A flexible micro-randomized trial design and sample size considerations

Xu, Jing; Yan, Xiaoxi; Figueroa, Caroline; Williams, Joseph Jay; Chakraborty, Bibhas

DOI
10.1177/09622802231188513

Publication date
2023

Document Version
Final published version

Published in
Statistical Methods in Medical Research

Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
1–18

DOI: 10.1177/09622802231188513
journals.sagepub.com/home/smm

Sage

A flexible micro-randomized trial design and sample size considerations

Jing Xu[1,2], Xiaoxi Yan[1], Caroline Figueroa[3,4], Joseph Jay Williams[5,6,7,8,9,10] and Bibhas Chakraborty[1,2,11,12]

Abstract

Technological advancements have made it possible to deliver mobile health interventions to individuals. A novel framework that has emerged from such advancements is the just-in-time adaptive intervention, which aims to suggest the right support to the individuals when their needs arise. The micro-randomized trial design has been proposed recently to test the proximal effects of the components of these just-in-time adaptive interventions. However, the extant micro-randomized trial framework only considers components with a fixed number of categories added at the beginning of the study. We propose a more flexible micro-randomized trial design which allows addition of more categories to the components during the study. Note that the number and timing of the categories added during the study need to be fixed initially. The proposed design is motivated by collaboration on the Diabetes and Mental Health Adaptive Notification Tracking and Evaluation study, which learns to deliver effective text messages to encourage physical activity among patients with diabetes and depression. We developed a new test statistic and the corresponding sample size calculator for the flexible micro-randomized trial using an approach similar to the generalized estimating equation for longitudinal data. Simulation studies were conducted to evaluate the sample size calculators and an R shiny application for the calculators was developed.

Keywords

mHealth, just-in-time adaptive intervention, micro-randomized trial, generalized estimating equation, longitudinaldata

1 Introduction

Mobile health (mHealth) is a term used to refer to the practice of medicine and health supported by mobile or wearables devices[1] that are increasingly indispensable in our daily lives. It provides convenient support to various health domains

[1]Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore
[2]Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore
[3]Faculty of Technology, Policy and Management, Delft University of Technology, The Netherlands
[4]School of Social Welfare, University of California, Berkeley, USA
[5]Department of Computer Science, University of Toronto, ON, Canada
[6]Department of Statistical Sciences, University of Toronto, ON, Canada
[7]Department of Psychology, University of Toronto, ON, Canada
[8]Vector Institute for Artificial Intelligence Faculty Affiliate, University of Toronto, ON, Canada
[9]Department of Mechanical and Industrial Engineering, University of Toronto, ON, Canada
[10]Department of Economics, University of Toronto, ON, Canada
[11]Department of Statistics and Data Science, National University of Singapore, Singapore
[12]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

Corresponding author:
Jing Xu, Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road Singapore 169857, Singapore.

Email: kenny.xu@duke-nus.ed.sg

including managing human immunodeficiency virus infection,[2] increasing physical activity,[3] supporting recovery from alcohol dependence[4] and smoking cessation.[5]

Mobile technology can be used to deliver *just-in-time adaptive interventions* (JITAIs), which aim to provide the right type or amount of support, at the right time,[6,7] according to an individual's evolving internal and contextual state. Nahum-Shani et al.[8] bridged the gap between the growing technological capabilities for delivering JITAIs and the research on the development and evaluation of these interventions.

The micro-randomized trial (MRT) design[9] has been proposed for testing the proximal effects of the intervention categories in JITAIs. The corresponding sample size calculation has been derived by Liao et al.[10] In an MRT, there are numerous decision time points for each participant throughout the study period. At each decision time point, a participant is randomly assigned to one of the available intervention options. There exist several research studies using the MRT design, for example, 'HeartSteps' for promoting physical activity among sedentary people,[11] 'Sense2Stop' for managing stress in newly abstinent smokers,[12] 'DIAMANTE'[13] for promoting physical activity among co-morbid diabetes and depression patients, 'StayWell'[14] for managing people's mental wellness during COVID-19 pandemic period, and so on.

In the existing MRT design, the categories of an intervention component are predetermined and fixed during the trial. With the goal to expand its scope, similar to adding new treatment arms during the study in platform trials (e.g. Ventz et al.[15] and Lee et al.[16]), we propose a novel flexible version of the MRT ('FlexiMRT') design that allows newly specified intervention categories to be added not only initially but also later in the study. Unlike the platform trials, however, the timings of adding new intervention categories and the number of intervention categories to be added at these times are pre-determined in FlexiMRTs. Going beyond prespecified categories is often needed to speed up research. For example, it enables using in-trial qualitative participant feedback and suggestions, to design and include new conditions, which can then be tested quantitatively. No matter how extensive, predeployment thinking or theory about what conditions to test can often miss better ideas, which only arise in the real-world context of diverse participants experiencing the intervention. For example, beyond health domains, rapid iterative cycles of design[17] is key to success in disciplines like human computer interaction and software product testing. Our methods support faster cycles of experimentation targeted at ecologically valid questions. In addition, novel techniques like crowdsourcing[18] enable applications in fields from social media to health to education where ideas are emerging after deployment. For example, Williams et al.[19] show the values of having new explanations of a concept continually added by people learning a topic, once they can be tested out quickly to assess efficacy. More broadly, applications of experimentation for successful apps, websites and products rely on cycles of continual testing and improvement.[20] Unlike the traditional long delay from trial to deployment, an app like Facebook or Google search will perpetually be having product teams add and modify conditions for testing,[21] on the order of thousands of experimental changes/deployments per day, since constantly changing product needs and context require more dynamic experimentation techniques. Our work aims to provide sounder methodological basis for such experiments in mHealth and beyond, to bring innovative techniques to bear on scientifically supported improvements for health interventions to help people.

In our proposed FlexiMRT design, for each intervention component, instead of estimating one proximal effect (i.e. the 'pooled' intervention categories versus the control category), the proximal effects of individual intervention categories versus no intervention (control) can be estimated. The individual effect sizes will be of interest if they are expected to be different in magnitudes or directions. Novel test statistics are derived to detect the proximal effects from the FlexiMRT data. We derive the corresponding sample size calculators so that the trials can be sized to either detect the proximal effect at a nominal power and a nominal type-I error rate, or to estimate the proximal effect within a certain precision or margin of error at a nominal confidence level. The latter approach is useful in a pilot study (e.g. the Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) students study by Figueroa et al.[22]) scenario, when prior information on the proximal effects may be limited.

This work is motivated by our collaboration on the DIAMANTE study.[13] The 'DIAMANTE' app sends multiple messages to encourage physical activity among co-morbid diabetes and depression patients from low-income and ethnic minority backgrounds served in the San Francisco Health Network.

We summarize the three main contributions of the article. First, we propose a novel FlexiMRT design, which allows for intervention categories to be added later in the study. Second, we develop two associated sample size calculation methods (one based on power and the other on precision) for the proposed design. Finally, based on the developed methodology, we create free online sample size calculators using R shiny (https://kennyxu.shinyapps.io/FlexiMRT-SS/) to facilitate wide dissemination. The relevant R functions are available in Github (https://github.com/Kenny-Jing-Xu/FlexiMRT-SS).

The rest of the article is organized as follows. Section 2 describes the proposed FlexiMRT design with the corresponding statistical analysis and sample size calculation approach. Section 3 demonstrates an application of the proposed sample size calculators based on an adapted student version of the DIAMANTE study. Section 4 investigates the performance of the

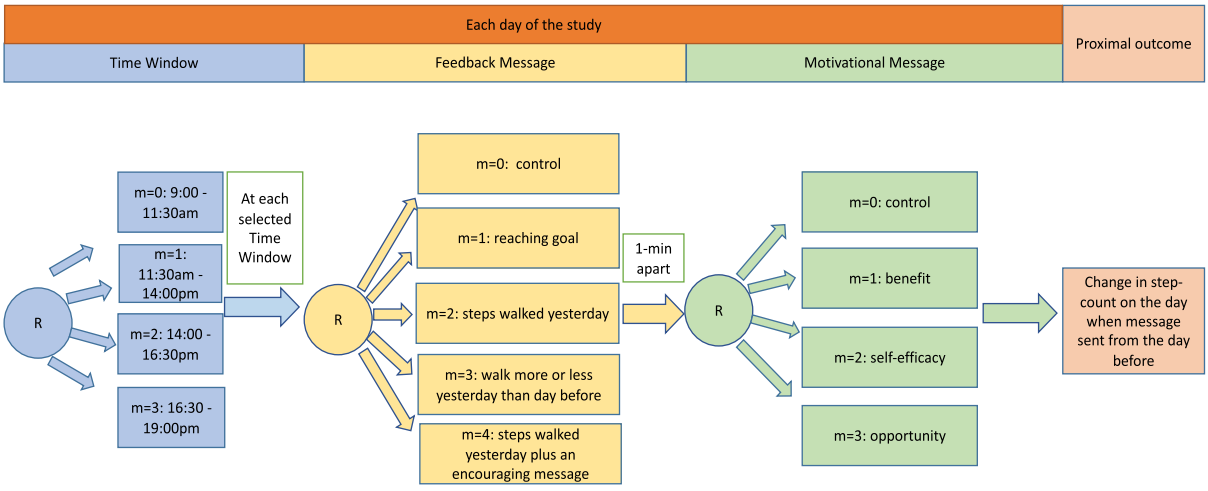

Figure 1. MRT design for DIAMANTE study, *R* = randomization and *m* = category. MRT: micro-randomized trial; DIAMANTE: Diabetes and Mental Health Adaptive Notification Tracking and Evaluation.

sample size calculators through simulation studies. The paper ends with a discussion in Section 5. The detailed derivations and the implementation of the proposed sample size calculators are deferred to the Supplemental Material.

2 FlexiMRT design: Statistical model estimation and sample size calculation

2.1 MRT design

MRT is a cutting-edge trial design suitable for time-varying, sequential, multi-component interventions, akin to a sequential full-factorial design. Within the study period, at decision time t, on day d, participant i is randomized to an intervention component denoted by A_{idt}, where $i = 1, \ldots, N$, $d = 1, \ldots, D$ and $t = 1, \ldots, T$. For example, $A_{idt} = 1$ if an intervention category is delivered with probability π_{dt}, otherwise $A_{idt} = 0$ if a control category is delivered. Though the extant MRT design[10] mainly focuses on components with binary categories only, we consider multiple categories in the present article. Let Y_{idt} denote the proximal outcome of participant i measured following time point t on day d.

