatment, but, in the study, subsequent treatment can occur during this time period.
5. A generative model is needed to calculate the required number of participants:
   - Only small, observational data from participants wearing the same sensor suite are usually available.
   - The sample size calculator should be robust to plausible deviations from the baseline generative model.

In the following we first discuss the smoking cessation study in greater detail. Next, we introduce the stratified micro-randomized trial (sMRT). We then define the causal treatment effect addressing challenge 1. Next, we construct a test statistic and associated theory that accommodate challenges 2–4. Subsequently, we develop a simulation-based method for determining the sample size that accommodates challenge 5.

2. Sense2Stop smoking cessation study. To focus on the experimental design and associated statistical challenges, we consider a simplified version of the smoking cessation study, *Sense2Stop*, in which we are involved through the Mobile Data to Knowledge Center (https://md2k.org/).[1] Sense2Stop is a 10 day mobile health intervention study beginning on each participant's smoking quit day. Participants wear both an AutoSense chest band (Ertin et al. (2011)) as well as bands on each wrist for 10 hours per day. An online pattern-mining algorithm uses the resulting sensor data to construct a binary time-varying stress classification (see Section 7 for an overview of how this algorithm uses episodes of time to construct the stress classifications) at each minute of sensor wearing throughout the entire day.

Each participant's smartphone contains a number of guided stress-reduction exercises that can be accessed 24/7. Participants are trained in the use of these exercises prior to their quit date. The treatment is a smartphone notification to remind the participant to access the app and to practice the stress-reduction exercises. Theoretically, a treatment can be delivered at any minute during the 10 hour day. Practically, treatment delivery is constrained by considerations of attendant burden and to times at which online stress classification is possible.

The trial design should enable us to address the scientific questions:

Is there an effect of the reminder treatment on near-term, proximal stress if the individual is currently experiencing stress? Does the effect of the reminder treatments vary with time in study?

3. Stratified micro-randomized trial. In general, the *stratified microrandomized trial* (sMRT) consists of a sequence of within-person decision times $t = 1, \ldots, T$, for example, occasions at which treatment may be randomized. In Sense2Stop there is a decision time

[1] *Simplified version* refers to omission of study details that obscure the core health and statistical science considerations (e.g., self-report protocol, methods used to reduce data loss due to technical failures and initial confusion in language).

each minute; that is, $T = 600 \times 10$ decision times. sMRTs are a generalization of the micro-randomized trial (Liao et al. (2016), Dempsey et al. (2015), Klasnja et al. (2015), Bidargaddi et al. (2018)) to accommodate stratification. The decision times are divided into strata, and the randomization occurs separately by strata. This ensures sufficient treatment and no treatment occurrences within each strata.

In Sense2Stop the stratification is motivated by our goal of collecting data to address the questions posed in the prior section. There are two strata, minutes at which a participant is classified as stressed and minutes at which the participant is not classified as stressed. Prior data indicated that participants are likely to experience many fewer minutes of stress than nonstress minutes per day, thus motivating the stratification.

In contrast to micro-randomized trials, in an sMRT, the stratification requires online monitoring of a time-varying stratification variable (e.g., minute-by-minute stress classification in Sense2Stop) as well as the development of randomization probabilities that, for each participant, depend on that participant's prior data. As a result sample size calculations are more complex than in the micro-randomized trial further complicating the fifth challenge listed in Section 1.

To describe the sMRT, and in particular the Sense2Stop sMRT, we use the following definitions:

Availability. At decision time t the mobile app assesses if the participant is unavailable for randomization. That is, at some time points it is inappropriate to provide treatment due to ethical, feasibility or burden considerations. In Sense2Stop if a participant receives a treatment reminder, then for the next 60 minutes the participant is unavailable for further treatment. This was done to limit burden and intrusiveness of smartphone notifications.

Feasibility constraints often are due to current sensing technology along with restrictions imposed by the goal of real time detection of the stratification variable. In Sense2Stop, for example, the classification algorithm only makes a real time classification of stress at minutes at which sufficient evidence of recent stress has accumulated. In particular, the Sense2Stop classification algorithm produces a smoothed probability of physiological stress across the minutes with an episodic pattern—the minute-by-minute probability increases then decreases then increases and so on. An episode is defined by the beginning of a positive-trend interval and peaks at the end of a positive-trend interval followed by the start of a negative-trend interval. To ensure the required sensitivity and specificity, the algorithm only makes a classification in the minute after the peak of an episode (see Figure 1). Only at these peak minutes is a participant considered available (provided no treatment has been delivered in the past 60 minutes). At all other times the participant is considered unavailable. For greater detail see the discussion in Section 7. The indicator $I_t = 1$ means that the participant is available at decision time t and $I_t = 0$ otherwise.

Stratification variable. The stratification variable is denoted X_t. In Sense2Stop there are two strata; $X_t = 1$ indicates t is within an episode which, at the peak of this episode, the participant was classified as stressed and $X_t = 0$ otherwise. As depicted in Figure 1, X_t is only observed in real time if t is the minute following the peak. This is also the one minute during the episode at which the participant is available as discussed above. In general, X_t may be categorical.

Treatment. At available decision times treatment, A_t, is randomized. In Sense2Stop A_t is binary with $A_t = 1$, if at minute t, the participant is randomized to receive a reminder to practice stress-reduction exercises and $A_t = 0$ otherwise.

Proximal response. Usually treatments are designed to have a proximal, near-term effect on a response variable. This proximal response, denoted here by $Y_{t,\Delta}$, is assumed to be a known function of the participant's data within a subsequent window of length Δ. In Sense2Stop the proximal outcome is the fraction of time classified as stressed over the subsequent $\Delta = 60$

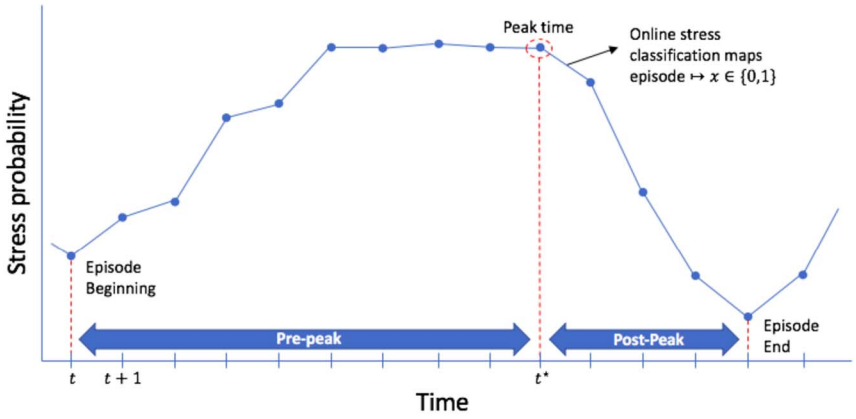


FIG. 1. *Illustrative example of the episodic pattern of smoothed Sense2Stop stress probabilities and its associated online classification algorithm. In the minute following $t^\star$, a stress classification is made. Subsequently, all minutes from the episode beginning to episode end are given the same classification. A participant can only be available in the minute following $t^\star$.*

minutes. Note, as discussed above, the real time stress classification is made at the peak of an episode (see Figure 1). Once a classification is made, then prior minutes in the same episode receive the same classification. This means that X_t is defined at all minutes t and thus can be used to form a proximal outcome. In particular,

$$Y_{t,\Delta} = \Delta^{-1} \sum_{s=1}^{\Delta} \mathbf{1}_{X_{t+s}=1}.$$

This choice of proximal response led to challenge 4.

Longitudinal data. The ordering of a participant's longitudinal data for use in the primary analysis is

$$\big(\{X_1, U_1, I_1\}, A_1, \{Y_2, X_2, U_2, I_2\}, A_2, \ldots, A_{T-1}, \{X_T, U_T, I_T\}\big),$$

where $U_t \in \{0, 1, 2\}$ indicates the episode phase at decision time t (i.e., "pre-peak," "peak" and "post-peak"). Let $H_t = (\{\{X_s, U_s, I_s\}, A_s\}_{s=1}^{t-1}, \{X_t, U_t, I_t\})$ denote the observed data at time t prior to randomization. In general, as in Sense2Stop, X_t, U_t and I_t may be impacted by prior treatment.

Randomization formula. At an available decision time t, the randomization probability, $\text{pr}(A_t = 1|H_t)$, is a known function of H_t, denoted by $p_t(1|H_t)$; else if $I_t = 0$ then $A_t = 0$ and $p_t(1|H_t)$ is set to 0. Note that $p_t(\cdot|H_t)$ need only be defined and is only used in the experiment, if t is an available decision time. Section F of the supplementary material (Dempsey et al. (2020)) provides a simplified version of the formula, used in Sense2Stop, for $p_t(a|h_t), t = 1, \ldots, T$ for any value of observed history.

The randomization probability is set to ensure an average number of treatments within a given time duration (e.g., within a day or a week). It is our experience that these constraints are almost always due to concerns about the intrusiveness and attendant burden. In designing Sense2Stop the team felt that an average of 1.5 treatment reminders per day within each strata would be well tolerated. The need for stratification and the average constraint of 1.5 treatment reminders per strata per day resulted in a randomization probability that depended on the entire observed history. This fact contributes to challenges 3 and 5.

REMARK 3.1 (Designing an sMRT). Appendices A, B and C are included to aid scientists interested in designing a sMRT.

4. Proximal effect of treatment. The primary question of interest is whether the treatment has a proximal effect, that is, whether there is an effect of treatment at decision time t on the mean proximal response $Y_{t,\Delta}$. Below we use potential outcomes (Robins (1986), Rubin (1978)) to make this question precise and, in particular, operationalize the questions relevant to Sense2Stop posed at the end of Section 2. Note, we are only interested in treatment effects conditional on availability ($I_t = 1$). We consider two types of effects: an effect that is defined conditionally on the value of the stratification variable X_t and $I_t = 1$ or an effect that is conditional only on $I_t = 1$, so marginal with respect to the distribution of X_t. For expositional simplicity we focus on the test for the conditional treatment effect in the remainder of this paper. Section I of the supplementary materials provides a parallel discussion in the case of the marginal treatment effect.

