**Psychological Methods:** *Its purpose is the dissemination of innovations in research design, measurement, methodology, and quantitative analysis to the psychological community, and to promote effective communication about related substantive and methodological issues. Convey how article is of interest to multiple subdisciplines of psychology.*

**Type of article:** Original article, 12,000 words max (approx. 40 double-spaced pages in 12-point Times New Roman), not including references, tables, figures, and appendixes. Use MathType to construct equations.

**Cover letter:** assurance that manuscript is not under review elsewhere, that any primary data have not been published previously or accepted for publication, and that the appropriate ethical guidelines were followed in the conduct of the research. Must specifically request a masked review if we wish – then make every effort to provide no clues as to author identity. May suggest individuals qualified to review manuscript.

**Abstract #1 (separate page, with keywords):** 250 words max, description for scientists and for purposes of document retrieval from reference databases – required at submission

**Keywords:** (up to 5) Micro-randomized trial; digital intervention; just-in-time adaptive intervention; intensive longitudinal data

**Abstract #2 (separate page):** 250 words max, communicate essence of the article and its value to applied researchers (who are not methodologists)

**Author footnote (separate page):** Acknowledgements of support and name and address of corresponding author. Provide information (2-4 sentences) on prior dissemination of the data and narrative interpretations of the data/research appearing in the manuscript (e.g., if some or all were presented at a conference or meeting, posted on a listserv/website, etc.)

**Title page (separate page):** Each author’s name and affiliation

**Supplemental Material:** Provide runnable source code as supplemental material to the article.

**The Micro-Randomized Trial for Developing Mobile Health Interventions: Data Analysis Methods**

Tianchen Qian, Michael A. Russell, Linda Collins, Pedrag Klasnja, Stephanie T. Lanza, Hyesun Yoo, Susan A. Murphy

**Abstract**

Although there is much excitement surrounding the use of mobile and wearable technology for the purposes of delivering interventions as people go through their day-to-day lives, data analysis methods for constructing and optimizing digital interventions lag behind. Here, we elucidate data analysis methods for primary and secondary analyses of micro-randomized trials (MRTs), an experimental design to optimize digital just-in-time adaptive interventions. We provide a definition of causal “excursion” effects suitable for use in digital intervention development. We introduce the weighted and centered least-squares (WCLS) estimator which provides unbiased causal excursion effect estimators for digital interventions from MRT data. We describe how the WCLS estimator along with associated test statistics can be obtained using standard statistical software such as SAS or R (R Core Team, 2019; SAS Institute Inc., 2019). Throughout we use HeartSteps, an MRT designed to increase physical activity among sedentary individuals, to illustrate potential primary and secondary analyses.

*Keywords*: Micro-randomized trials (MRT), digital health interventions, just-in-time adaptive interventions (JITAI), intensive longitudinal data

**The Micro-Randomized Trial for Developing Mobile Health Interventions: Data Analysis Methods**

Mobile technologies—including tablets, smartphones, and wearable sensors—have become ubiquitous in daily life. Because of their ability to engage users “in the moment,” they provide unprecedented opportunity to deliver interventions at the times and in the contexts individuals are most likely to benefit. However, data analysis methods for designing such digital just-in-time adaptive interventions (JITAIs) lag behind technology’s current capabilities to reach individuals, with the result that scientific knowledge is lacking about when, how often, in which context, and what intervention content should be delivered so that the intervention is effective without overburdening the individual. The micro-randomized trial (MRT) is an experimental trial for obtaining this knowledge. In an MRT, intervention components are randomized and delivered during the flow of people’s daily lives (see the companion article, Walton et al. (submitted), for more details and examples of MRTs).

As discussed in Walton et al. (submitted), the MRT was developed for use in constructing and optimizing JITAI components in digital interventions (Klasnja et al., 2015; Nahum-Shani et al., 2017). We provide a brief review of the MRT below along with an overview of the HeartSteps MRT (Klasnja et al., 2018) for use in clarifying ideas. In data analyses following a MRT, a natural primary analysis would focus on whether there is a marginal effect of the intervention component. Possible secondary analyses might include moderation analyses. However, in the MRT setting where intervention components are repeatedly delivered over time but only in settings in the individual’s life in which it is appropriate to deliver the intervention component, the meaning of marginal effect and moderation effects requires careful thought. In this article, we provide precise definitions via the concept of causal “excursion” effects. To do this we use the potential outcomes framework (Robins, 1986; Rubin, 1978). Next we review and elucidate the weighted and centered least-squares (WCLS) estimator (Boruvka, Almirall, Witkiewitz, & Murphy, 2018) which provides estimators and test statistics for conducting primary and secondary analyses using data from MRTs. Data from the HeartSteps MRT is used to illustrate these analyses.