A participant may not be in a position to receive an intervention at a decision point during the study. For example, it is not safe for an mHealth app to deliver an intervention to a participant who is driving. A participant is considered unavailable for randomization at such decision points, therefore, no intervention or control is delivered automatically. Let I_{idt} denote the availability indicator, that is, $I_{idt} = 1$ if participant i is available at decision point t on day d and $I_{idt} = 0$, otherwise. Note that $A_{idt} = 0$ if $I_{idt} = 0$ and the randomization probability is denoted by $\pi_{dt} = P(A_{idt} = 1 \mid I_{idt} = 1)$.

MRTs consider the longitudinal data, that is, $\boldsymbol{O}_i = (\boldsymbol{S}_i^{\top}, I_{i11}, A_{i11}, Y_{i11}, \ldots, I_{iDT}, A_{iDT}, Y_{iDT})$ for the observations of participant i, where $\boldsymbol{S}_i$ denotes the baseline covariates vector. We assume that $\boldsymbol{O}_i$, $i = 1, \ldots, N$ are independent and identically distributed. Note that the sample size calculation focuses on detecting the proximal effect of intervention.

2.2 DIAMANTE study

The FlexiMRT design is motivated by the DIAMANTE study that employs an MRT design with multi-category components and is summarized in Figure 1. This is a six-month study with one decision time point per day. On each day, each available participant is randomized to one of the categories of each of three components, namely, *Time Window* (four categories), *Feedback Message* (five categories), and *Motivational Message* (six categories). Thus, the DIAMANTE study allocates interventions according to a $4 \times 5 \times 4$ factorial design each day. The two different messages are sent 1 min apart.

The Time Window component ($\boldsymbol{A}^T$) specifies when to send the messages, that is, category-0 (9:00 a.m.–11:30 a.m.), category-1 (11:30 a.m.–14:00 p.m.), category-2 (14:00 p.m.–16:30 p.m.) and category-3 (16:30 p.m.–19:00 p.m.).

The Feedback Message component ($\boldsymbol{A}^F$) has a reference category-0 (i.e. no message) and four intervention categories, that is, category-1 (reaching goal), category-2 (steps walked yesterday), category-3 (walked more or less yesterday than the day before) and category-4 (steps walked yesterday plus a positive/negative message).

The Motivational Message component ($\boldsymbol{A}^M$) has a reference category-0 (no message) and three intervention categories, that is, category-1 (benefit), category-2 (self-efficacy) and category-3 (opportunity).

The proximal outcome (Y) is the daily steps change, that is, the step-count on the calendar day when intervention messages were sent minus the step-count on the previous calendar day, measured by the participants' phone pedometer, whose formula is defined in Section 2.3.

The availability indicator (I) at a given decision time codes one if a participant is available for intervention at that decision time and zero otherwise. The DIAMANTE study assumes 100% availability. This is reasonable because the proximal outcome is the change in daily steps following the messages, thus allowing participants sufficient time to respond to messages at the randomly selected time windows.

The current phase of the DIAMANTE trial is a conventional MRT where all the intervention categories are introduced upfront. At the macro level, this study involves a randomized control trial with three groups, namely, the uniform randomization (UR) group, the adaptive learning (AL) group and the control group, where the MRT design is embedded within the UR and AL groups. The participants of the UR group receive different messages at different times with equal probability while the participants of the AL group receive them with probabilities learned adaptively by a reinforcement learning algorithm aimed at maximizing the proximal effects. Though both the UR and AL groups can be modified into a FlexiMRT design, the proposed sample size calculators can be only used for the UR group.

For the statistical model and sample size calculators proposed through Sections 2.3 to 2.5, we consider the long format of longitudinal datasets. The observation on participant i on day d at decision time point t is

$$\{I_{idt}, A^T_{1idt}, A^T_{2idt}, A^T_{3idt}, A^F_{1idt}, A^F_{2idt}, A^F_{3idt}, A^F_{4idt}, A^M_{1idt}, A^M_{2idt}, A^M_{3idt}, Y_{idt}\}$$

where $\{A^T_1, A^T_2, A^T_3\}$, $\{A^F_1, A^F_2, A^F_3, A^F_4\}$ and $\{A^M_1, A^M_2, A^M_3\}$ represent the intervention category indicators of the Time Window, Feedback Message and Motivational Message components, respectively.

2.3 Statistical model

In this section, we first extend the statistical model for MRT[9] proposed by Liao et al.[10] to accommodate multi-category intervention components. Instead of estimating the proximal effect between two categories (active vs. control), we estimate such effects for multiple active categories with reference to control. The multi-category approach can be used to recognize whether the differences in effect sizes among the intervention categories are in magnitudes only (see Section 3) or if opposite direction of signs exist. A regression model considering the proximal effect of a single component can be defined as below. We assume one decision time point per day, that is, $T = 1$ for participant i on day d, with the proximal outcome denoted by Y_{id}. Note that the decision is to allocate a participant to one of the intervention categories. The allocation only depends on the randomization probabilities at the same time point, and does not depend on the outcome and randomization probabilities at the previous time points.

We define the proximal effect size of the message category m (vs. the control category 0) of a particular component on day d, denoted by a function $b_m(d; \boldsymbol{\beta}_m)$, as

$$b_m(d; \boldsymbol{\beta}_m) = Z^\top_{md}\boldsymbol{\beta}_m = \left[1, d-1, \ldots, (d-1)^{p_m-1}\right]\boldsymbol{\beta}_m \tag{1}$$

where $m = 1, \ldots, M$, $d = 1, \ldots, D$ (the study period in days) and parameter $\boldsymbol{\beta}_m = (\beta_{m1}, \ldots, \beta_{mp_m})^\top$.

Note that $\boldsymbol{Z}^\top_{md}$ in equation (1) corresponds to one decision time point per day as in DIAMANTE, which can be generalized by involving the number of decision time points T per day. We define the proximal effect size of category m at time point t of day d from a particular component, denoted by $b_m(d, t; \boldsymbol{\beta}_m)$, as

$$b_m(d, t; \boldsymbol{\beta}_m) = Z^\top_{mdt}\boldsymbol{\beta}_m = \left(1, \left[\frac{(d-1)T + t - 1}{T}\right], \ldots, \left[\frac{(d-1)T + t - 1}{T}\right]^{p_m-1}\right)\boldsymbol{\beta}_m$$

where $t = 1, \ldots, T$. The proximal effects of different categories on different days can vary. They may follow the constant, linear or quadratic trends, corresponding to $p_m = 1, 2$ or 3, respectively. Alternatively, the proximal effect trend for category-m can also be described as having a combination of the linear and constant trends, where it increases or decreases linearly until a turning point on day d^m_{turn} and plateaus afterwards, as demonstrated in the StayWell study results.[14] We call it the 'linear-plateau' trend. In this case, we can define the proximal effect size using a linear spline, that is

$$b_m(d, t; \boldsymbol{\beta}_m) = Z^\top_{mdt}\boldsymbol{\beta}_m = \left(1, \left[\frac{\min\left[d^m_{\text{turn}} - 1, d-1\right]T + t - 1}{T}\right]\right)\boldsymbol{\beta}_m \tag{2}$$

where we have $\boldsymbol{\beta}^\top_m = (\beta_{m1}, \beta_{m2})$.

Next, we consider the scenario where new message categories are added during the study. Let M_0 be the number of categories introduced initially on day d_0 (e.g. $d = 1$), M_1 be the number of message categories added on day d_1 (first adding day after the beginning of the study) and so on, and finally, M_k be the number of categories added at the last adding day d_k, where we have $d_0 \leq 1 < d_1 < \cdots < d_k \leq D$. Therefore, the total number of message categories $M = \sum_{j=0}^{k} M_j$. Note that for the experimental design and sample size calculation purposes, M_j and d_j, where $j = 1, \ldots, k$, are pre-determined. We assume that the participants have the same length of time until each new category is added. This ensures that all added messages have the same duration for each participant to estimate their proximal effects. Note that the effect size of each of new message category is undefined before their adding days. Suppose $d = d_1$, $m = M_0 + 1$, $T = 1$ and $p_m = 2$, then the proximal effect of category $M_0 + 1$ on day d_1 can be computed by $b_{M_0+1}(d_1; \boldsymbol{\beta}_{M_0+1}) = \beta_{(M_0+1)1} + \beta_{(M_0+1)2}(d_1 - 1)$ based on equation (1), where $b_{M_0+1}(d_1; \boldsymbol{\beta}_{M_0+1})$ is undefined before day d_1. Given that participant i is available for randomization on day d, we denote the message categories by $\boldsymbol{A}_{id} = (A_{i1d},\ldots,A_{iM_0d},\ldots,A_{i(\sum_{j=0}^{k-1} M_j+1)d},\ldots,A_{i(\sum_{j=0}^{k} M_j)d})^\top$ and assume that they follow a multinomial distribution, that is, $Multinomial(1\text{-}\sum_m \pi_{md}, \pi_{1d},\ldots,\pi_{(\sum_{j=0}^{k} M_j)d})$, where $\pi_{(\cdot)}$ denotes the randomization probability corresponding to $A_{i(\cdot)}$. Note that before the first adding day (i.e. $d < d_1$), the message categories $M_0 + 1$ to $\sum_{j=0}^{k} M_j$ are not available. Thus, the category indicators $A_{i(M_0+1)d},\ldots, A_{i(\sum_{j=0}^{k} M_j)d}$ and their corresponding randomization probabilities $\pi_{i(M_0+1)d},\ldots, \pi_{i(\sum_{j=0}^{k} M_j)d}$ have zero values. Therefore, to allow for the flexible addition of new intervention categories during the trial, we further remove the restriction on fixed allocation in the model. We then denote the proximal effects of all the message categories on day d to be

$$\boldsymbol{Z}_{1d}^\top\boldsymbol{\beta}_1, \ldots, \boldsymbol{Z}_{M_0d}^\top\boldsymbol{\beta}_{M_0}, \ldots, \boldsymbol{Z}_{\sum_{j=0}^{k-1} M_j+1d}^\top\boldsymbol{\beta}_{\sum_{j=0}^{k-1} M_j+1}, \ldots, \boldsymbol{Z}_{\sum_{j=0}^{k} M_jd}^\top\boldsymbol{\beta}_{\sum_{j=0}^{k} M_j}$$