4.1. *Proximal treatment effect, potential outcomes and reference distribution.* As stated above, we use potential outcomes (Robins (1986), Rubin (1978)) to define the conditional proximal effect. The overbar is used to denote a sequence through a specified treatment occasion, $\bar{a}_t = (a_1, \ldots, a_t)$, for instance, denotes the sequence of realized actions up to and including decision time t. The potential observations at decision time t are $\{X_t(\bar{a}_{t-1}), U_t(\bar{a}_{t-1}), I_t(\bar{a}_{t-1})\}_{\bar{a}_{t-1}\in\{0,1\}^{t-1}}$. For example, at time 2 the potential observations are $\{X_2(a_1), U_2(a_1), I_2(a_1)\}_{a_1\in\{0,1\}}$. In the case of Sense2Stop, availability is defined as

$$I_t(\bar{a}_{t-1}) = \begin{cases} 1 & \text{if } \sum_{s=1}^{\Delta} a_{t-s} = 0 \text{ and } U_t(\bar{a}_{t-1}) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The potential outcomes for the proximal response at time t are $\{Y_{t,\Delta}(\bar{a}_{t+\Delta-1})\}_{\bar{a}_{t+\Delta-1}\in\{0,1\}^{t+\Delta-1}}$. Each individual has $2^{t+\Delta-1}$ potential outcomes at time t.

At the individual level the effect of providing treatment versus not providing treatment at time t is a difference in potential outcomes for the proximal response and is given by

$$Y_{t,\Delta}(\bar{a}_{t-1}, 1, a_{t+1}, \ldots, a_{t+\Delta-1}) - Y_{t,\Delta}(\bar{a}_{t-1}, 0, a_{t+1}, \ldots, a_{t+\Delta-1}). \tag{1}$$

In general, there are $2^{t+\Delta-2}$ treatment differences for each individual, each corresponding to a treatment pattern for $(\bar{a}_{t-1}, a_{t+1}, \ldots, a_{t+\Delta-1})$. However, participants' availability constrains the number of possible treatment patterns. In particular, our hypotheses only concern differences of potential outcomes corresponding to treatment at available times. In the Sense2Stop study, for instance, we are interested in treatment differences between potential outcomes for which if $a_t = 1$, then $(a_{t+1}, \ldots, a_{t+\Delta})$ is equal to $\bar{0}$, since following treatment, the participant is unavailable for further treatment for the next $\Delta = 60$ minutes.

Recall that the "fundamental problem of causal inference" (Imbens and Rubin (2015), Pearl (2009)) is that we cannot observe any one of these individual differences. Thus, we consider averages of potential outcomes in defining treatment effects. In addition, to define the treatment effect, we specify a reference distribution,[2] that is, the distribution of treatments prior to time t, $\bar{a}_{t-1}$. Moreover, if $\Delta > 1$, then we must also define a second reference distribution over treatments after time t, $(a_{t+1}, \ldots, a_{t+\Delta-1})$. Overall, the treatment effect at time t will be an average of the differences in (1) both over the distribution across individual's potential outcomes as well as over the reference distributions for the treatments and respecting the constraints imposed by availability. To define the proximal treatment effect, we must select these reference distributions.

[2]Here reference distribution is unrelated to the notion of reference sets in randomization inference; see Rosenberger, Uschner and Wang (2019).

The question is, "Which reference distributions should be used?" The choice of which distribution to use for $(a_{t+1}, \ldots, a_{t+\Delta-1})$ might differ by the type of inference desired. For example, in Sense2Stop we further operationalize the questions posed at the end of Section 2 by setting the treatments $a_{t+1}, \ldots, a_{t+\Delta-1}$ to 0. In this case the treatment effect is:

The effect on the fraction of time stressed in the next hour of (a) providing a notification at time t to practice stress-reduction exercises and no notifications within the next hour versus (b) no notification at time t and no notifications within the next hour.

In this paper we set treatment at the subsequent $\Delta - 1$ times equal to 0 as described above. In order to select the reference distribution for $\bar{a}_{t-1}$, we follow common practice in observational mobile health studies; here, longitudinal methods such as GEEs and random effects models (Liang and Zeger (1986)) might be used to model how a time-varying variable, such as physical activity, varies with current mood. In this case the mean model in these analyses is marginal over the past distribution of mood. A similar strategy in the randomized setting is to use the past treatment randomization probabilities as the reference distribution.

With the reference distribution set to the randomization probabilities for past treatment and set to no treatment for the subsequent $\Delta - 1$ times, the average causal effect at time t can be viewed as an *excursion*. That is, participants get to time t under treatment according to the randomization probabilities, then at time t (if available) the effect is the contrast between two opposing excursions into the future. In one excursion we treat at time t and then do not treat for $\Delta - 1$ further times; in the opposing excursion we do not treat at time t nor do we treat for $\Delta - 1$ subsequent times.

Using the above reference distribution, the conditional, proximal treatment effect at time t, $\beta(t; x)$, is

$$\begin{aligned}&\left(\mathbb{E}\left[\sum_{\bar{a}_{t-1}}\left(\prod_{j=1}^{t-1} p_j(a_j|H_j(\bar{a}_{j-1}))\right)(Y_{t,\Delta}(\bar{a}_{t-1}, 1, \bar{0})\right.\right.\\&\quad\left.\left.- Y_{t,\Delta}(\bar{a}_{t-1}, 0, \bar{0}))I_t(\bar{a}_{t-1})1_{X_t(\bar{a}_{t-1})=x}\right]\right)\\&\Big/\left(\mathbb{E}\left[\sum_{\bar{a}_{t-1}}\left(\prod_{j=1}^{t-1} p_j(a_j|H_j(\bar{a}_{j-1}))\right)I_t(\bar{a}_{t-1})1_{X_t(\bar{a}_{t-1})=x}\right]\right),\end{aligned}$$

where the expectation, $\mathbb{E}$, is over the distribution of the potential outcomes and $\bar{0}$ is a row vector of length $\Delta - 1$.

Beyond scientific considerations a further statistical consideration in selecting a reference distribution is that if the reference distribution is far from the randomization distribution, then treatment effects may be very difficult to estimate; see Section B in the supplementary material for a discussion. For the reminder of this paper, the proximal effects are defined using the randomization distribution for past treatments $(\bar{a}_{t-1})$ and $(a_{t+1}, \ldots, a_{t+\Delta-1})$ are set to 0 (no treatment).

4.2. *Proximal effect of treatment and observable data.* The following three assumptions are used to express the causal treatment effect, $\beta(t; x)$, in terms of the observable data.

ASSUMPTION 4.1. We assume consistency, positivity and sequential ignorability (Robins (1986)):

- Consistency: For each $t \leq T + \Delta$, $\{X_t(\bar{A}_{t-1}), I_t(\bar{A}_{t-1})\} = \{X_t, I_t\}$. That is, the observed values equal the corresponding potential outcomes.

- Positivity: If the joint density $\{H_t = h, A_t = a\}$ is greater than zero, then $\text{pr}(A_t = a_t | H_t = h_t) > 0$.
- Sequential ignorability: For each $t \leq T$, the potential outcomes, $\{X_2(a_1), I_2(a_1), \ldots, X_{T+\Delta}(\bar{a}_{T+\Delta-1})\}_{\bar{a}_{T+\Delta-1} \in \{0,1\}^{T+\Delta-1}}$, are independent of A_t conditional on the history H_t.

Sequential ignorability and, assuming all of the randomization probabilities are bounded away from *zero* and *one*, positivity are guaranteed for an sMRT by design. Consistency is a necessary assumption for linking the potential outcomes as defined here to the data. When an individual's outcomes may be influenced by the treatments provided to other individuals, consistency may not hold. In such instances a group-based conceptualization of potential outcomes is used (Hong and Raudenbush (2006), VanderWeele et al. (2013)). In particular, if the mobile intervention includes treatments that aim to produce social ties between participants, then consistency as stated above will not hold. For simplicity, we do not consider such mobile interventions here.

LEMMA 4.2. *Under Assumption 4.1, the conditional treatment effect satisfies*

$$\begin{aligned}\beta(t;x) = {} & \mathbb{E}_{\mathbf{p}}\left[\mathbb{E}_{\mathbf{p}}\left[\prod_{j=t+1}^{t+\Delta-1} \frac{1_{A_j=0}}{p_j(A_j|H_j)} Y_{t,\Delta} | A_t = 1, H_t\right] \middle| X_t = x, I_t = 1\right] \\ & - \mathbb{E}_{\mathbf{p}}\left[\mathbb{E}_{\mathbf{p}}\left[\prod_{j=t+1}^{t+\Delta-1} \frac{1_{A_j=0}}{p_j(A_j|H_j)} Y_{t,\Delta} | A_t = 0, H_t\right] \middle| X_t = x, I_t = 1\right]\end{aligned} \tag{2}$$

for all $x \in \{0, \ldots, k\}$ where each expectation is with respect to the distribution of the data collected using the randomization probabilities specified in the design of the sMRT (indicated by the subscript $\mathbf{p}$ *on the expectations).*

Note that the above products, for example, $\prod_{j=t+1}^{t+\Delta-1} \frac{1_{A_j=0}}{p_j(A_j|H_j)}$, are set to 1 if $\Delta = 1$. Proof of Lemma 4.2 can be found in the Section G of the supplementary material (Dempsey et al. (2020)). We now focus on designing an sMRT where the primary purpose is testing whether the treatment effect at any time point differs from 0.

5. Test statistic. Our main objective is the development of a sample size formula that will ensure sufficient power to detect alternatives to the null hypothesis of no proximal treatment effect. For the conditional proximal effect the null hypothesis is $H_0 : \beta(t;x) = 0, t = 1, \ldots, T$ and $x \in \{0, \ldots, k\}$.

The proposed sample size formulas are simulation-based and will follow from consideration of the distribution of test statistics under alternatives to the above null hypothesis. The sample size will be denoted by N. Our test statistic will generalize the test statistics developed by Boruvka et al. (2018) to accommodate stratification as well as the fact that the response $Y_{t,\Delta}$ covers a time interval during which subsequent treatment may be delivered (in Boruvka et al. (2018), $\Delta = 1$ throughout). Moreover, sample size calculations are informed by the novel conceptual insight that these estimators can be interpreted as L_2 projections (see Remark 4 in Section 5.1).