The working model can be written as

$$\begin{aligned}
Y_{id} = {} & \boldsymbol{B}_d^\top\boldsymbol{\alpha} \\
& + (A_{i1d} - \pi_{1d})\boldsymbol{Z}_{1d}^\top\boldsymbol{\beta}_1 + \cdots + (A_{iM_0d} - \pi_{M_0d})\boldsymbol{Z}_{M_0d}^\top\boldsymbol{\beta}_{M_0} \\
& + (A_{i(M_0+1)d} - \pi_{(M_0+1)d})\boldsymbol{Z}_{(M_0+1)d}^\top\boldsymbol{\beta}_{(M_0+1)} \\
& + \cdots + (A_{i(M_0+M_1)d} - \pi_{(M_0+M_1)d})\boldsymbol{Z}_{(M_0+M_1)d}^\top\boldsymbol{\beta}_{(M_0+M_1)} \\
& \vdots \\
& + (A_{i(\sum_{j=0}^{k-1} M_j+1)d} - \pi_{(\sum_{j=0}^{k-1} M_j+1)d})\boldsymbol{Z}_{(\sum_{j=0}^{k-1} M_j+1)d}^\top\boldsymbol{\beta}_{(\sum_{j=0}^{k-1} M_j+1)} \\
& + \cdots + (A_{i(\sum_{j=0}^{k} M_j)d} - \pi_{(\sum_{j=0}^{k} M_j)d})\boldsymbol{Z}_{(\sum_{j=0}^{k} M_j)d}^\top\boldsymbol{\beta}_{(\sum_{j=0}^{k} M_j)} \\
& + \epsilon_{id}
\end{aligned}$$

$\boldsymbol{B}_d^\top\boldsymbol{\alpha}$ is a function of d and covariates that are unaffected by intervention categories. For example, we can define a $(q-1)$th-order function of d for $B_d^\top\boldsymbol{\alpha}$ with parameters $\boldsymbol{\alpha}^\top = (\alpha_1, \ldots, \alpha_q)$ and $\boldsymbol{B}_d^\top = (1, d-1, \ldots, (d-1)^{q-1})$. $\boldsymbol{\beta}_m^\top = (\beta_{m1}, \ldots, \beta_{mp_m})$ is the parameter vector of interest for the mth intervention category. ϵ_{id} is the error term, and $\boldsymbol{\epsilon}_i{=}(\epsilon_{i1}, \ldots, \epsilon_{iD})^\top$ is assumed to follow a multivariate normal distribution with zero mean, variance σ^2 and correlation coefficient ρ, for $i = 1, \ldots, N$. Note that the parameter ρ can take nonzero values, meaning that observations from the same participants are not necessarily independent. For each of the proximal effects $\boldsymbol{Z}_{md}^\top\boldsymbol{\beta}_m$, $m = 1, \ldots, M$, the intervention indicator A_{imd} is centred by the randomization probability π_{md}, as by Liao et al.[10] and Boruvka et al.[23] This centering procedure gives zero expected value for all $(A_{imd} - \pi_{md})\boldsymbol{Z}_{md}^\top\boldsymbol{\beta}_m$, $m = 1, \ldots, M$, and makes it easier to interpret the term $\boldsymbol{B}_d^\top\boldsymbol{\alpha}$ of the working model, that is, $\boldsymbol{B}_d^\top\boldsymbol{\alpha} = E(Y_{id} \mid I_{id} = 1)$, where I_{id} is the availability indicator for participant i on day d with expectation $E(I_{id}) = \tau_d$.[24]

We define the model parameter for $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iD})^\top, i = 1, \ldots, N$ by $\boldsymbol{\theta}{=}(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, where $\boldsymbol{\beta}^\top{=}(\boldsymbol{\beta}_1^\top,\ldots,\boldsymbol{\beta}_M^\top)$. Following Liao et al.[10] and Boruvka et al.,[23] we derive the least squares (LS) estimator $\hat{\boldsymbol{\theta}}$ obtained by minimizing the squared error criterion

$$SEC(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} I_{id}\left[Y_{id} - \boldsymbol{X}_{id}^\top\boldsymbol{\theta}\right]^2 \tag{3}$$

where N is the sample size and $\boldsymbol{X}$ is the design matrix for the regression model of $\boldsymbol{Y}$. The LS estimator can be obtained by solving the equation

$$\frac{\partial}{\partial\boldsymbol{\theta}}SEC(\boldsymbol{\theta}) = \frac{-2}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} I_{id}\left[Y_{id} - \boldsymbol{X}_{id}^\top\boldsymbol{\theta}\right]\boldsymbol{X}_{id} = 0$$

Therefore, the LS estimator is

$$\hat{\boldsymbol{\theta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{D} I_{id} \boldsymbol{X}_{id} \boldsymbol{X}_{id}^{\top} \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{D} I_{id} Y_{id} \boldsymbol{X}_{id} \tag{4}$$

that is, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^{\top}, \hat{\boldsymbol{\beta}}^{\top})^{\top}$ with dimension $(q + \sum_{m=1}^{M} p_m) \times 1$, where $\hat{\boldsymbol{\alpha}}$ is a vector that includes the first q elements of $\hat{\boldsymbol{\theta}}$ while $\hat{\boldsymbol{\beta}}$ includes the rest of the $\sum_{m=1}^{M} p_m$ elements of $\hat{\boldsymbol{\theta}}$.

2.4 Test statistics

This section proposes the hypothesis tests and derives the test statistics for the model parameters estimated in Section 2.3. We propose a null hypothesis that the proximal effect sizes of all intervention categories are zeros. In other words, the intervention components do not have any effects on the proximal outcome, that is,

$$H_0 : b_m(d; \boldsymbol{\beta}_m) = 0, \text{for all } m \text{ and } d, \text{ or } \boldsymbol{\beta} = \mathbf{0}$$

where $\mathbf{0}$ represents a zero vector with $\sum_{m=1}^{M} p_m$ elements. An alternative hypothesis is that the proximal effect sizes of intervention categories are not all zeros. In other words, the intervention components have some effects on the proximal outcome, that is,

$$H_1 : b_m(d; \boldsymbol{\beta}_m) \neq 0, \text{for some } m \text{ and } d, \text{ or } \boldsymbol{\beta} \neq \mathbf{0}$$

where m indicates the m^{th} message category, that is, $m = 1, \ldots, M_0, \ldots, M_0 + M_1, \ldots, \sum_{j=0}^{k} M_j = M$, $\boldsymbol{\beta}_m = (\beta_{m1}, \ldots, \beta_{mp_m})^{\top}$; d is the day during the study period, that is, $d = 1, \ldots, D$; and $\boldsymbol{Z}_{md}$ is defined in equation (1). We have the control category at $m = 0$. Note that the above hypothesis test is conducted over the entire study period, not at each decision time point separately.

In order to derive the test statistic distributions, the following assumptions are required.

1. Let $\boldsymbol{\Theta}$ be the parameter space for $\boldsymbol{\theta}$, where $\boldsymbol{\Theta}$ is a compact subset of $R^{q+\sum_{m=1}^{M} p_m}$.
2. $E(SEC(\boldsymbol{\theta}))$ exists and has a unique minimum value at $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$.
3. $SEC(\boldsymbol{\theta})$ is continuous, bounded, and differentiable in the neighbourhood of $\tilde{\boldsymbol{\theta}}$.
4. The matrix $\sum_{d=1}^{D} E(I_{id} \boldsymbol{X}_{id} \boldsymbol{X}_{id}^{\top})$ in equation (5) is invertible.

First, we present the following lemma about consistency and asymptotic normality for the LS estimator $\hat{\boldsymbol{\theta}}$. The proof is deferred to the Supplemental Material.

Lemma 2.1. *The LS estimator $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\tilde{\boldsymbol{\theta}}$. Under standard moment conditions and Assumptions 1 to 4, we have $\sqrt{N}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \rightarrow Normal(0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$.*

When the sample size N is large, the sample mean within the estimator (4) is replaced by its expectation, that is,

$$\tilde{\boldsymbol{\theta}} = \left[\sum_{d=1}^{D} E\left(I_{id} \boldsymbol{X}_{id} \boldsymbol{X}_{id}^{\top} \right) \right]^{-1} \sum_{d=1}^{D} E\left(I_{id} Y_{id} \boldsymbol{X}_{id} \right) \tag{5}$$

that is, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}^{\top}, \tilde{\boldsymbol{\beta}}^{\top})^{\top}$. In other words, it turns out that $\hat{\boldsymbol{\theta}} \rightarrow \tilde{\boldsymbol{\theta}}$ when $N \rightarrow \infty$. More details about equation (5) are covered in the Supplemental Material. The asymptotic covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is defined by

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left[\sum_{d=1}^{D} E(I_{id} \boldsymbol{X}_{id} \boldsymbol{X}_{id}^{\top}) \right]^{-1} E\left(\sum_{d=1}^{D} I_{id} \tilde{\epsilon}_{id} \boldsymbol{X}_{id} \sum_{d=1}^{D} I_{id} \tilde{\epsilon}_{id} \boldsymbol{X}_{id}^{\top} \right) \left[\sum_{d=1}^{D} E(I_{id} \boldsymbol{X}_{id} \boldsymbol{X}_{id}^{\top}) \right]^{-1} \tag{6}$$

where $E(\sum_{d=1}^{D} I_{id} \tilde{\epsilon}_{id} \boldsymbol{X}_{id} \sum_{d=1}^{D} I_{id} \tilde{\epsilon}_{id} \boldsymbol{X}_{id}^{\top})$ is defined in the Supplemental Material. Thus the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ converges to normal, that is, $\sqrt{N}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) \rightarrow \text{Normal}(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$; see the Supplemental Material for its derivation. The asymptotic covariance matrix can be expressed as $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{Q}^{-1} \boldsymbol{W} \boldsymbol{Q}^{-1}$ with square matrices $\boldsymbol{Q}$ and $\boldsymbol{W}$ that are the lower right $\sum_{m=1}^{M} p_m \times \sum_{m=1}^{M} p_m$ blocks of $\sum_{d=1}^{D} E(I_{id} \boldsymbol{X}_{id} \boldsymbol{X}_{id}^{\top})$ and $E(\sum_{d=1}^{D} I_{id} \tilde{\epsilon}_{id} \boldsymbol{X}_{id} \sum_{d=1}^{D} I_{id} \tilde{\epsilon}_{id} \boldsymbol{X}_{id}^{\top})$, respectively.

As $\hat{\boldsymbol{\beta}}$ follows a normal distribution when the null hypothesis is true and N is large, in a similar fashion as by Tu et al.,[25] the test statistic function $C_N(\cdot)$, that is, $C_N(\hat{\boldsymbol{\beta}}) = N \hat{\boldsymbol{\beta}}^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \hat{\boldsymbol{\beta}}$ follows a χ^2 distribution with a degrees of freedom (df)

of $\sum_{m=1}^{M} p_m$ (i.e. the length of $\boldsymbol{\beta}$). For example, assuming a quadratic trend (i.e. $p_m = 3$ for all $m = 1, \ldots, M$), the df of $N\hat{\boldsymbol{\beta}}^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{\beta}}$ for the motivation component ($M = 3$) of the DIAMANTE study is 9. Unlike the Wald statistic, the chi-square statistic allows the hypothesis to involve not only a single parameter but also a vector parameter. We have $\boldsymbol{\delta} = \boldsymbol{\beta}/\bar{\sigma}$, where $\bar{\sigma}^2$ is the average of the residual variance over all the decision time points, that is, $\bar{\sigma}^2 = \sum_{d=1}^{D} \mathrm{Var}(\epsilon_{id})/D$, therefore, $C_N(\hat{\boldsymbol{\beta}}) = C_N(\hat{\boldsymbol{\delta}}) = N\hat{\boldsymbol{\delta}}^{\top}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}/\bar{\sigma}^2)^{-1}\hat{\boldsymbol{\delta}}$. Tu et al. proposed a sample size calculation method based on power under the GEE approach for longitudinal data. Leveraging on this, we define the power function as follows. If H_0: $\boldsymbol{\beta} = 0$ is true, then the type-I error rate is defined by

$$\Pr\left(X_{\sum_{m=1}^{M} p_m} > \chi^2_{\sum_{m=1}^{M} p_m,\alpha}\right) = \alpha \tag{7}$$

where $X_{\sum_{m=1}^{M} p_m}$ presents a random variable following a central chi-squared distribution with df $\sum_{m=1}^{M} p_m$ while $\chi^2_{\sum_{m=1}^{M} p_m,\alpha}$ represents its $1-\alpha$ quantile. We reject H_0 at level α if $X_{\sum_{m=1}^{M} p_m} > \chi^2_{\sum_{m=1}^{M} p_m,\alpha}$. If H_1: $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \neq 0$ is true, then

$$\Pr\left(X_{\sum_{m=1}^{M} p_m, C_N(\tilde{\delta})} > \chi^2_{\sum_{m=1}^{M} p_m,\alpha}\right) = \text{Power} \tag{8}$$

where $X_{\sum_{m=1}^{M} p_m, C_N(\tilde{\delta})}$ represents a random variable following a chi-squared distribution with df $\sum_{m=1}^{M} p_m$. Note that[25] does not define the distribution of $C_N(\hat{\boldsymbol{\delta}})$ for a small sample size.

When N is small, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is replaced by its sample estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$, which is derived by Mancl and DeRouen,[26] and the test statistic follows Hotelling's T^2 distribution; see, for example, Hotelling[27] and Li and Redden.[28] We define the small-sample estimator by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$=$\hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}^{-1}$. Let

$$\hat{e}_{id} = Y_{id} - \boldsymbol{X}_{id}^{\top}\hat{\boldsymbol{\theta}} \tag{9}$$

$$\hat{\boldsymbol{e}}_i^{\top} = (\hat{e}_{i1}, \ldots, \hat{e}_{iD}) \tag{10}$$

$$\boldsymbol{X}_i^{\top} = \begin{bmatrix} \boldsymbol{X}_{i1}^{\top} I_{i1} \\ \vdots \\ \boldsymbol{X}_{iD}^{\top} I_{iD} \end{bmatrix}_{D\times(q+\sum_{m=1}^{M} p_m)} \tag{11}$$

$$\boldsymbol{H}_i = \boldsymbol{X}_i^{\top}[\sum_{i=1}^{N} \boldsymbol{X}_i\boldsymbol{X}_i^{\top}]^{-1}\boldsymbol{X}_i \tag{12}$$

The matrix $\hat{\boldsymbol{Q}}^{-1}$ is given by the lower right $\sum_{m=1}^{M} p_m \times \sum_{m=1}^{M} p_m$ block of $[\sum_{i=1}^{N} \boldsymbol{X}_i\boldsymbol{X}_i^{\top}/N]^{-1}$; the matrix $\hat{\boldsymbol{W}}$ is given by the lower right $\sum_{m=1}^{M} p_m \times \sum_{m=1}^{M} p_m$ block of $[\sum_{i=1}^{N} \boldsymbol{X}_i(\boldsymbol{I}_{D\times D} - \boldsymbol{H}_i)^{-1}\hat{\boldsymbol{e}}_i\hat{\boldsymbol{e}}_i^{\top}(\boldsymbol{I}_{D\times D} - \boldsymbol{H}_i)^{-1}\boldsymbol{X}_i^{\top}]/N$, where $\boldsymbol{I}_{D\times D}$ is the identity matrix with dimension $D \times D$.

Liao et al.[10] suggested that the test statistic follows a Hotelling's T^2 distribution with dimension $\sum_{m=1}^{M} p_m$ and df $N-q-1$, that is,

$$\hat{C}_N(\hat{\boldsymbol{\delta}}) = N\hat{\boldsymbol{\beta}}^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{\beta}} \sim T^2_{\sum_{m=1}^{M} p_m, N-q-1} = \frac{\sum_{m=1}^{M} p_m(N-q-1)}{N-q-\sum_{m=1}^{M} p_m} F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m}$$

Thus

$$\frac{N-q-\sum_{m=1}^{M} p_m}{\sum_{m=1}^{M} p_m(N-q-1)} \hat{C}_N(\hat{\boldsymbol{\delta}}) \sim F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m}$$

if H_0: $\boldsymbol{\beta} = 0$ is true, and the type-I error rate (α) is defined by

$$\Pr\left(F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m} > f_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m,\alpha}\right) = \alpha \tag{13}$$

where $F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m}$ represents a random variable following a central F distribution with df $\sum_{m=1}^{M} p_m$ and $N - q - \sum_{m=1}^{M} p_m$ while $f_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m,\alpha}$ represents its $1-\alpha$ quantile. We reject H_0 at level α if $F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m} >$

$f_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m, \alpha}$. If H_1: $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \neq 0$ is true, then

$$\frac{N-q-\sum_{m=1}^{M} p_m}{\sum_{m=1}^{M} p_m (N-q-1)} \hat{C}_N(\hat{\boldsymbol{\delta}}) \sim F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m, \tilde{C}_N(\tilde{\boldsymbol{\delta}})}$$

that is, a non-central F distribution with non-centrality parameter $\tilde{C}_N(\tilde{\boldsymbol{\delta}}) = N\tilde{\boldsymbol{\delta}}^{\top}(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}/\bar{\sigma}^2)^{-1}\tilde{\boldsymbol{\delta}}$, where $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ is the sample estimate of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ when N is large. Therefore

$$\Pr\left(F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m, \tilde{C}_N(\tilde{\boldsymbol{\delta}})} > f_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m, \alpha}\right) = \text{Power} \tag{14}$$

where $F_{\sum_{m=1}^{M} p_m, N-q-\sum_{m=1}^{M} p_m, \tilde{C}_N(\tilde{\boldsymbol{\delta}})}$ represents a random variable following a non-central F distribution with df $\sum_{m=1}^{M} p_m$ and $N-q-\sum_{m=1}^{M} p_m$ and non-centrality parameter $\tilde{C}_N(\tilde{\boldsymbol{\delta}})$. Note that Liao et al.[10] did not provide any mathematical proofs for the distribution of the test statistic.

In this article, we further suggest an alternative distribution for $\hat{C}_N(\hat{\boldsymbol{\delta}})$ (see Corollary 2.1.1) and provide certain mathematical derivations in the Supplemental Material.

Corollary 2.1.1. *According to Lemma* 2.1*, under a finite sample, the test statistic* $\hat{C}_N(\hat{\boldsymbol{\delta}}) = N\hat{\boldsymbol{\beta}}^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{\beta}}$ *follows a Hotelling's* $T^2_{\sum_{m=1}^{M} p_m, N}$ *distribution.*

This distribution can be defined by

$$\hat{C}_N(\hat{\boldsymbol{\delta}}) = N\hat{\boldsymbol{\beta}}^{\top}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1}\hat{\boldsymbol{\beta}} \sim T^2_{\sum_{m=1}^{M} p_m, N} = \frac{\sum_{m=1}^{M} p_m(N)}{N-\sum_{m=1}^{M} p_m+1} F_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1}$$

under H_0. The corresponding type-I error rate (α) can be defined by

$$\Pr\left(F_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1} > f_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1, \alpha}\right) = \alpha \tag{15}$$

We reject H_0 at level α if

$$F_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1} > f_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1, \alpha}$$

In additon if H_1: $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} \neq 0$ is true, then the power function can be defined by

$$\Pr\left(F_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1, \tilde{C}_N(\tilde{\boldsymbol{\delta}})} > f_{\sum_{m=1}^{M} p_m, N-\sum_{m=1}^{M} p_m+1, \alpha}\right) = \text{Power} \tag{16}$$

Simulation studies show that the proposed test statistic provides better power and coverage probability estimates than the test statistic of Liao et al.[10] when the sample size is small.