In the following we describe L_2 projections and provide the test statistics. First, in the conditional setting, the test statistic is based on an empirical projection of $\{\beta(t;x)\}_{t=1,\ldots,T;x\in\{0,\ldots,k\}}$ on the space spanned by a q_c by 1 vector of features involving t and x, denoted by $f_t(x)$. We denote the projection by $f_t(x)'\beta_c$. The β_c weights in this projection are given by

$$\beta_c^\star = \arg\min_{\beta_c} \mathbb{E}_{\mathbf{p}}\left[\sum_{t=1}^{T} I_t \tilde{p}_t(1|X_t)\big(1 - \tilde{p}_t(1|X_t)\big)\big(\beta(t;X_t) - f_t(X_t)'\beta_c\big)^2\right],$$

where $\{\tilde{p}_t(1|x)\}_{t=1,\ldots,T;x\in\{0,\ldots,k\}}$ are prespecified probabilities used to define the weighting across time and stratification distribution in the projection. The expectation $\mathbb{E}_{\mathbf{p}}$ is taken with respect to the joint distribution of $\{(X_t, I_t)\}_{t=1}^{T}$ generated using the randomization probabilities in the sMRT design. If desired, one can set $\tilde{p}_t(1|x) = 1/2$ for all t, x. See Section 6.1 for further comments on the choice of the $\tilde{p}_t(1|x)$'s and $f_t(x)$.

In some settings there will be sufficient a priori information (e.g., data on individuals from a similar population) that will permit the test statistic to use *control variables*. These variables are used to help reduce the variance of the estimators with the goal that the resulting test statistic is more powerful in detecting particular alternatives to the null hypothesis. See Section 6.1 for further discussion on the choice of control variables. For example, in Sense2Stop a natural control variable would be the fraction of time stressed in the hour prior to time t as this pretime t variable is likely highly correlated with the fraction of time stressed in the hour subsequent to time t, $Y_{t,60}$. Given a q' by 1 vector of "control variables" $g_t(H_t)$, define $g_t(H_t)'\alpha_c^\star$ as an L_2 projection of $\mathbb{E}_{\mathbf{p}}[w_{ct}(H_{t+\Delta-1})Y_{t,\Delta}|I_t = 1, H_{t+\Delta-1}]$; in particular,

$$\alpha_c^\star = \arg\min_{\alpha} \mathbb{E}_{\mathbf{p}}\left[\sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1})\big(Y_{t,\Delta} - g_t(H_t)'\alpha_c\big)^2\right],$$

where $w_{ct}(H_{t+\Delta-1}) = \frac{\tilde{p}_t(A_t|X_t)\prod_{s=1}^{\Delta-1}\mathbf{1}[A_{t+s}=0]}{\prod_{s=0}^{\Delta-1} p_{t+s}(A_{t+s}|H_{t+s})}$. Note, one can choose $g_t(H_t)$ equal to the scalar, 1. This use of control variables to reduce variance in the response is used to address challenge 2 listed in the Section 1.

Recall, the proposed test statistic is based on an estimator of $\beta_c^\star$. Here, we consider an estimator of $\beta_c^\star$ which is the minimizer of the following weighted-centered least-squares criterion, minimized over (α_c, β_c):

$$\mathbb{P}_n\left[\sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1})\big(Y_{t,\Delta} - g_t(H_t)'\alpha_c - \big(A_t - \tilde{p}_t(1|X_t)\big) f_t(X_t)'\beta_c\big)^2\right], \tag{3}$$

where $\mathbb{P}_n[\phi(H_{t+\Delta-1})]$ is defined as the average of a function, $\phi(H_{t+\Delta-1})$, over the sample. The centering refers to the centering of the treatment indicator A_t in the above weighted least-squares criterion. The centering idea is from Boruvka et al. (2018), Liao et al. (2016) (unlike here, Boruvka et al. (2018) aimed to consistently model the treatment effect). Here, we aim to estimate the projection for use in the test statistic; the centering allows us to simultaneously consistently estimate the cofficients in each of the two projections. The consistent estimation of $\beta_c^\star$ addresses challenge 2 listed in Section 1. Centering in the construction of the test statistic preserves the null and avoids introducing causal bias which addresses challenge 3 listed in Section 1. Indeed, preserving the null is difficult because both the stratification variable x and the randomization probabilities may be influenced by prior treatment.

Under finite moment and invertibility assumptions, the minimizers $(\hat{\alpha}_c, \hat{\beta}_c)$, are consistent, asymptotically normal estimators of $(\alpha_c^\star, \beta_c^\star)$. The limiting variance of $\sqrt{N}(\hat{\beta}_c - \beta_c^\star)$ is given by $Q_c^{-1} W_c Q_c^{-1}$ where

$$W_c = \mathbb{E}_{\mathbf{p}}\left[\sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1})\epsilon_{ct}\big(A_t - \tilde{p}_t(1|X_t)\big) f_t(X_t) \times \sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1})\epsilon_{ct}\big(A_t - \tilde{p}_t(1|X_t)\big) f_t(X_t)'\right],$$

$$\epsilon_{ct} = Y_{t,\Delta} - g_t(H_t)'\alpha_c^\star - \big(A_t - \tilde{p}_t(1|X_t)\big) f_t(X_t)'\beta_c^\star \quad \text{and}$$

$$Q_c = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{p}}\big[I_t \tilde{p}_t(1|X_t)\big(1 - \tilde{p}_t(1|X_t)\big)\big) f_t(X_t) f_t(X_t)'\big].$$

See Section G.2 in the supplementary material (Dempsey et al. (2020)) for technical details.

The proposed sample size formula is based on the test statistic

$$T_{cN} = N\hat{\beta}_c' \hat{Q}_c \hat{W}_c^{-1} \hat{Q}_c \hat{\beta}_c, \tag{4}$$

where N is the sample size, $\hat{W}_c$ and $\hat{Q}_c$ are empirical estimators of W_c and Q_c (i.e., replace $\mathbb{E}_{\mathbf{p}}$ with $\mathbb{P}_n$) and $\hat{\epsilon}_{ct} = Y_{t,\Delta} - g_t(H_t)'\hat{\alpha}_c - (A_t - \tilde{p}_t(1|X_t))\, f_t(X_t)'\hat{\beta}_c$. Here, we have implicitly assumed that $\hat{W}_c$ is invertible. The following lemma provides the distribution of T_{cN}:

LEMMA 5.1 (Asymptotic distribution of T_{cN}). *Under finite moment and invertibility assumptions*,

$$N(\hat{\beta}_c - \beta_c^\star)' \hat{Q}_c \hat{W}_c^{-1} \hat{Q}_c (\hat{\beta}_c - \beta_c^\star) \longrightarrow_d \chi^2_{q_c}.$$

The above lemma implies that the distribution of the test statistic T_{cN} is approximately a noncentral chi-squared distribution with noncentrality parameter $\lambda = N\gamma_c$ where

$$\gamma_c = (\beta_c^\star)' Q_c {W_c}^{-1} Q_c \beta_c^\star. \tag{5}$$

However, from a technical perspective, T_{cN} is very similar to the quadratic form test statistics based on weighted regression used in generalized estimating equations (GEEs) method (Liang and Zeger (1986), Diggle et al. (2002)). In this field much work has been done on how to best adjust these test statistics and their distribution when the sample size N might be small (Liao et al. (2016), Mancl and DeRouen (2001)). The adjustments are based on the intuition that the quadratic form is akin to the multivariate T-test statistic used to test whether a vector of means is equal to 0 and thus Hotelling's T-squared distribution is used to approximate the distribution when N may be small.

To develop the sample size formula, we follow the lead of the well-developed GEE literature and use a noncentral Hotelling's T-squared distribution with degrees of freedom $(d_1 = q_c, d_2 = N - (q' + q_c))$ to approximate the distribution of T_{cN}. Recall, q' is the dimension of α_c and q_c is the dimension of β_c. See Section G in the supplementary material (Dempsey et al. (2020)) for a discussion of how for large N, we recover the chi-squared distribution given in Lemma 5.1. Recall that if a random variable X has noncentral Hotelling's T-squared distribution with degrees of freedom (d_1, d_2) and noncentrality parameter λ, then $\frac{d_2}{d_1(d_1+d_2-1)} X$ has noncentral F-distribution with the same degrees of freedom and noncentrality parameter (Hotelling (1931)). Thus, the rejection region for the test $H_0 : \beta(t; x) = 0, t = 1, \ldots, T$ and $x \in \{0, \ldots, k\}$ can be written as

$$\left\{ T_{cN} > \frac{q_c(N - (q' + 1))}{N - (q' + q_c)} F^{-1}_{q_c, N-(q'+q_c);0}(1 - \alpha_0) \right\} \tag{6}$$

with α_0 a specified significance level. For details regarding further small sample size adjustments used when analyzing the data, see Section J in the supplementary material (Dempsey et al. (2020)).

5.1. *Remarks.* Next, we discuss components of the proposed test statistic.

1. *Specification of the weights.* The weight $w_{ct}(H_{t+\Delta-1})$ plays multiple roles:

   - First, the term $\tilde{p}_t(A_t|X_t)/p_t(A_t|H_t)$ is similar to the inverse probability of treatment weighting in causal inference (Robins (1986)) in that it facilitates estimation of a marginal effect, marginal over the history H_t given strata $X_t = x$ and availability $I_t = 1$.

- Second, choice of the numerator of the user-defined weight $\tilde{p}_t(A_t|X_t)$ determines the L_2 projection of the treatment effect if $\beta(t;x)$ is not equal to a linear combination of $f_t(x)$. In these settings the numerator of the weight determines the $\beta_c^\star$ coefficients in the projection. See below for further comments regarding the L_2 projection. These user-defined weights are distinct from randomization probabilities in the sMRT. Note, however, that if the randomization probabilities in the sMRT $p_t(1|H_t)$ only depend on t and X_t then we can set $\tilde{p}_t(A_t|X_t) = p_t(1|H_t)$.
- Third, the remaining terms $\prod_{s=0}^{\Delta-1}[\mathbf{1}[A_{t+1}=0]/p_{t+s}(A_{t+s}|H_{t+s})]$ adjust for the fact that the reference distribution at the subsequent $\Delta-1$ times is different from the sMRT randomization protocol.

2. *No use of a nonindependence working correlation matrix.* As discussed, the estimating equation underlying the test statistic is similar to a generalized estimating equation (GEEs) (Liang and Zeger (1986), Diggle et al. (2002)). While this might motivate inclusion of nonindependence working correlation matrix to further reduce the variance of estimator and thus increase the power of the test (Mancl and Leroux (1996)), the inclusion of a nonindependence working correlation matrix generally introduces causal bias (Boruvka et al. (2018), Liao et al. (2016)). Similar biases occur with the use of nonindependence working correlation matrices in the inverse probability of treatment weighting literature (Vansteelandt (2007), Tchetgen Tchetgen et al. (2012)) or in GEEs where a time-varying response is modeled by time-varying covariates (Pepe and Anderson (1994)).
3. *Use of sandwich estimator of the variance.* The test statistic accounts for the within person correlation in the longitudinal response via use of a sandwich estimator (i.e., $\hat{Q}_c\hat{W}_c^{-1}\hat{Q}_c$) for the covariance matrix. Unfortunately, the power of the test will depend on the within-person correlation in responses, thus the simulation-based sample size formula developed below requires modeling the correlation. Under the null hypothesis that $\beta(t;x)=0, t=1,\ldots,T$ and $x\in\{0,\ldots,k\}$, the test statistic and associated rejection region has the desired asymptotic type I error rate regardless of the underlying true within-person correlation (assuming W_c is invertible).
4. *Use of a L_2-projection to form the test statistic.* Recall that if under the alternative hypothesis $\beta(t;x)$ is not equal to a linear combination of $f_t(x)$, then the L_2-projection of $\beta(t;x)$ depends on the feature vector, the pattern of availability across time and the distribution of the stratification variable across time. Figure 2 provides a visualization. Here, consider different uses of the feature d_t denoting the day in study. The red line is the complex, true treatment effect $\beta(t;x)$. The black line is the projected effect onto feature vector $f_t=(1,d_t)$ when there is a quadratic pattern across time in availability ($\mathbb{E}[I_t|X_t=x]$) and the stratification distribution $P(X_t=x)$ is constant through time. Similar interpretations hold for the blue, dotted blue and dotted black lines as indicated in Figure 2. The four projections are distinct in the top graph in Figure 2, illustrating how the joint distribution of availability and the stratification variable affects the L_2-projection.

   None of the four projections fully reflect the true alternative $\beta(t;x)$, but all four roughly pick up the departure of the true $\beta(t;x)$ from the null hypothesis. While it is tempting to consider higher dimensional and more flexible feature spaces so as to more fully reflect the variety of possible alternatives to the null hypothesis, these come at a cost in the additional degrees of freedom in the F statistic. This may lead to an increase in sample size for a given desired power. This tradeoff is discussed at length in Section H in the supplementary material (Dempsey et al. (2020)); we suggest sizing a study for primary hypothesis tests using the *least* complex alternative possible. In the case of Sense2Stop, we decided to use a projection onto $f_t=(1,d_t,d_t^2)$. The dotted black line in Figure 2 captures most of the variation in $\beta(t;x)$ under plausible time patterns in the distribution of availability and the stratification variable. The test statistic targets this low dimensional alternative so as to address challenge 2 listed in Section 1.

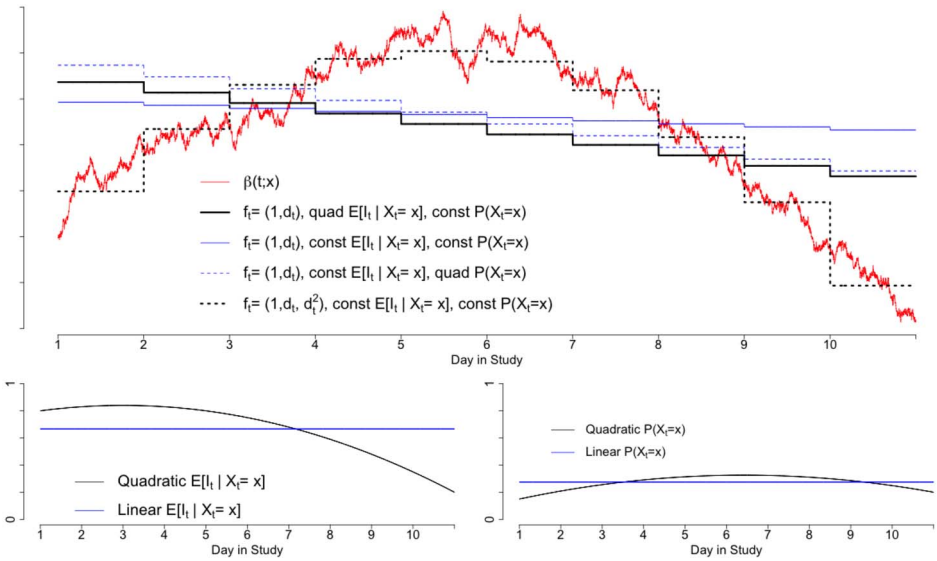


FIG. 2. *Illustration of the L_2-projection of $\beta(t;x)$ onto feature vector f_t. The reference distribution $\tilde{p}_t(1|x)$ is constant in (t,x). The feature vector is nonparametric in binary x and set within each strata to $f_t = (1, d_t)$ or $f_t = (1, d_t, d_t^2)$ where d_t is equal to the number days in study; expected availability given $X_t = x$ is constant in t time or is quadratic in t with the same average. The distribution of X_t is constant in t or is quadratic in t with the same average.*

6. Sample size formulae. To plan the sMRT, we need to determine the sample size, N, needed to detect a specific alternative with a given power $(1 - \beta_0)$ at a given significance level (α_0). The sample size is the smallest value N such that

$$1 - F_{q_c, N-(q'+q_c); N\gamma_c}\big(F^{-1}_{q_c, N-(q'+q_c);0}(1-\alpha_0)\big) \geq 1 - \beta_0. \tag{7}$$

$F_{d_1,d_2;\lambda}$ and $F^{-1}_{d_1,d_2;\lambda}$ denote the cumulative and inverse distribution functions, respectively, for the noncentral F-distribution with degrees of freedom (d_1, d_2) and noncentrality parameter λ. Calculation of the sample size N is nontrivial due to the unknown form of the noncentrality parameter, $N\gamma_c$ (where γ_c is defined in (5)). This is in contrast to micro-randomized trials where, under nonstochastic randomization probabilities and certain working assumptions, Liao et al. (2016) were able to find an analytic form for the noncentrality parameter $N\gamma_c$.

We outline a simulation-based sample size calculation, starting with general overview and comments in Section 6.1 and employ this calculator to design the smoking cessation study in Section 7.

6.1. *Simulation-based sample size calculation.* As discussed above, explicit calculation of the sample size N is nontrivial due to the unknown form of the noncentrality parameter. Here, we propose a three-step simulation-based sample size calculator.

In the first step equation (5) and information elicited from the scientist is used to calculate, via Monte-Carlo integration, γ_c in the noncentrality parameter. The resulting value, $\hat{\gamma}_c$, is plugged in to equation (7) to solve for an *initial* sample size $\hat{N}_0$. In the second step we use a binary search algorithm to search over a neighborhood of $\hat{N}_0$; in our simulations we found the binary search quickly resulted in a solution. For each sample size N required by the binary search algorithm, K samples each of N simulated participants are run. Within each simulation the rejection region for the test is given by equation (6) at the specified significance level. The average number of rejected null hypotheses across the K simulations is the estimated power

for the sample size N. The sample size is the minimal N with estimated power above the prespecified threshold $1 - \beta_0$.

In the last, third, step we conduct a variety of simulations to assess the robustness of the sample size calculator to any assumptions and to make adjustments to ensure robustness. See our use of these simulations to test robustness for Sense2Stop in Section 7.

The sample size calculator uses the following information for $t = 1, \ldots, T; x \in \{0, \ldots, k\}$:

1. Desired type 1 and type 2 error rates,
2. Targeted alternative $\beta(t; x)$,
3. Selected probabilities $\{\tilde{p}_t(1|x)\}$,
4. Selected "control variables" $g_t(H_t)$,
5. The randomization formula used to determine $p_t(1|h)$ given a history h, and
6. A generative model for $\{H_t\}_{t=1,\ldots,T}$.

We provide general comments concerning the choice of the above items and then build the sample size calculator for the Sense2Stop study of Section 7.

First, we elicit information from the scientist to construct a specific alternative form for $\beta(t; x)$. A simple approach is to consider linear alternatives, $\{\beta(t; x) = f_t(x)'\beta_c^\star\}_{t=1,\ldots,T;x\in\{0,\ldots,k\}}$, so that the L_2 projection and the alternative coincide. Stratification variables are often categorical (X is categorical); as a result, we model the alternative separately for each value of $X = x; x \in \{0, \ldots, k\}$. Furthermore, if we suspect that the effect will be generally decreasing (with study time) due to habituation, then we might consider a vector feature, f_t that represents a linear in time t trend. Or we might believe that the effect of the treatments might be low at the beginning of the study and then increase as participants learn how to use the treatment and then decrease due to habituation; here, we might consider a vector feature f_t that results in a quadratic trend. Both quadratic and linear trends are presented in Figure 2.

The less complex the projection (smaller q_c) of the alternative $\beta(t; x)$, the smaller the required sample size N becomes. On the other hand, the use of a simple projection for the alternative may not reflect the true alternative $\beta(t; x)$ very well (see Section H in the supplementary material (Dempsey et al. (2020)) for a discussion of this tradeoff). This led to the suggestion in Section 5.1 for sizing a study for primary hypothesis tests using the *least* complex alternative possible.

To select the probabilities $\{\tilde{p}_t(1|x)\}_{t=1,\ldots,T;x\in\{0,\ldots,k\}}$, recall that these probabilities define the weighting across time and across the stratification distribution of the alternative when operationalized as an L_2 projection. To see this, suppose we decide to target a constant-across-time alternative and select $f_t(X_t) = (\mathbf{1}_{X_t=1}, \mathbf{1}_{X_t=2}, \ldots, \mathbf{1}_{X_t=k})'$. If we set the reference probabilities to be constant in t and x then $\beta_c^\star = (\beta_{c,1}^\star, \beta_{c,2}^\star, \ldots, \beta_{c,k})$ where

$$\beta_{c,x}^\star = \left[\sum_{t=1}^{T} \mathbb{E}[I_t \mathbf{1}_{X_t=x}]\right]^{-1} \left[\sum_{t=1}^{T} \mathbb{E}[I_t \mathbf{1}_{X_t=x}]\beta(t; x)\right].$$

In this case $\beta_{c,x}$ is an average treatment effect across time weighted by the fraction of time the participant is available and in stratification level x. In our work we usually set $\tilde{p}_t(1|x)$ to be constant in (t, x) so as to more easily discuss the targeted alternative with collaborators.

Next, a decision should be made about which control variables $g_t(H_t)$ should be included in the construction of the test statistic. One might want to include in the q' by 1 vector $g_t(H_t)$ many variables so as to maximally reduce variance and thus increase the size of the noncentrality parameter in (5); indeed, for fixed q', the larger the noncentrality parameter, the smaller the sample size N. However, from equation (7) we see that fixing all other quantities, the sample size N increases with increasing q'. So intuitively, there is a tradeoff between

increasing the size of the noncentrality parameter by including more variables in $g_t(H_t)$ with the resulting reduction in degrees of freedom in the denominator of the F test caused by increasing q', the number of variables in $g_t(H_t)$. See Section H in the supplementary material (Dempsey et al. (2020)) for further discussion. This tradeoff is directly related to balancing between small sample bias and power, challenge 2 from Section 1. Below, for Sense2Stop, we calculate the sample size with the vector of control variables $g_t(H_t)$ set equal to $f_t(X_t)$; this maintains a hierarchical regression yet keeps q' as small as possible. Incidentally, this simplifies the development of the generative model as additional time-varying variables are not included.