2.5 Sample size calculation

Here, we propose a power-based sample size formula for FlexiMRT. This method requires a working knowledge of the standardized proximal effect sizes of intervention categories (i.e. $b_m(d; \boldsymbol{\delta}_m) = \boldsymbol{Z}_{md}^{\top}\boldsymbol{\delta}_m$, where $\boldsymbol{\delta}_m = \boldsymbol{\beta}_m/\bar{\sigma}$, for $m = 1, \ldots, M$) as input parameters. Thus, given the desired power and the intended minimum detectable standardized $\boldsymbol{\beta}$ ($\boldsymbol{\delta}$), the sample size N can be computed by solving for either equations (8), (14) or (16), depending on the choice of the test statistic. For example, assuming a linear-plateau trend, $\boldsymbol{\delta}$ can be derived by the standardized initial $b_m(d_j; \boldsymbol{\delta}_m)$ and average $\frac{1}{D-(d_j-1)}\sum_{d=d_j}^{D} b_m(d; \boldsymbol{\delta}_m)$ proximal effect sizes and the turning point day d^m_{turn}, where the m-th category is added on day d_j, for $m = 1, \ldots, M$ and $j = 0, 1, \ldots, k$. The calculated N is, therefore, the minimum integer that gives the estimated power not lower than its nominal value.

In Figure 2, we illustrate the operating characteristics of the required sample size as a function of the design parameters, assuming constant trend proximal effect and using the Hotelling's T^2 test statistic (equation (14)). In general, N increases with the number of intervention categories M in a particular component, but decreases with the number of decision time points and study days. Although the difference is slight, N decreases when more categories are added later than earlier, that is, fewer M_0 and more M_1 categories, for a fixed number of total categories (e.g. $M = M_0 + M_1 = 4$).

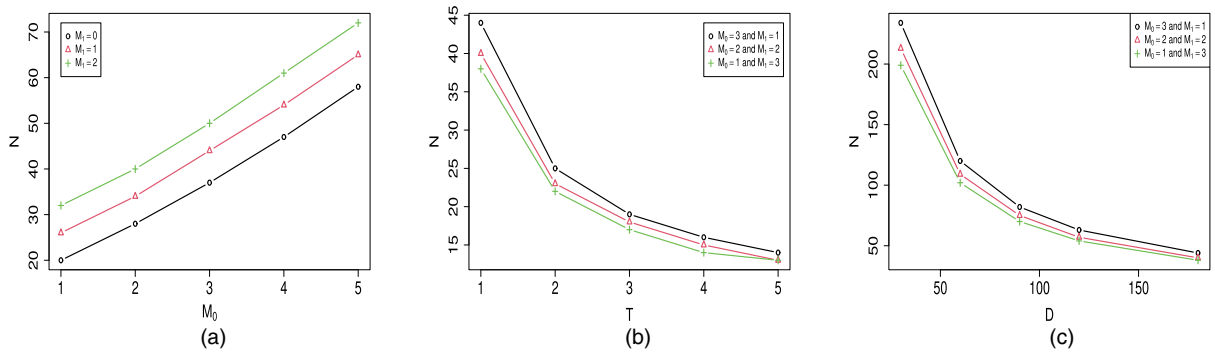

Figure 2. Sample size N versus: (a) initial intervention categories M_0 given different number of categories M_1 added halfway through the study; (b) number of decision time points T per day and (c) study days D given $M = 4$ and different M_0 and M_1 allocations.

Figure 2(a) also illustrates the benefit (i.e. avoiding the unnecessary increases in both the sample size and the study duration) of allowing additional categories to be added later in the trial. Take the StayWell study as an example (contact the authors of Figueroa et al.[14] for relevant details), where two consecutive MRTs were conducted, the first with two categories and the latter with an additional category (three categories in total) at the beginning of each trial. Assuming all other design parameters are specified in Figure 2(a), the total sample size needed for the two trials is $N = 65$. In contrast, a FlexiMRT design would have suggested using the $M_0 = 2$ and $M_1 = 1$ variation, which only gives a sample size of $N = 34$.

Note the power-based method should only be used if the goal is to perform a hypothesis test, as is typical in a confirmatory study. However, due to the novelty and recency of MRT, prior information on the proximal effects of the individual categories are often limited. A pilot MRT to assess the feasibility and acceptability of the intervention categories may be necessary before conducting the full-scale study, similar to pilot studies for sequential multiple assignment randomized trials (SMARTs).[29] To operationalize the sample size calculation for pilot MRT, we suggest the sample size to be calculated based on certain desired level of precision for the $\boldsymbol{\beta}$ estimates; see the Supplemental Material for the derivation, Kelley et al.[30] or Maxwell et al.[31] for discussions in the classical settings, and Yan et al.[32] for the rationale and benefit in the context of pilot SMARTs.

3 DIAMANTE study with student population example

A smaller study[22] was conducted where 93 students from the University of California, Berkeley used the DIAMANTE app (66 and 27 participants in the UR and AL groups, respectively) for 45 days. A total of 44 decision time points and proximal outcome measures (corresponding to 45 days) were collected from each participant. Here, we use the results from the UR group to demonstrate the power-based sample size calculator. The precision-based demonstration appears in the Supplemental Material.

To calculate the required sample size, we first estimate the proximal effects of the intervention categories using only the complete cases in the dataset and we assume $\tau = 100\%$ availability as reasoned in Section 2.2. For each participant, we delete the days when the messages were not sent due to technical errors or when the outcome measures were not collected due to non-response. The technical missingness can be dealt with using imputation[33] if desired.

Here, we give an example using the Motivational Message component that has three intervention categories ($M = 3$), 'benefit', 'self-efficacy' and 'opportunity', all proposed at the beginning of the trial. The randomization probability ($\boldsymbol{\pi}$) for each category (including the control category) was 0.25. According to the working model of Y_{id} described in Section 2.3, we consider the constant ($q = p_m = 1$) , linear ($q = p_m = 2$) and quadratic ($q = p_m = 3$) trends for the intervention categories, where $m = 1, 2, 3$. The regression coefficients ($\boldsymbol{\beta}$) can be estimated by equation (4). We denote the initial and average proximal effect sizes by $\boldsymbol{\beta}^0$ and $\bar{\boldsymbol{\beta}}^d$, respectively, where $\boldsymbol{\beta}^d$ is defined by equation (1). The corresponding standardized effect sizes are denoted by $\boldsymbol{\delta}^0$ ($\boldsymbol{\beta}^0/\bar{\sigma}$) and $\bar{\boldsymbol{\delta}}^d$ ($\bar{\boldsymbol{\beta}}^d/\bar{\sigma}$), respectively, where $\bar{\sigma}^2$ is the average of the residual variance over all the decision time points.

Under the constant trend assumption, the initial and average proximal effect sizes are the same, because the effect sizes are marginalized over the study days and the historical variables are not considered in the model. We have $\boldsymbol{\beta} = (357, 589, 526)^\top = \boldsymbol{\beta}^0 = \bar{\boldsymbol{\beta}}^d$ and $\bar{\sigma} = 4869$ or $\boldsymbol{\delta}^0 = \bar{\boldsymbol{\delta}}^d = (0.073, 0.121, 0.108)^\top$. Assuming 100% availability, 44 decision time points, type-I error rate $\alpha = 5\%$, 80% power, and using the more conservative Hotelling's T^2 distributed test statistic

Table 1. The sample sizes calculation based on the data analysis results of the Diabetes and Mental Health Adaptive Notification Tracking and Evaluation (DIAMANTE) study with the university students.

τ	Trend	Category	β^0	$\bar{\beta}^d$	$\bar{\sigma}$	δ^0	$\bar{\delta}^d$	N
100%	Constant	Benefit	357	357	4869	0.073	0.073	117
		Self-efficacy	589	589		0.121	0.121	
		Opportunity	526	526		0.108	0.108	
	Linear	Benefit	609	338	4867	0.125	0.069	116
		Self-efficacy	441	598		0.091	0.123	
		Opportunity	869	512		0.178	0.105	
	Quadratic	Benefit	662	378	4866	0.136	0.078	101
		Self-efficacy	718	621		0.148	0.128	
		Opportunity	1394	530		0.287	0.109	
	Constant	Pooled	494	494	4870	0.101	0.101	72
	Constant	Benefit	357	357	4869	0.073	0.073	163
		Self-efficacy	589	589		0.121	0.121	
		Opportunity	526	526		0.108	0.108	
		Message 4	300	300		0.062	0.062	
		Message 5	300	300		0.062	0.062	
70%	Constant	Benefit	357	357	4869	0.073	0.073	230
		Self-efficacy	589	589		0.121	0.121	
		Opportunity	526	526		0.108	0.108	
50%	Constant	Benefit	357	357	4869	0.073	0.073	319
		Self-efficacy	589	589		0.121	0.121	
		Opportunity	526	526		0.108	0.108	

in equation (14), the sample size required to detect a three-category component with average standardized proximal effect sizes 0.073, 0.121 and 0.108 is 117. This calculation approach was mentioned by Figureroa et al.[22]

Assuming linear and quadratic trends, we have the initial proximal effect size vectors $(609, 441, 869)^\top$ and $(662, 718, 1394)^\top$, the average proximal effect size vectors $(338, 598, 512)^\top$ and $(378, 621, 530)^\top$, $\bar{\sigma} = 4867$ and 4866, respectively. The corresponding standardized initial proximal effect size vectors are $(0.125, 0.091, 0.178)^\top$ and $(0.136, 0.148, 0.287)^\top$, while the corresponding standardized average proximal effect size vectors are $(0.069, 0.123, 0.105)^\top$ and $(0.078, 0.128, 0.109)^\top$, respectively. We observed the quadratic trend, where the shape of the proximal effect size of the 'benefit' category is concave down with local maximum at the 36-th decision time point while both the 'self efficacy' and 'opportunity' categories are concave up with local minimum at the 17-th and 26-th decision time points, respectively. Therefore, the calculated sample sizes corresponding to the linear and quadratic trends are 116 and 101, respectively. The required sample size under the quadratic trend turns out to be the smallest because the estimated initial and average proximal effect sizes using the quadratic trend are larger than the effect sizes using the constant and linear trends, given the same length of study periods.