Generally, the randomization formula has been determined by considerations of treatment burden, availability and whether it is critical for the scientific question that the randomization depend on a time-varying stratification variable such as a prediction of risk. Treatment burden considerations might impose a constraint, such as on average around n treatments per strata should occur over a specified time period (e.g., an average of n treatments per day); also, the randomization formula might be developed so as to limit the variance in the number of treatments in the specified time period. In the Sense2Stop study the randomization probability $p_t(1|H_t)$ is set to limit treatment to an average of 1.5 treatments per day when classified as stressed and when not classified as stressed.

The sample size formula requires the specification of a generative model for the history H_t which achieves the specified alternative treatment effect. However, existing datasets that include the use of the required sensor suites and thus can be used to guide the form of the generative model are often small and do not include treatment. In Sense2Stop, for example, we require a generative model for the multivariate distribution of $\{X_t, U_t, I_t, A_t\}_{t=1}^T$ of which only the distribution of A_t given $(H_t, I_t = 1)$ is known (e.g., $p_t(1|H_t)$). We have access to a small, observational, no-treatment data set that included the required sensor suites and thus can be used to guide the form of the generative model. Because the dataset is small, in Section 7 we construct a low-dimensional Markovian generative model. Here, and in general, the prior data does not include treatments. Thus, we use the prior data to develop a generative model under no treatment.

The relatively simple generative model allows us to use only a few summary statistics from this small noisy dataset. This, of course, may lead to bias (i.e., the simple generative model may not adequately reflect the true data generating mechanism). This bias would be problematic if the bias results in sample sizes for which the power to detect the desired effect is below the specified power. Thus, we also use the small dataset to guide our assessment of robustness of the sample size calculator. In Section 7.4.3 a complex generative model is proposed by exploratory data analysis.

We follow the three steps outlined at the beginning of this subsection to provide a sample size N. Our calculator also provides standardized effect sizes. Table 6 in Section K of the supplementary material (Dempsey et al. (2020)) provides standardized treatment effect sizes, defined as, $d(t; x) = \beta(t; x)/\bar{\sigma}_x$. The average conditional variance, $\bar{\sigma}_x^2 = (1/T) \times \sum_{t=1}^T \mathbb{E}[\text{Var}(Y_{t,\Delta}|I_t = 1, A_t, H_t)|I_t = 1, X_t = x]$, is calculated using the alternative effect $\beta(t; x)$ and the generative model.

7. Sense2Stop. In the following we take the general three-step procedure and walk through how to adapt it to the specifics of the Sense2Stop study and form the sample size calculator. Recall, the last step involves a variety of simulations to assess robustness to the assumption underlying the generative model; this step is provided in Section 7.4.

As noted previously, Sense2Stop is a 10 day study; the first day is the "quit day," the day the participant quits smoking. Recall that participants wear the AutoSense sensor suite (Ertin et al. (2011)) which provides a variety of physiological data streams. During the conduct of

the study, the stratification variable, X_t, is constructed online. $X_t = 1$ if at minute t there is "sufficient evidence" that the participant is in a stress episode; otherwise $X_t = 0$. That is, until there is sufficient evidence whether that the participant is in a stress episode, X_t remains unknown. Further information on these episodes follows. First, every minute, a support vector machine (SVM) algorithm is applied to a number of ECG and respiration features constructed from the prior one minute stream of sensor data. The output of the SVM is then transformed to obtain a stress "likelihood" in (0, 1); see Hovsepian et al. (2015) for details. This output (in (0, 1)) across the minute intervals is further smoothed to obtain a smoother stress likelihood time series. Next, a Moving Average Convergence Divergence approach is used to identify minutes at which the trend in the stress likelihood is going up and when it is going down; see Sarker et al. (2016) for details. The beginning of an episode is marked by the start of a positive-trend interval in the stress "likelihood." Recall, the *peak* of an episode is the end of a positive-trend interval followed by the start of a negative-trend interval. If the area under the curve from the beginning of the episode to the minute that the peak of the episode is detected exceeds a threshold, then at this time the individual is classified as stressed for all minutes t in the episode (i.e., $X_t = 1$). At all other times the participant belongs to the not classified as stressed strata (i.e., $X_t = 0$). Figure 1 visualizes this episodic pattern. The threshold is based on prior data from lab experiments and was evaluated on independent test datasets (from both lab and field) in terms of the F1 score (a combination of sensitivity and specificity (Wikipedia (2017))) for use in detecting physiological stress.

Next, we build the simulation-based calculator assuming the primary hypothesis is H_0 : $\beta(t; x) = 0; t = 1, \ldots, T; x \in \{0, 1\}$ and the test statistic is as given in (4). Small sample corrections are used in constructing the test statistic as discussed in Section 5; see Section J in the supplementary material (Dempsey et al. (2020)) for additional details.

7.1. *Simulation-based calculator.* We start by choosing inputs for the sample size formula as outlined in Section 6.1. We set the desired type 1 and type 2 error rates to be 0.05 and 0.20 respectively. We next specify the targeted alternative $\beta(t; x) = f_t(x)'\beta_c^\star$ for $\beta_c^\star \in \mathbb{R}^{q_c}$. The scientific team suspected that if there is an effect of the mindfulness reminders, then this effect might increase as participants begin to practice the mindfulness exercises and then the effect may decrease due to habituation. Thus, we select $f_t(X_t)' = (f_t' \cdot \mathbf{1}_{X_t=0}, f_t' \cdot \mathbf{1}_{X_t=1})$ where $f_t' = (1, \lfloor \frac{t-1}{600} \rfloor, \lfloor \frac{t-1}{600} \rfloor^2)$. This leads to a nonparametric treatment effect model in the stratification variable X_t and a piecewise constant treatment effect model in time given $X_t = x$ that is quadratic as a function of "day in study." In this case the dimension of the L_2 projection is $q_c = 3 \cdot 2 = 6$, $\beta_c^\star = (\beta_{c,0}^\star, \beta_{c,1}^\star) \in \mathbb{R}^6$ and the targeted alternative is $\beta(t; x) = f_t'\beta_{c,x}^\star$ for $x = 0, 1$. Next, to elicit enough information from the scientist to specify $\beta_c^\star$, we ask scientists to specify for each level of X, (1) an initial conditional effect, (2) the day of maximal effect ($t_x^\star$) and (3) the average conditional treatment effect $\bar{\beta}_{c,x} = T^{-1} \sum_{t=1}^{T} \beta(t; x)$. This set of conditions uniquely identifies the subvector $\beta_{c,x}^\star$; therefore, the conditions over each level of X combine to uniquely identify the vector $\beta_c^\star = (\beta_{c,0}^\star, \beta_{c,1}^\star)$ as desired. For Sense2Stop we will target the same alternative for both levels of the stratification variable X_t; thus, $\beta_{c,0}^\star = \beta_{c,1}^\star$. To set this common alternative, we use the following values: the day of maximal effect is day 6 and the initial conditional effect is 0. We consider three possible common values of $\bar{\beta}_{c,0} = \bar{\beta}_{c,1}$ denoted $\bar{\beta}$ in Table 2.

Here, we set the control variables to $g_t(H_t) = f_t(X_t)$. Furthermore, suppose the formula for randomization probability depends only on past values of the time-varying variable X_t, availability I_t and treatments A_t. We use the formula for $p_t(a|h_t)$ provided in Section F of the supplementary material (Dempsey et al. (2020)). One of the inputs to the randomization formula at an available decision point t is the expected number of episodes during the remaining part of the day that will be classified as stressed ($X = 1$) and the expected number of

episodes during the remaining day that will not be classified as stressed ($X = 0$). The generative model developed below is used to provide this input. See Section F in the supplementary material (Dempsey et al. (2020)) for further details and the specification of other inputs to this randomization formula.

7.1.1. *Generative model*. We briefly overview the procedure for constructing the generative model. We then move on to the specifics, highlighting the rationale behind each decision. First, the stratification variable process X_t is a state-space stochastic process. A natural candidate for such processes with small, finite state spaces are Markov chains. These are computationally tractable and easy to discuss with the scientific team. Second, availability is tied to the episodic nature of the stress classifier as decribed at the beginning of Section 7 with pre-peak, peak and post-peak phases to each episode. To handle this, we specify a "structured Markov chain" (X_t, U_t) where t is at the minute level, X_t denotes the episode type ("Stress," "Nonstress") and U_t denotes the current episode phase ("pre-peak," "peak" and "post-peak"). The episodic nature of the data is due to the complexities of the underlying physiology of stress and the particulars of the stress classifier.

We now use a subset of the data collected in an observational, no treatment, smoking cessation study of 61 cigarette smokers (here on called the "Minnesota dataset") (Saleheen et al. (2015)) to inform the generative model of longitudinal trajectory $\{X_t, I_t\}_{t=1}^{T}$. Of the 61 participants 50 had sufficiently high-quality electrocardiogram data to construct the episodes and infer the stress classification. This subset is reported in Sarker et al. (2017). From this data, we calculate the sample moments:

1. For each episode type (i.e., $x \in \{0, 1\}$), the probability that the next episode will be a stress episode, that is, a 2 by 1 vector $\bar{W}$;
2. For each episode type (i.e., $x \in \{0, 1\}$), the average episode length, that is, a 2 by 1 vector $\bar{Z}$.

The sample moments are $\bar{W} = (0.067, 0.519)$ and $\bar{Z} = (10.9, 12.0)$.

Using these sample moments, we construct a no-treatment transition matrix for the joint process $V_t = (X_t, U_t), t = 1, \ldots, 600$. Each episode ends in state $V_t = (x, 2)$ for $x \in \{0, 1\}$ and transitions to the beginning of the next episode, $V_{t+1} = (x', 0)$ for $x' \in \{0, 1\}$. We restrict the transition matrix such that for $x \in \{0, 1\}$:

- $(x, 0)$ can *only* transition to states $(x, 0)$ or $(x, 1)$ (i.e., stay in state "pre-peak" or transition to state "peak") from one minute to the next minute.
- $(x, 1)$ transitions immediately to $(x, 2)$ with probability one (i.e., $\text{pr}(V_{t+1} = (x, 2)|V_t = (x, 1)) = 1$); in other words, the process inhabits the "peak" state for *only* one minute.
- $(x, 2)$ can *only* transition to states $(x, 2)$, $(0, 0)$ *or* $(1, 0)$ (i.e., stay in state "post-peak" or end the episode and begin a new one).