Suppose it is reasonable to assume that the effects of the intervention categories are roughly equal, or that the interest lies in the 'pooled' effect of all the intervention categories. We may use the conventional two-category approach,[10] where the components are collapsed into for example, no message versus a 'pooled' intervention message. The estimated standardized average proximal effect size for the Motivational Message component will be 0.101, under a constant trend, and the corresponding sample size required will be 72. However, the equal intervention category effects assumption may not hold and the interest may be in identifying the individual effect sizes. Then the multi-category approach as demonstrated earlier is more useful.

Suppose the smaller study considered adding more intervention categories later in the study. For example, suppose two additional categories are added to the Motivational Message component at halfway (i.e. the 23rd decision point), based on the feedback received from the participants since the trial started. Suppose we use the same proximal effect sizes for the first three categories and $\bar{\sigma}$ as estimated above, and assume the proximal effect sizes of both added categories follow constant trends with average value 300 and 100% availability, the Hotelling's T^2 distributed test statistic calculates a sample size of 163, for $\alpha = 5\%$ and 80% power. Suppose the availability is not always 100%; for example, some participants may turn off the activity notification on some particular days. The calculated sample size increases to 230 and 319 for 70% and 50% expected availabilities, respectively.

The data analysis and the sample size calculation results described above are summarized in Table 1.

Table 2. The N and power estimates when the working model assumptions are correct.

		Average standardized proximal effect size					
		0.10	0.06	0.10	0.06	0.10	0.06
Availability	Test statistics	N		Formulated power		Monte Carlo power	
100%	$\chi^2_{\sum_m p_m}$	46	127	0.81	0.80	0.81	0.81
	Hotelling's $T^2_{\sum_m p_m, N}$	54	135	0.81	0.80	0.81	0.81
	Hotelling's $T^2_{\sum_m p_m, N-q-1}$	54	135	0.80	0.80	0.77	0.78
70%	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.81	0.78
	Hotelling's $T^2_{\sum_m p_m, N}$	73	190	0.80	0.80	0.78	0.80
	Hotelling's $T^2_{\sum_m p_m, N-q-1}$	73	190	0.80	0.80	0.80	0.81

4 Simulation study

In this section, we present simulation studies to investigate the performance of the proposed power-based sample size (N) formulas. Let the study period be $D = 180$ and the number of decision time points per day be $T = 1$. Let there be a control category, and $M = 4$ intervention categories, where $M_0 = 3$ categories are added at the beginning, and $M_1 = 1$ category is added halfway through the study at $d_1 = 91$. We calculate N under the correctly specified or some mis-specific models for each simulation study, with nominal power $P = 80\%$ and type-I error rate $\alpha = 5\%$, and generate 1000 Monte Carlo (MC) data sets. The performance of each N formula is measured by comparing the difference between the formulated and MC power estimates. The data generation steps are listed asfollows:.

Step 1. The availability indicators I_{id} follow the Bernoulli distributions, that is, $I_{id} \sim \text{Bernoulli}(\tau_d)$ for each participant i and on each day d. We set the availability rate τ_d at 100% and 70%.

Step 2. The intervention categories $\boldsymbol{A}_{id} = (A_{i1d}, \ldots, A_{i(\sum_{j=0}^{k} M_j)d})^\top$ follow the multinomial distributions, that is, $\boldsymbol{A}_{id} \sim \text{Multinomial}(1 - \sum_m \pi_{md}, \pi_{1d}, \ldots, \pi_{(\sum_{j=0}^{k} M_j)d})$, where π_{md} and $1 - \Sigma_m \pi_{md}$ are the randomization probabilities of the intervention category-m and the control category on day d, respectively. We set the initial randomization probability as $\pi_{md} = 0.25$, and $\pi_{md} = 0.2$ after d_1.

Step 3. The error terms $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iD})^\top$ follow the multivariate normal distributions, that is, $\boldsymbol{\epsilon}_i \sim \text{MVN}(\mathbf{0}_{D\times 1}, \text{COV}(\boldsymbol{\epsilon}_i))$, where $\text{COV}(\boldsymbol{\epsilon}_i)$ is the $D \times D$-dimensional covariance matrix with diagonal entries σ^2 and off-diagonal entries $\rho\sigma^2$. We set $\sigma = 1$ and $\rho = 0$.

Step 4. The proximal outcome is computed by $Y_{id} = \boldsymbol{X}_{id}^\top \boldsymbol{\theta} + \epsilon_{id}$, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ and $\boldsymbol{X}_{id}^\top = [\boldsymbol{B}_d^\top, (A_{i1d} - \pi_{1d})\boldsymbol{Z}_{1d}^\top, \ldots, (A_{i4d} - \pi_{4d})\boldsymbol{Z}_{4d}^\top]$. We set $\boldsymbol{B}_d = (1, \min[28 - 1, d - 1])$, and consider a linear-plateau trend for the standardized proximal effect size for each intervention category with turning point on the 28-th day, that is, $d^m_{\text{turn}} = 28$ for m=1, 2 and 3 and $d^m_{\text{turn}} = 118$ for $m = 4$.

The standardized proximal effect size of the m-th category satisfies $\delta_m(d, \boldsymbol{\delta}_m) = \boldsymbol{Z}_{md}^\top \boldsymbol{\delta}_m$, where $\boldsymbol{\delta}_m = \boldsymbol{\beta}_m / \bar{\sigma}$. We define the initial standardized proximal effect sizes to be 0.001 and the average standardized proximal effect sizes to be 0.1 and 0.06.

4.1 Correctly specified working model

Table 2 gives the calculated N, and the corresponding formulated and MC power estimates. We observe that the power estimates are fairly similar. In general, when the working model is correctly specified, the proposed sample size calculator provides accurate power.

4.2 Mis-specified trend for proximal effect

Next, we consider the robustness of the sample size calculator when the model is mis-specified. We observe that the formulated power estimates are very close to the nominal power, but the MC power estimates are lower. This is because the mis-specified proximal trends give smaller Ns in contrast to the Ns in Table 2. Comparing Tables 2 and 3, mis-specifying

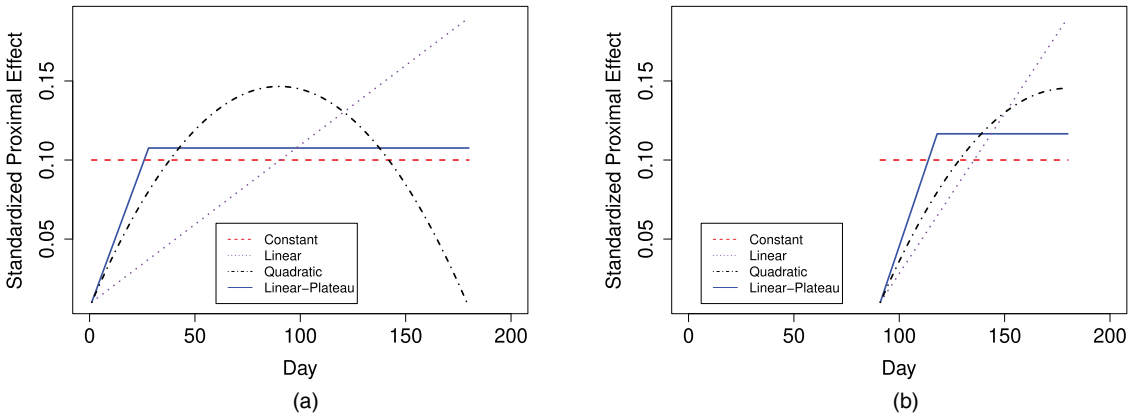

Figure 3. The plots of standardized proximial effect size over the course of the study period: (a) intervention category proposed at the beginning; (b) intervention category proposed on the half-way.

as the constant trend results in the largest difference in N, mainly due to the shape differences (see Figure 3). In contrast, the linear and quadratic trends give similar Ns.

4.3 Mis-specified number of intervention categories

Suppose we mis-specified M_1 such that no categories are added later in the study ($M_1 = 0$) and underestimated the number of initial categories $M_0 = 1$ in Table 4. The formulated power estimates are very close to the nominal power, but because the Ns are calculated assuming fewer categories, the MC power estimates are much lower than 80%. The MC power estimates (i.e. around 30%–40%) performed the worst when $M_0 = 1$ and $M_1 = 0$, as they are the furthest away from the true values (i.e. $M_0 = 3$ and $M_1 = 1$). Note that when $M_1 = 0$, the *design structure* essentially reduces to a conventional MRT. However, the sample size calculation under a conventional MRT approach will always assume $M_0 = 1$ by 'pooling' the intervention categories into one proximal effect (Liao et al.,[10]), that is, the sample sizes for $M_0 = 3$ and $M_1 = 0$ is the same as that for $M_0 = 1$ and $M_1 = 0$. The MC power estimates under $M_0 = 3$ and $M_1 = 0$ are around 70% while the MC power estimates under $M_0 = 4$ and $M_1 = 0$ are around 85%. However, in order to detect the proximal effect sizes of the multiple intervention categories that are not all added initially, the proposed method provides a better solution of estimating powers than the Liao's method.[10]

4.4 Mis-specified error term distribution

We calculate the Ns and the corresponding powers assuming that the error terms are generated from the multivariate normal distribution, that is, $\boldsymbol{\epsilon}_i \sim \text{MVN}(\mathbf{0}_{D\times 1}, \text{COV}(\boldsymbol{\epsilon}_i))$, where a constant variance σ^2 is used in the $D \times D$-dimensional covariance matrix. However, for the data generation, we let the true variance of ϵ_{id} to be σ_d^2, for d=1, … , D, which varies over d. For example, σ_d^2 is linearly increasing (i.e. $\sigma_d^2 = 0.9^2 + 0.0021 \times (d-1)$) and decreasing (i.e. $\sigma_d^2 = 1.19^2 - 0.0021 \times (d-1)$), where $\bar{\sigma}^2 = \frac{1}{D}\sum_{d=1}^{D} \sigma_d^2 = 1$. We observe that both the formulated and MC power estimates remain close to the nominal power in Table 5, meaning our sample size calculator is still accurate when the outcome variances are not constant over the study.

Suppose the error term follows a multivariate normal distribution with standardized normal for marginal distribution and non-zero correlation coefficients, that is, $\boldsymbol{\epsilon}_i = (\epsilon_{i1},\ldots,\epsilon_{iD})^\top$ follows a normal AR(1) process, for example, $\epsilon_{id} = \phi\epsilon_{i,d-1} + \nu_{id}$, where $\phi = 0.5$ and -0.5 and ν_d are i.i.d Normal(0, 0.75). When calculating N assuming no autocorrelation, we observe that both the formulated and MC power estimates are very close to the nominal power in Table 6. Therefore, the required N seems robust against the autocorrelation structure of the outcome measures over time.

4.5 Mis-specified trend of availability

It is common to calculate N under a simple assumption of constant trend for availability. However, the availability trend may not actually be constant in real data. Here, we calculate N assuming a constant trend of 70% availability ($\tau_d = 0.7$). The datasets are however generated based on linearly increasing (i.e. $\tau_d = 0.5 + 0.0022 \times (d-1)$) and decreasing (i.e. $\tau_d = 0.9 - 0.0022 \times (d-1)$) availability trends, for $d = 1, \ldots, D$, where we have $\bar{\tau} = \frac{1}{D}\sum_{d=1}^{D} \tau_d = 0.7$. In Table 7, both

Table 3. The N and power estimates when the standardized proximal effect size trends are mis-specified.

			Average standardized proximal effect size					
			0.1	0.06	0.1	0.06	0.1	0.06
Trend	Availability	Test statistics	N		Formulated power		Monte Carlo power	
Constant	100%	$\chi^2_{\sum_m p_m}$	39	107	0.81	0.80	0.73	0.69
		Hotelling's $T^2_{\sum_m p_m,N}$	43	111	0.80	0.80	0.64	0.70
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	44	111	0.81	0.80	0.66	0.66
	70%	$\chi^2_{\sum_m p_m}$	55	152	0.80	0.80	0.70	0.68
		Hotelling's $T^2_{\sum_m p_m,N}$	60	157	0.81	0.80	0.69	0.71
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	60	157	0.80	0.80	0.64	0.69
Linear	100%	$\chi^2_{\sum_m p_m}$	41	116	0.81	0.80	0.76	0.74
		Hotelling's $T^2_{\sum_m p_m,N}$	49	124	0.81	0.80	0.77	0.77
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	49	124	0.80	0.80	0.73	0.76
	70%	$\chi^2_{\sum_m p_m}$	58	166	0.80	0.80	0.73	0.77
		Hotelling's $T^2_{\sum_m p_m,N}$	66	174	0.80	0.80	0.74	0.75
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	66	174	0.80	0.80	0.73	0.73
Quadratic	100%	$\chi^2_{\sum_m p_m}$	40	115	0.80	0.80	0.74	0.73
		Hotelling's $T^2_{\sum_m p_m,N}$	51	126	0.80	0.80	0.73	0.76
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	52	126	0.81	0.80	0.76	0.71
	70%	$\chi^2_{\sum_m p_m}$	57	165	0.80	0.80	0.74	0.78
		Hotelling's $T^2_{\sum_m p_m,N}$	68	175	0.80	0.80	0.74	0.76
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	69	175	0.81	0.80	0.77	0.74

the formulated and MC power estimates are close to the nominal power, showing that mis-specifying the availability rate trend over time has little impact on the power, and consequently the required N.

5 Discussion

In this article, we propose a novel FlexiMRT design wherein it is possible to allow categories of intervention components to be added not only at the beginning, but also later in the study, with the adding times and the numbers of categories to be added at these times being pre-determined. We derive the associated sample size calculation methods based on both power and precision of the proximal effect size estimates. For both the methods, the required sample sizes increase with the number of intervention categories, and decrease when either the study period is longer or the number of decision time points per day is larger. We also observe that the required sample size decreases when more categories are added later in the study than at the beginning, given a fixed total number of intervention categories and assuming constant proximal effect trends. The proposed methods give the MC estimates of power and coverage probability close to the corresponding nominal values, provided that the specified working model is 'not too far' from the true one. A sample size calculated by incorrectly using the conventional MRT sample size calculator that only considers two-category components[10] does not provide sufficient power to detect the proximal effects.

As mentioned, the DIAMANTE study is the primary motivation behind this article. We have shown how the proposed sample size calculation method can be applied for the UR group of the DIAMANTE student study. However, it is worth

Table 4. The N and power estimates when M_0 and M_1, and consequently M are mis-specified.

			Average standardized proximal effect size					
			0.1	0.06	0.1	0.06	0.1	0.06
M_0 and M_1	Availability	Test statistics	N		Formulated power		Monte Carlo power	
$M_0 = 1$ and $M_1 = 0$	100%	$\chi^2_{\sum_m p_m}$	21	58	0.81	0.81	0.42	0.39
		Hotelling's $T^2_{\sum_m p_m, N}$	24	61	0.81	0.81	0.30	0.34
		Hotelling's $T^2_{\sum_m p_m, N-q-1}$	24	61	0.80	0.80	0.26	0.36
	70%	$\chi^2_{\sum_m p_m}$	30	82	0.81	0.80	0.41	0.41
		Hotelling's $T^2_{\sum_m p_m, N}$	33	85	0.81	0.80	0.31	0.38
		Hotelling's $T^2_{\sum_m p_m, N-q-1}$	33	85	0.80	0.80	0.29	0.38
$M_0 = 3$ and $M_1 = 0$	100%	$\chi^2_{\sum_m p_m}$	39	109	0.80	0.80	0.71	0.73
		Hotelling's $T^2_{\sum_m p_m, N}$	46	115	0.81	0.80	0.68	0.71
		Hotelling's $T^2_{\sum_m p_m, N-q-1}$	46	115	0.81	0.80	0.68	0.70
	70%	$\chi^2_{\sum_m p_m}$	56	155	0.81	0.80	0.73	0.75
		Hotelling's $T^2_{\sum_m p_m, N}$	62	161	0.80	0.80	0.71	0.71
		Hotelling's $T^2_{\sum_m p_m, N-q-1}$	62	161	0.80	0.80	0.71	0.72
$M_0 = 4$ and $M_1 = 0$	100%	$\chi^2_{\sum_m p_m}$	50	140	0.80	0.80	0.84	0.85
		Hotelling's $T^2_{\sum_m p_m, N}$	58	148	0.80	0.80	0.84	0.85
		Hotelling's $T^2_{\sum_m p_m, N-q-1}$	59	148	0.81	0.80	0.82	0.86
	70%	$\chi^2_{\sum_m p_m}$	72	200	0.80	0.80	0.86	0.84
		Hotelling's $T^2_{\sum_m p_m, N}$	80	208	0.81	0.80	0.84	0.81
		Hotelling's $T^2_{\sum_m p_m, N-q-1}$	80	208	0.80	0.80	0.84	0.85

noting that the calculated sample size is aimed not to select the optimal category, but to detect whether at least one of the intervention categories is more effective than the control category.

We are currently collecting data through the main DIAMANTE study, and performing interim data analysis, which can be used to further improve the experimental design. When the data collection is complete, we can apply linear mixed model approach to investigate the between-person heterogeneity of the proximal effect of an intervention component. This is similar to the method of Qian et al.[34] Based on the results of Bidargaddi et al.,[35] we could investigate, for example, whether a message category is more effective at mid-day on weekends than other decision time points.

The proposed FlexiMRT design only allows a pre-specified number of new intervention categories to be added at specified time points during the study. However, no additional participants are recruited when new categories are added. These features distinguish FlexiMRT from the platform clinical trials,[15] where the number of treatment arms added and the corresponding times are not known a priori, and additional participants are recruited when more arms are added. The proposed method, therefore, requires adequate planning to pre-determine the number of intervention categories and when to add them. Extending the procedure to be able to add arbitrary new categories into the FlexiMRT design can be an important future direction. This design can be further extended using a decision-theoretic framework similar to Lee et al.[16] to investigate when to add or not add a message category based on the observed proximal outcomes. Alternatively, a message category can be dropped early if it is unlikely to have any effect. The study can be stopped if the efficacy of a category is

Table 5. The N and power estimates when σ_d^2 is not constant over d.

			Average standardized proximal effect size					
			0.1	0.06	0.1	0.06	0.1	0.06
Trend of σ_d^2	Availability	Test statistics	N		Formulated power		Monte Carlo power	
Increasing	100%	$\chi^2_{\sum_m p_m}$	46	127	0.81	0.80	0.80	0.80
		Hotelling's $T^2_{\sum_m p_m,N}$	54	135	0.81	0.80	0.78	0.77
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	54	135	0.80	0.80	0.77	0.80
	70%	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.78	0.80
		Hotelling's $T^2_{\sum_m p_m,N}$	73	190	0.80	0.80	0.78	0.78
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	73	190	0.80	0.80	0.78	0.78
Decreasing	100%	$\chi^2_{\sum_m p_m}$	46	127	0.81	0.80	0.81	0.80
		Hotelling's $T^2_{\sum_m p_m,N}$	54	135	0.81	0.80	0.78	0.79
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	54	135	0.80	0.80	0.79	0.80
	70%	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.82	0.80
		Hotelling's $T^2_{\sum_m p_m,N}$	73	190	0.80	0.80	0.78	0.81
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	73	190	0.80	0.80	0.78	0.81

Table 6. The N and power estimates when $\epsilon_{id} = \phi\epsilon_{i,d-1} + v_{id}$, where $\phi = 0.5$ and -0.5 and v_d are i.i.d Normal(0, 0.75).

			Average standardized proximal effect size					
			0.1	0.06	0.1	0.06	0.1	0.06
ϕ	Availability	Test statistics	N		Formulated power		Monte Carlo power	
0.5	100%	$\chi^2_{\sum_m p_m}$	46	127	0.81	0.80	0.82	0.82
		Hotelling's $T^2_{\sum_m p_m,N}$	54	135	0.81	0.80	0.78	0.77
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	54	135	0.80	0.80	0.78	0.80
	70%	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.79	0.81
		Hotelling's $T^2_{\sum_m p_m,N}$	73	190	0.80	0.80	0.79	0.78
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	73	190	0.80	0.80	0.77	0.79
−0.5	100%	$\chi^2_{\sum_m p_m}$	46	127	0.81	0.80	0.80	0.81
		Hotelling's $T^2_{\sum_m p_m,N}$	54	135	0.81	0.80	0.83	0.79
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	54	135	0.80	0.80	0.80	0.77
	70%	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.80	0.82
		Hotelling's $T^2_{\sum_m p_m,N}$	73	190	0.80	0.80	0.78	0.82
		Hotelling's $T^2_{\sum_m p_m,N-q-1}$	73	190	0.80	0.80	0.79	0.78

Table 7. The N and power estimates when τ_d is not constant over d, that is, linearly increasing or decreasing over d.