We label each episode depending on the value x. We use the approximation, $U_t \neq 1$ implies $I_t = 0$. In this minute the episode is classified as stressed or not classified as stressed. Define $\tilde{Z}_{(x,u)}$ to be the length of the phase u in an episode of type x after the chain enters state (x, u). Then $\tilde{Z}_{(x,1)} = 0$ for each x because as soon as the chain enters the peak ($u = 1$) of an episode, the chain departs. Otherwise, set $\tilde{Z}_{(x,u)} = (\bar{Z}_x - 3)/2$ for $u = 0$ and $u = 2$[3].

We set the no-treatment transition probability matrix to

$$P^{(0)}_{(x,u),(x,u)} = \tilde{Z}_{x,u}/(\tilde{Z}_{x,u} + 1) \quad \text{and} \quad P^{(0)}_{(x,1),(x,2)} = 1.0$$

[3]We subtract three as we are guaranteed one pre-peak, one peak and one post-peak minute in each episode. Dividing by two splits the remaining average time evenly between pre-peak and post-peak phases of an episode.

TABLE 1
$P^{(0)}$: *Transition Matrix for the Markov chain, V_t, under No Treatment*

		Nonstress			Stress		
		Pre-peak	Peak	Post-peak	Pre-peak	Peak	Post-peak
Nonstress	Pre-peak	0.80	0.20	0.00	0.00	0.00	0.00
	Peak	0.00	0.00	1.00	0.00	0.00	0.00
	Post-peak	0.19	0.00	0.80	0.01	0.00	0.00
Stress	Pre-peak	0.00	0.00	0.00	0.82	0.18	0.00
	Peak	0.00	0.00	0.00	0.00	0.00	1.00
	Post-peak	0.09	0.00	0.00	0.09	0.00	0.82

for $x \in \{0, 1\}$ and $u \in \{0, 2\}$, and then set

$$P^{(0)}_{(x,2),(0,0)} = (1 - \bar{W}_x)(1 - P_{(x,2),(x,2)}) \quad \text{and} \quad P^{(0)}_{(x,2),(1,0)} = \bar{W}_x(1 - P_{(x,2),(x,2)})$$

for $x \in \{0, 1\}$ (recall that $\bar{W}_x$ is the estimated probability that the next episode will be a stress episode). All other entries of $P^{(0)}$ are set to zero. Thus $P^{(0)}$ is a deterministic function of the moments $\bar{W}$ and $\bar{Z}$. See Table 1 for the transition matrix $P^{(0)}$.

The transition matrix $P^{(0)}$ specified in Table 1 has stationary distribution ($\pi_{(0,0)} = 0.394$, $\pi_{(0,1)} = 0.080$, $\pi_{(0,2)} = 0.394$, $\pi_{(1,0)} = 0.061$, $\pi_{(1,1)} = 0.011$, $\pi_{(1,2)} = 0.061$).

7.2. *Generative model under treatment.* Next, we form the generative model under treatment. We make the simplifying assumption that following treatment (i.e., $A_t = 1$), V_{t+j} evolves as a discrete-time Markov chain but with respect to a different transition matrix $P^{(1)}_t$ for each of the subsequent $j = 1, \ldots, 60$ minutes. After the hour, assuming a subsequent treatment notification is not provided, the time-varying stratification variable returns to evolution as a Markov chain with transition matrix $P^{(0)}$. Thus,

$$\begin{aligned}&\operatorname{pr}\big(V_t = (x, u) | V_{t-1} = \big(x', u'\big), H_{t-1}\big)\\ &\quad = \begin{cases} \big[P^{(0)}\big]_{(x',u'),(x,u)} & \text{if } A_{t-s} = 0, s = 1, \ldots, 60,\\ \big[P^{(1)}_t\big]_{(x',u'),(x,u)} & \text{otherwise.}\end{cases}\end{aligned}$$

Because the alternative $\beta(t; x)$ is constant within each day, we will construct a transition matrix, $P^{(1)}_t$, that will only depend on t through the day of decision t. Thus, we use the notation $P^{(1)}_{d(t)}$ instead of $P^{(1)}_t$ where $d(t)$ is the day of decision time t.

Recall the treatment effect is the effect of providing a notification at time t to practice stress-reduction exercises and no more notifications within the next hour versus no notification at time t and no notifications over the next hour on the percent of time stressed in the next hour. Thus, the reference policy sets the treatments $a_{t+1}, \ldots, a_{t+\Delta-1}$ to 0 and the expected proximal response under the reference policy can be computed analytically for any combination of x and a ($\Delta = 60$). See Section K.1 of the supplementary material (Dempsey et al. (2020)) for derivations of the below analytic forms. When $a = 1$, under the proposed generative model the above expectation is equal to $\Delta^{-1} \sum_{s=1}^{\Delta} \sum_{u \in \{0,1,2\}} [(P^{(1)}_{d(t)})^s]_{(x,1),(1,u)}$. When $a = 0$, the expectation is equal to the fraction of time stressed within the next hour under the reference policy of no actions for that hour

$$\Delta^{-1} \sum_{s=1}^{\Delta} \sum_{u \in \{0,1,2\}} [(P^0)^s]_{(x,1),(1,u)}.$$

Given the alternative $\beta(t;x)$ for a particular day, we set $P^{(1)}_{d(t)}$ equal to

$$\arg\min_{Q\in\mathcal{P}} \sum_{x\in\{0,1\}} \left(\Delta^{-1} \sum_{s=1}^{\Delta} \sum_{u\in\{0,1,2\}} ([Q^s]_{((x,1),(1,u))} - [(P^{(0)})^s]_{((x,1),(1,u))}) - \beta(t;x) \right)^2,$$

where $\mathcal{P}$ denotes the set of transition matrices which satisfy the constraints discussed above. The set $\mathcal{P}$ can be parameterized in order to use general-purpose, box-constrained optimization methods to calculate $P^{(1)}_{d(t)}$ efficiently. For all calculations we initialize with inputs equivalent to the transition matrix $P^{(0)}$. Using this procedure, the maximum squared distance across all alternatives $\beta(t;x)$ considered in this paper is 2.71×10^{-11} (i.e., low approximation error).

7.3. *Generating the simulated data.* The prior section yields the no-treatment and treatment transition matrices (i.e., $P^{(0)}$ and $\{P^{(1)}_d\}_{d=1}^{10}$), given the specified alternative $\{\beta(t;x)\}$. We briefly show how to use this information along with the randomization probability formula to generate synthetic data arising from a stratified micro-randomized trial. First, we generate data for each day independently. On a given day at time t, we first generate V_t using the transition equation in Section 7.2. We then assess availability, I_t, which is a deterministic function of the current value of V_t and the past 60 minute history of actions $\{A_{t-s}\}_{s=1,\ldots,60}$. That is, $I_t = \mathbf{1}[\sum_{s=1}^{60} A_{t-s} = 0] \times \mathbf{1}[U_t = 1]$. We adjust availability in the first hour of each day to be only a function of whether an intervention was already provided that day. Given $I_t = 1$, we take the history H_t and generate the action at time t, A_t, using the given randomization probability formula $p_t(1|H_t)$ found in Section F of the supplementary material (Dempsey et al. (2020)). In order to compute the proximal response $Y_{t,\Delta}$ for every minute over the 10 hour day (i.e., $t = 1, \ldots, 600$), we simulate an additional eleventh hour during which participants cannot receive treatment (i.e., participants are unavailable). The above procedure generates synthetic data for one participant in a stratified micro-randomized trial.

7.3.1. *The test statistic.* The above provides the generative model for use in the simulation-based sample size calculator. Next, consider the choice of the test statistic for use in calculating the sample size. In the test statistic (4) we set the time t reference probability $\tilde{p}_t(1|x)$ equal to $\sum_{x=0,1} \pi_{(x,1)}(1.5/[(600 - 1.5 \cdot 60)\pi_{(x,1)}]) = 5.88 \times 10^{-3}$. Recall that the numerator of the weight, w_{ct}, in (3) is $\tilde{p}_t(A_t|x)\prod_{s=1}^{\Delta-1}\mathbf{1}[A_{t+s} = 0]$. The probability, $\tilde{p}_t(1|x)$ is equal to the daily average number of treatments while in state x divided by the daily average number of times the participant is available and in state x, marginalized over the state x. In the denominator the term $1.5 \cdot 60$ is subtracted off the total number of decision points due to the availability constraints following treatment.

The test statistic (4) with the above choice of reference probabilities, and the above generative model are used to generate the sample sizes in Table 2. The column labeled "Sample Size" in this table provides the estimated sample size to detect a specified alternative for the conditional proximal effect given power of 80% and significance level 5% for the Sense2Stop study. Recall that our input for the day of maximal effect is day 5 and the input for the initial

TABLE 2
Estimated sample size, N, achieved power, and achieved type I error

	Sample size	Power	Type I error
$\bar{\beta} = 0.030$	50	80.6%	5.1%
$\bar{\beta} = 0.025$	67	80.7	4.4
$\bar{\beta} = 0.020$	127	80.6	5.6

conditional effect is 0 for both levels of the time-varying variable X_t. The average treatment effects $\{\bar{\beta}_x = T^{-1}\sum_{t=1}^{T}\beta(t;x)\}_{x=0,1}$ are assumed equal across levels X and set to $\bar{\beta}$; in the tables below three values of $\bar{\beta}$ are considered. Achieved significance levels under the null are included.

7.4. *Evaluation of simulation calculator for the smoking cessation study.* Recall that the relatively simple generative model allowed us to use only a very few statistics from the Minnesota dataset described in Section 7.1.1. There are two concerns: (1) whether or not the participants in the Minnesota study are representative of the future participants in Sense2Stop, and (2) whether or not the generative model built upon a few sample moments adequately captures the variation in the unknown longitudinal distribution $(X_t, U_t, I_t)_{t=1}^{T}$ under no treatment. If (1) does not hold, then the sample moments may be biased (scenario A). If (1) holds and (2) does not, it may be that prior scientific knowledge can suggest potential deviations that are difficult to estimate given the small size of prior data (scenario B). If (1) holds and (2) does not, alternatively, it may be that we can account for additional variation via fitting more complex models to the Minnesota dataset (scenario C). Any of these scenarios may problematic if it results in sample sizes for which the power to detect the desired effect is below the specified power. Therefore, we construct a feasible set of alternative generative models to which the sample size calculator should be robust. Note, this is not an exhaustive list, but highlights three important scenarios we expect to occur in practice.

7.4.1. *Misspecification of transition matrix $P^{(0)}$.* For scenario A, we consider situations where the generative model for the future Sense2Stop participants can be constructed in the same manner; however, the correct moment inputs for Sense2Stop are deviations from the sampled moments of the Minnesota study. Let $B_{(\epsilon,\epsilon')}$ denote an (ϵ,ϵ')-ball around the inputs $(\bar{W}, \bar{Z})$; that is,

$$B_{(\epsilon,\epsilon')} = \big\{(W,Z) \,|\, \|W - \bar{W}\|_\infty \le \epsilon \text{ and } \|Z - \bar{Z}\|_\infty \le \epsilon'\big\}.$$

For each $(W, Z) \in B_{(\epsilon,\epsilon')}$, we wish to compute the achieved power under the alternative generative model where V_t under no treatment evolves as a Markov chain with transition matrix P constructed from inputs W and Z; however, this is computationally prohibitive. Simulation suggests power to be a smooth, nonincreasing function of both ϵ and ϵ', so instead we focus on computing power for the following subset of $B_{(\epsilon,\epsilon')}$:

$$\Omega_{(\epsilon,\epsilon')} = \{(W,Z) | W \in \bar{W} \pm \big\{(\epsilon,-\epsilon),(\epsilon,\epsilon)\big\} \quad \text{and} \quad Z \in \bar{Z} \pm \big\{(\epsilon',-\epsilon'),(\epsilon',\epsilon')\big\}.$$

For each pair $(W, Z) \in \Omega_{(\epsilon,\epsilon')}$ we compute the associated transition matrix P; then we compute the sequence of transition matrices $P_{d(t)}^{(1)}$ which maintain the correct alternative treatment effect. We define the power for $B_{(\epsilon,\epsilon')}$ to be the minimum power across $(W, Z) \in \Omega_{(\epsilon,\epsilon')}$.

Selection of (ϵ, ϵ') is driven by observed variation in the Minnesota dataset. For selection of ϵ', we note the standard deviation of nonstress and stress episode durations in the Minnesota dataset is 6.89 and 6.48, respectively. Moreover, the standard errors in the sample moment $\bar{Z}$ were only 0.12 and 0.28, respectively. Thus, we chose $\epsilon' \in \{2, 4\}$. To select ϵ, we observe the standard error for the moment estimates $\bar{W}$ are 0.005 and 0.03 for nonstress and stress episodes, respectively. Thus, we set $\epsilon \in \{0.01, 0.02\}$.

Table 3 presents achieved power under the previously calculated sample sizes for $\Omega_{(0.02,4)}$ and $\Omega_{(0.01,2)}$, respectively. For both $(\epsilon, \epsilon') = (0.01, 2)$ and $(\epsilon, \epsilon') = (0.02, 4)$, the achieved power is significantly below the prespecified 80% level for all three choices of the average treatment effect $\bar{\beta}$.

TABLE 3
Misspecification of transition matrix $P^{(0)}$: minimum achieved power over set of matrices in $\Omega_{\epsilon,\epsilon'}$

	$(\epsilon, \epsilon') =$	
	(0.02, 4)	(0.01, 2)
$\bar{\beta} = 0.030$	57.5%	61.5%
$\bar{\beta} = 0.025$	43.9	52.2
$\bar{\beta} = 0.020$	40.4	65.6

7.4.2. *Deviations from a time-homogenous transition matrix under no treatment.* For scenario B we consider a deviation suggested by prior knowledge, namely, that stress dynamics are different over the weekend from the weekday. Given the prior Minnesota study was small, this proposed deviation is not data driven but scientifically motivated. This suggests a different type of misspecification of the transition matrix $P^{(0)}$, that of time-inhomogeneity; as before the treatment effect is still correctly specified. In particular, suppose that the assumed transition matrix, $P^{(0)}$, is correct for weekdays but not for weekends; in particular, suppose in reality that the transition matrix under no treatment on the weekend is $P^{(0)}_{\text{weekend}} \neq P^{(0)}$. The weekend is defined as $d(t) = 6$ and 7 (i.e., all participants enter the study on a Monday). We specify $P^{(0)}_{\text{weekend}}$ via inputs $(\bar{W}_{\text{weekend}}, \bar{Z}_{\text{weekend}})$ which we set to two possible values,

$$\underbrace{\big((0.04, 0.45), (10.9, 12.0)\big)}_{\text{weekend inputs (1)}} \quad \text{or} \quad \underbrace{\big((0.10, 0.60), (10.9, 12.0)\big)}_{\text{weekend inputs (2)}}.$$

Using the inputs, we construct two alternate versions of what the true transition matrix $P^{(0)}_{\text{weekend}}$ might be. For input (1) the individual is less likely to enter a stress episode over the weekend; for input (2) the individual is more likely to enter a stress episode over the weekend. In both cases the average episode lengths are assumed equal to $\bar{W}$.

Table 4 presents achieved power under these alternative generative models. We see that the achieved power is below the prespecified 80% threshold in each case except for $\bar{\beta} = 0.020$ under weekend input 1. If the scientist thought such deviations feasible, then the above analysis suggests for Sense2Stop that the sample size be set to ensure a least 80% power over a set of feasible choices for time-inhomogeneous choices for the no-treatment transition matrix.

7.4.3. *Deviations from a Markovian generative model.* For scenario C we fit a semi-Markov generative model to the small, observational Minnesota study. This accounts for additional variation in the prior study, but the resulting generative model may not represent

TABLE 4
Estimated power under generative model with time-inhomogeneous Markov chain

	Estimated power	
	Weekend Input 1	Weekend Input 2
$\bar{\beta} = 0.030$	79.2	69.8
$\bar{\beta} = 0.025$	72.5	66.0
$\bar{\beta} = 0.020$	81.5	76.4

TABLE 5
Parameter estimates for the logistic regression. Response is indicator of current episode being a stress episode

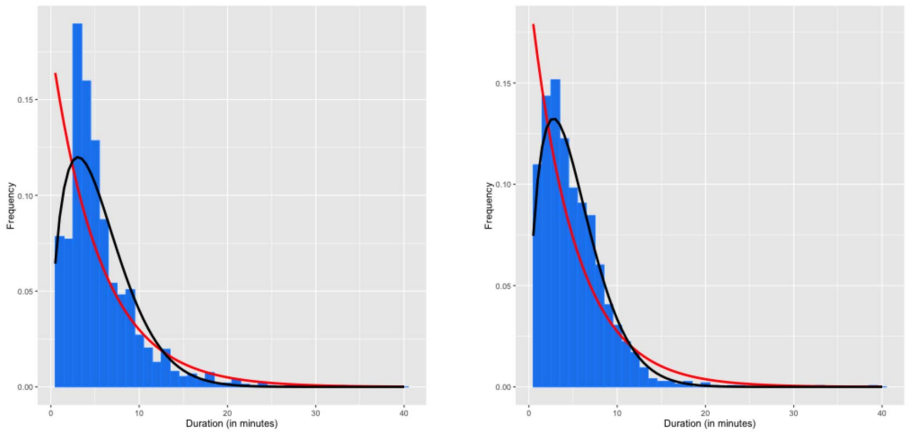
Parameter	Estimate	Std. Error	95% LCL	95% LCL
Intercept	−2.83	0.10	−3.03	−2.63
1L Stress Ep.	2.75	0.20	2.37	3.14
2L Stress Ep.	0.71	0.22	0.27	1.14

behavior for future Sense2Stop participants. That is, the model may overfit the Minnesota study data and not generalize well to the future Sense2Stop participants. After presenting data analysis for the semiMarkovian deviation, we then assess robustness of the sample size calculator to this data-driven deviation.

We start by considering the episodic transition rule. The Markovian model assumes that the episode transitions only depend on the prior episode classification. We test this by fitting a logistic regression with episode classification as the response variable and lagged values of episode classification as well as additional summaries of past history, including prior episode durations and time of day as covariates. Analysis suggests that neither time of day nor prior episode duration were statistically significant. We used forward selection to determine the number of lagged values of episode classification, leading to inclusion of two lags. Table 5 presents the estimates of the logistic regression along with robust standard errors and confidence intervals.

This model leads to slightly distinct behavior of the transition rules. For example, given the prior episode was a stress episode, the probability of the next episode being a *stress* episode ranges from 0.480 (two-lagged prior episode was nonstress) to 0.652 (two-lagged prior episode was stress). Given the prior episode was a nonstress episode, the probability of the next episode being a *stress* episode ranges from 0.056 (two-lagged prior episode was nonstress) to 0.107 (two-lagged prior episode was stress). Table 5 leads to a different Markovian model in which the state is $(X_t, U_t, L_t^{(1)})$ where $L_t^{(1)}$ denotes the classification of the prior episode.

We next examine the pre- and post-peak durations. Figure 3 shows histograms of the pre and postpeak durations in the analyzed subset of data along with empirical Bayes estimates



(a) Pre-peak duration (b) Post-peak duration

FIG. 3. *Histograms of duration for pre/post-peak durations for Minnesota study. Empirical bayes pdfs for exponential (red) and Weibull (black) densities are overlayed.*

TABLE 6
Parameter estimates for each Weibull survival regression

	Pre-peak			Post-peak		
Parameter	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
Intercept	1.78	0.016	0.000	1.59	0.02	0.000
0L Stress Ep.	−0.20	0.037	0.000	0.45	0.07	0.000
1L Stress Ep.	–	–	–	−0.21	0.058	0.004
2L Stress Ep.	–	–	–	−0.16	0.07	0.020
Log(scale)	−0.24	0.015	0.000	−0.31	0.05	0.000

of the probability density functions under both exponential and Weibull distribution specifications. We recognize the durations are discrete and the above distributions are continuous. These are fit for simplicity. When generating the episode duration, we generate a random variable from the continuous distribution and take the integer part of that random variable. It is evident from the figures that the Weibull distribution is more appropriate.

Table 6 presents the parameter estimates for this over-fit model to the duration data assuming a Weibull distribution.[4] Like the episodic transition rules, the post and pre-peak durations now depend on the current episode classification as well as the prior episode classifications. The exploratory data analysis suggests a semiMarkovian model in which the pre-/post-peak durations are Weibull distributed, and the state is given by $(X_t, U_t, L_t^{(1)}, L_t^{(2)})$ where $L_t^{(i)}$ denotes the classification of the ith prior episode. For the pre-peak model, the one and two-lagged indicators of a stress episode ("1L" and "2L Stress Ep." in Table 6) were insignificant and thus excluded from the model.