		Average standardized proximal effect size					
		0.1	0.06	0.1	0.06	0.1	0.06
Trend of τ_d	Test statistics	N		Formulated power		Monte Carlo power	
Increasing	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.80	0.79
	Hotelling's $T^2_{\sum_m p_m,N}$	73	190	0.80	0.80	0.82	0.79
	Hotelling's $T^2_{\sum_m p_m,N-q-1}$	73	190	0.80	0.80	0.80	0.78
Decreasing	$\chi^2_{\sum_m p_m}$	65	182	0.80	0.80	0.79	0.79
	Hotelling's $T^2_{\sum_m p_m,N}$	73	190	0.80	0.80	0.78	0.77
	Hotelling's $T^2_{\sum_m p_m,N-q-1}$	73	190	0.80	0.80	0.77	0.80

recognized early. These approaches suggested by Magirr et al.[36] can be used to evaluate the intervention categories of an MRT design efficiently through a series of interim analyses.

The proposed sample size calculators can be extended to allow for adaptive randomization probabilities, determined by the proximal outcomes and message categories at previous decision time points,[37] with the aim of sending more effective messages to the participants. In other words, the sequential outcomes and randomization probabilities of the current message categories would depend on the outcomes and message categories from previous decision time points, in order to achieve experimental objectives. In this type of design, we can estimate not only the current proximal effect, but also the delayed effect, of each category of a particular intervention component. Similar to Dempsey et al.,[38] one can also consider stratifying strategies, where the inverse probability weighting techniques may be incorporated into the test statistic construction, enabling it to deal with more complex dependencies, for example, a setting where the proximal outcome may depend on both treatments of today and yesterday. Another possible future direction can be extending the sample size calculators to also account for the binary outcomes, similar to the method of Qian et al.[39]

Acknowledgements

The authors would like to thank the reviewers and editors for the valuable comments and feedback. We also acknowledge the feedback from our colleague, Dr Raju Maiti. We would like to acknowledge Chris Karr who helped developing and managing the DIAMANTE automated text messaging service and passive data collection. This work has been partially supported by Khoo Bridge Funding Award (Duke-NUS-KBrFA/2021/0040) and the start-up grant from the Duke-NUS Medical School, Singapore, as well as an Academic Research Fund Tier 2 grant (MOE-T2EP20122-0013) from the Ministry of Education to Dr Bibhas Chakraborty. The DIAMANTE trial has been funded by an R01 grant to Dr Adrian Aguilera (University of California, Berkeley) and Dr Courtney Lyles (University of California, San Francisco) who designed the DIAMANTE application, 1R01 HS25429-01 from the Agency for Healthcare Research and Quality. We acknowledge Dr Adrian Aguilera and Dr Courtney Lyles, the Principal Investigators of the DIAMANTE study, for involving us in this important mHealth study and for sharing the UC-Berkeley student data.

Data availability

The students dataset from the adapted DIAMANTE study used to demonstrate the proposed sample size calculators in Section 3 is available from the co-author, Dr Caroline Figueroa, `C.Figueroa@tudelft.nl`, upon reasonable request.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been partially supported by Khoo Bridge Funding Award (Duke-NUS-KBrFA/2021/0040) and the start-up grant from the Duke-NUS Medical School, Singapore, as well as an Academic Research Fund Tier 2 grant (MOE T2EP20122-0013) from the Ministry of Education to Dr Bibhas Chakraborty. The DIAMANTE trial has been funded by an R01 grant to Dr Adrian Aguilera (University of California, Berkeley) and Dr Courtney Lyles (University of California, San Francisco) who designed the DIAMANTE application, 1R01 HS25429-01 from the Agency for Healthcare Research and Quality.

ORCID iDs

Jing Xu https://orcid.org/0000-0003-0687-9004
Bibhas Chakraborty https://orcid.org/0000-0002-7366-0478

Supplemental material

Supplemental materials for this article are available online.

References

1. Adibi S. *Mobile Health: A Technology Road Map*. Cham: Springer, 2015.
2. Lewis M, Uhrig J, Bann C et al. Tailored text messaging intervention for HIV adherence: a proof-of-concept study. *Health Psychol* 2013; **32**: 248–253.
3. King A, Castro C, Buman M et al. Behavioral impacts of sequentially versus simultaneously delivered dietary plus physical activity interventions: the CALM trial. *Ann Behav Med* 2013; **46**: 157–168.
4. Alessi S, Petry N. A randomized study of cellphone technology to reinforce alcohol abstinence in the natural environment. *Addiction* 2013; **108**: 900–909.
5. Free C, Phillips G, Galli L et al. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. *PLoS Med* 2013; **10**: e1001362.
6. Intille S. Ubiquitous computing technology for just-in-time motivation of behavior change. *Stud Health Technol Inform* 2004; **107**: 1434–1437.
7. Patrick K, Griswold W, Raab F et al. Health and the mobile phone. *Am J Prev Med* 2008; **35**: 177–181.
8. Nahum-Shani I, Smith S, Spring B et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med* 2018; **52**: 446–462.
9. Klasnja P, Hekler E, Shiffman S et al. Micro-randomized trials: an experimental design for developing just-in-time adaptive interventions. *Health Psychol* 2015; **34**: 1220–1228.
10. Liao P, Klasnja P, Tewari A et al. Sample size calculations for micro-randomized trials in mHealth. *Stat Med* 2016; **35**: 1944–1971.
11. Klasnja P, Smith S, Seewald N et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. *Ann Behav Med* 2019; **53**: 573–582.
12. Liao P, Dempsey W, Sarker H et al. Just-in-time but not too much: determining treatment timing in mobile health. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2018; **2**: 179.
13. Aguilera A, Figueroa C, Hernandez-Ramos R et al. An mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE study. *BMJ Open* 2020; **10**: e034723.
14. Figueroa C, Hernandez-Ramos R, Boone C et al. A text messaging intervention for coping with social distancing during COVID-19 (StayWell at home): protocol for a randomized controlled trial. *JMIR Res Protoc* 2021; **10**: e23592.
15. Ventz S, Cellamare M, Parmigiani G et al. Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostatistics* 2017; **19**: 199–215.
16. Lee K, Wason J and Stallard N. To add or not to add a new treatment arm to a multiarm study: a decision-theoretic framework. *Stat Med* 2019; **38**: 3305–3321.
17. Mei B, May L, Heap R et al. Rapid development studio: an intensive, iterative approach to designing online learning. *Sage Open* 2021; **11**: 1–9.
18. DeVries R, Truong K, Kwint S et al. Crowd-designed motivation: motivational messages for exercise adherence based on behavior change theory. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 2016, pp. 297–308.
19. Williams J, Kim J, Rafferty A et al. AXIS: generating Explanations at scale with learnersourcing and machine learning. In *Proceedings of the 3rd (2016) ACM Conference on Learning @ Scale* 2016; 379–388.
20. Koning R, Hasan S and Chatterji A. Experimentation and start-up performance: evidence from A/B testing. *Manage Sci* 2022; **68**: 6434–6453.
21. Fabijan A, Dmitriev P, McFarland C et al. Experimentation growth: evolving trustworthy A/B testing capabilities in online software companies. *J Softw: Evol Process* 2018; **30**: e2113.
22. Figueroa C, Deliu N, Chakraborty B et al. Daily motivational text-messages to promote physical activity in university students: results from a micro-randomized trial. *Ann Behav Med* 2022; **56**: 212–218.
23. Boruvka A, Almirall D, Witkiewitz K et al. Assessing time-varying causal effect moderation in mobile health. *J Am Stat Assoc* 2018; **113**: 1112–1121.
24. Seewald N, Smith S, Lee A et al. Practical considerations for data collection and management in mobile health micro-randomized trials. *Stat Biosci* 2019; **11**: 355–370.
25. Tu X, Kowalski J, Zhang J et al. Power analyses for longitudinal trials and other clustered designs. *Stat Med* 2004; **23**: 2799–2815.
26. Mancl L, DeRouen T. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**: 126–134.
27. Hotelling H. The generalization of student's ratio. *Ann Math Stat* 1931; **2**: 360–378.
28. Li P, Redden D. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Stat Med* 2015; **34**: 281–296.

29. Almirall D, Compton S, Gunlicks-Stoessel M et al. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Stat Med* 2012; **31**: 1887–1902.
30. Kelley K, Maxwell S and Rausch J. obtaining power or obtaining precision: delineating methods of sample-size planning. *Eval Health Prof* 2003; **26**: 258–287.
31. Maxwell SE, Kelley K and Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol* 2008; **59**: 537–563.
32. Yan X, Ghosh P and Chakraborty B. Sample size calculation based on precision for pilot sequential multiple assignment randomized trial (SMART). *Biometrical J* 2021; **63**: 247–271.
33. Buuren SV, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011; **45**: 1–67.
34. Qian T, Klasnja P and Murphy S. Linear mixed models with endogenous covariates: modeling sequential treatments with application to a mobile health study. *Stat Sci (With Discussion)* 2020; **35**: 375–390.
35. Bidargaddi N, Almirall D, Murphy S et al. 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* 2018; **6**: e10123.
36. Magirr D, Jaki T and Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**: 494–501.
37. Wason J, Trippa L. A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials a comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Stat Med* 2014; **33**: 2206–2221.
38. Dempsey W, Liao P, Kumar S et al. The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments. *Ann App Stat* 2020; **28**: 2687–2708.
39. Qian T, Yoo H, Klasnja P et al. Estimating time-varying causal excursion effect in mobile health with binary outcomes. *Biometrika* 2021; **108**: 507–527.