Next, we test robustness of the sample size calculator to the semiMarkovian deviations described above. To test the calculator, we generate data using the no-treatment semiMarkov model specified in Section L in the supplementary material (Dempsey et al. (2020)). The data is simulated so that the treatment effect used by the calculator is correct. See Section L in the supplementary material (Dempsey et al. (2020)) for a discussion of how this was achieved. Table 7 presents achieved power under these alternative generative models. We see that the achieved power is well above the prespecified 80% threshold in each case. Therefore, the sample size calculator is robust to such complex deviations from the Markovian generative model. For the given alternative $\beta(t; x)$ and semiMarkov generative model, we calculate the standardized effects. These are provided in Table 7 in Section K of the supplementary material (Dempsey et al. (2020)).

7.5. *Adjustments to the simulation-based calculator.* In Section 7.4 we evaluated the simulation calculator built in Section 7.1. Here, we make adjustments to the simulation calcu-

TABLE 7
Estimated power under semiMarkov generative

	Estimated power
$\bar{\beta} = 0.030$	93.6
$\bar{\beta} = 0.025$	88.0
$\bar{\beta} = 0.020$	93.6

[4]Models are fit to duration minus one as pre- and post-peak durations are guaranteed to be greater than one. Thus, we are modeling the duration in the state above the minimum value of one.

TABLE 8
Estimated sample size, N, and computed power under $\epsilon = 2$ and $\epsilon' = 0.01$

	Sample size	Minimum power
$\bar{\beta} = 0.030$	69	81.9%
$\bar{\beta} = 0.025$	107	80.4
$\bar{\beta} = 0.020$	208	80.5

lator to ensure robustness. First, we note that the simulation calculator is robust to the potential semiMarkovian deviation discussed in Section 7.4.3. Next, we make the decision that we are not concerned with lack of robustness to deviations from a time-homogenous transition matrix as discussed in Section 7.4.2. Therefore, we focus on making the simulation calculator robust to misspecification of Markov transition matrix as discussed in Section 7.4.1.

Analysis in Section 7.4.1 suggests for Sense2Stop that the sample size should be set to ensure at least 80% power *over a set of feasible choices for the transition matrix* $P^{(0)}$. We fix $(\epsilon, \epsilon') = (0.01, 2)$ to be our tolerance to misspecification of the inputs. For each set of inputs $(W, Z) \in \Omega_{0.01,2}$, we compute a sample size using the simulation calculator built in Section 7.1. The maximum of this set of computed sample sizes is chosen to ensure tolerance to misspecification of the transition matrix. Table 8 presents the sample size under this procedure as well as the achieved *minimum power* over the set $\Omega_{\epsilon,\epsilon'}$.

We have now used the three-step procedure to form a sample size calculator for the smoking cessation study example. For illustration suppose we wish to detect an average conditional treatment effect $\bar{\beta}$ equal to 0.025. Based on the above discussion a sample size N of 107 would be recommended to ensure power above the prespecified 80% threshold across a set of feasible deviations from the assumed generative model.

8. Conclusion. In this paper we introduced the stratified micro-randomized trial (sMRT) and provided a definition and discussion of proximal treatment effects along with the dependence of this definition on a reference distribution. We proposed a simulation-based approach for determining sample size and used this approach to determine the sample size for a simplified version of the MD2K smoking cessation study. We expect that similar trial designs would be applicable in areas such as marketing and advertising in which each client is tracked and provided incentives, for example, treatments repeated over time, and it is of interest to determine in which contexts particular treatments are most effective.

An alternative test to our projection-based method is a randomization-based test. In Bojinov and Shephard (2019), exact randomization based p-values are constructed for testing causal effects in single time series experiments. The approach relies solely on random assignment of treatment paths rather than the distribution of the test statistic for the validity of the test (Rosenberger, Uschner and Wang (2019)). Randomization inference, however, targets a sharp null that the treatment has no effect on the distribution of *all* the time-varying endogenous variables (i.e., in our setting across availability I_t, phase U_t, stratification X_t and response $Y_{t,\Delta}$ variables). Our inferential target is more restrictive; our goal is to assess if the conditional mean for a specific outcome $Y_{t,\Delta}$ given availability (i.e., $I_t = 1$) and stratification variable (i.e., $X_t = x$) is equal to zero jointly across time $t = 1, \ldots, T$ and strata $x = 0, 1$. The authors would be very interested in future work that extends the randomization test framework to our inferential target.

While the focus here is sample size considerations, stratified micro-randomized studies yield data for a variety of interesting secondary data analyses. For example, understanding

predictors of future availability is of general interest as keeping participants engaged in the mobile health intervention is often of high concern. Moreover, there is interest in using the data in constructing "dynamic treatment regimes" (e.g., just-in-time adaptive interventions (Spruijt-Metz and Nilsen (2014))). The stratified micro-randomized trial improves such analyses by reducing causal confounding.

SUPPLEMENTARY MATERIAL

Supplementary material for "The stratified micro-randomized trial design: Sample size considerations for testing nested causal effects of time-varying treatments" (DOI: 10.1214/19-AOAS1293SUPP; .pdf). This supplement provides important details on the Sense2Stop stratified micro-randomized trial design, additional comments on sMRT design choices, proofs and technical derivations, and sample size calculations for the marginal proximal effect.

REFERENCES

BIDARGADDI, N., ALMIRALL, D., MURPHY, S., NAHUM-SHANI, I., KOVALCIK, M., PITUCH, T., MAAIEH, H. and STRECHER, V. (2018). To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR Mhealth Uhealth* **6** e10123. https://doi.org/10.2196/10123.

BOJINOV, I. and SHEPHARD, N. (2019). Time series experiments and causal estimands: Exact randomization tests and trading. *J. Amer. Statist. Assoc.* **114** 1665–1682. https://doi.org/10.1080/01621459.2018.1527225

BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274

DEMPSEY, W., LIAO, P., KLASNJA, P., NAHUN-SHANI, I. and MURPHY, S. A. (2015). Randomised trials for the fitbit generation. *Significance* **12** 20–23.

DEMPSEY, W., LIAO, P., KUMAR, S. and MURPHY, S. A (2020). Supplement to "The stratified micro-randomized trial design: Sample size considerations for testing nested causal effects of time-varying treatments." https://doi.org/10.1214/19-AOAS1293SUPP.

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

ERTIN, E., STOHS, N., KUMAR, S., RAIJ, A., AL'ABSI, M. and AUTOSENSE, S. S. (2011). Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems* 274–287, New York, NY, USA.

FREE, C., PHILLIPS, G., GALLI, L., WATSON, L., FELIX, L., EDWARDS, P., PATEL, V. and HAINES, A. (2013). The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: A systematic review. *PLoS Med.* **10** 1–45.

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 https://doi.org/10.1198/016214506000000447

HOTELLING, H. (1931). The generalization of student's ratio. *Annals of Mathematical Sciences* **2** 360–378.

HOVSEPIAN, K., AL'ABSI, M., ERTIN, E., KAMARCK, T., NAKAJIMA, M. and KUMAR, S. (2015). cstress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp* '15 493–504. ACM, New York, NY, USA.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

KLASNJA, P., HEKLER, E. B., SHIFFMAN, S., BORUVKA, A., ALMIRALL, D., TEWARI, A. and MURPHY, S. A. (2015). Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* **34** 1220–1228.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LIAO, P., KLASNJA, P., TEWARI, A. and MURPHY, S. A. (2016). Micro-randomized trials in mHealth. *Stat. Med.* **35** 1944–1971. MR3513494 https://doi.org/10.1002/sim.6847

MANCL, L. A. and DEROUEN, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57** 126–134. MR1833298 https://doi.org/10.1111/j.0006-341X.2001.00126.x

MANCL, L. A. and LEROUX, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics* **52** 500–511.

PEARL, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. MR2545291 https://doi.org/10.1214/09-SS057

PEPE, M. S. and ANDERSON, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simulation Comput.* **23** 939–951.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758 https://doi.org/10.1016/0270-0255(86)90088-6

ROSENBERGER, W. F., USCHNER, D. and WANG, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Stat. Med.* **38** 1–12. MR3887263 https://doi.org/10.1002/sim.7901

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

SALEHEEN, N., ALI, A. A., HOSSAIN, S. M., SARKER, H., CHATTERJEE, S., MARLIN, B., ERTIN, E., AL'ABSI, M. and KUMAR, S. (2015). puffmarker: A multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the* 2015 *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, *UbiComp* '15 999–1010, ACM, New York, NY, USA. Available at http://doi.acm.org/10.1145/2750858.2806897.

SARKER, H., TYBURSKI, M., RAHMAN, M. M., HOVSEPIAN, K., SHARMIN, M., EPSTEIN, D. H., PRESTON, K. L., FURR-HOLDEN, C. D., MILAM, A. et al. (2016). Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the* 2016 *CHI Conference on Human Factors in Computing Systems*, *CHI* '16 4489–4501. ACM, Santa Clara, CA, USA.

SARKER, H., HOVSEPIAN, K., CHATTERJEE, S., NAHUM-SHANI, I., MURPHY, S. A., SPRING, B., ERTIN, E., AL'ABSI, M., NAKAJIMA, M. et al. (2017). From markers to interventions: The case of just-in-time stress intervention. In *Mobile Health Sensors*, *Analytic Methods*, *and Applications* (J. M. Regh, S. A. Murphy and S. Kumar, eds.) Springer, Berlin.

SPRUIJT-METZ, D. and NILSEN, W. (2014). Dynamic models of behavior for just-in-time adaptive interventions. *Pervasive Computing*, *IEEE* **13** 13–17.

TCHETGEN TCHETGEN, E. J., GLYMOUR, M. M., WEUVE, J. and ROBINS, J. (2012). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. *Epidemiology* **23** 644–646.

VANDERWEELE, T. J., HONG, G., JONES, S. M. and BROWN, J. L. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *J. Amer. Statist. Assoc.* **108** 469–482. MR3174634 https://doi.org/10.1080/01621459.2013.779832

VANSTEELANDT, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scand. J. Stat.* **34** 478–498. MR2368794 https://doi.org/10.1111/j.1467-9469.2006.00555.x

Wikipedia. F1 score—Wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/wiki/F1_score. [Online; accessed 23-May-2017].