**Brief Review of Micro-Randomized Trials (MRTs)**

MRT are conducted with the goal of providing intensive longitudinal data that can be used to assess the causal effects of, as well as develop, one or more JITAI components. Each component is associated with different intervention options that might be provided at any of multiple *decision points* during an individual’s day-to-day life. For example, one of the components examined in the HeartSteps MRT was the *activity suggestions* component for which the intervention options were a *contextually tailored activity suggestion*[[1]](#footnote-1) or *no suggestion*. The activity suggestions component was randomly assigned at 5 decision points per day. An important objective of conducting an MRT is to construct decision rules and identify tailoring variables to determine which intervention option to provide at a decision point, that is, to identify the set of tailoring variables and decision rules that will result in an effective JITAI.

In an MRT, each individual is repeatedly randomized among the different options of an intervention component with known probability at each decision point. This repeated, intensive randomization means that over the course of an MRT, an individual may be randomized hundreds or even thousands of times. As discussed in the companion article (Walton et al., submitted), intervention components are often designed to have greatest impact on a near-term *proximal outcome*. The proximal outcome is observed after each randomization. Prior to each decision point, sensor data as well as self-report data may be observed.

As an individual goes through their everyday life, there may be times when only the intervention option of “no treatment” is appropriate. This is formalized with the notion of *availability*. For example, often the delivery of the intervention involves an audible and visual cue. If sensors on the phone and/or wearables indicate that the individual might be operating a moving vehicle at a decision point, then to avoid potentially dangerous distractions the individual might be deemed unavailable for an intervention. Individuals also may be determined as unavailable if, for example, they are not able to receive text messages (e.g., phone set to airplane mode). At times of unavailability, the only appropriate intervention is the “no treatment” option. Availability is one of the time-varying variables observed prior to each decision point.

**HeartSteps**

The HeartSteps study involves a 6-week MRT designed to inform the optimization of the HeartSteps digital intervention to increase physical activity (Klasnja et al., 2018; Walton et al., submitted). HeartSteps combines a wristband activity tracker that monitors individuals’ steps throughout the day in concert with a mobile phone application. Two intervention components were experimented on in the HeartSteps MRT—a planning support component and the contextually tailored activity suggestions component described above. The *planning support* component consisted of support for planning in the evening for activity on the next day. The intervention options were *planning* and *no planning*[[2]](#footnote-2). For simplicity, much of the exposition below focuses on the activity suggestions component.

Activity suggestions were randomized at 5 decision points each day: morning commute, lunch time, mid-afternoon, evening commute, and after dinner. The exact time-of-day of the five decision points was specified by each individual at the beginning of the study and could be different between individuals. At each decision point, if the individual was available, the probability of delivering a contextually tailored activity suggestion (as opposed to no suggestion) was .6. The proximal outcome for the activity suggestion component was the step count of the individual in the 30-minute window following a decision point.

In the case of the activity suggestions component, as mentioned above, an individual was deemed unavailable at a decision point if the individual’s speed indicated that the individual might be driving. Furthermore, because the content of the activity suggestions involved suggestions for new activities, an individual was deemed unavailable if they were currently walking or running or they had just finished an activity bout in the previous 90 seconds prior to the decision point.

A natural primary analysis for the HeartSteps MRT is to test whether there is a casual effect of delivering an activity suggestion versus not delivering any suggestion on the proximal outcome—i.e., the subsequent 30-minute step count. This question is conceptually similar to testing for the main effect of a factor (Collins, Dziak, & Li, 2009) as this causal effect is marginal over the other intervention components; see below. Important concerns in the development of the HeartSteps activity suggestion component include habituation (Rankin et al., 2009) and/or treatment burden (Clawson, Pater, Miller, Mynatt, & Mamykina, 2015; Eysenbach, 2005; Ho & Intille, 2005; Klasnja et al., 2008; Shaw et al., 2013; Yardley et al., 2016). If the participant is habituating to the activity suggestions or finding the intervention burdensome, the causal effect would be expected to deteriorate over time. Thus, a natural secondary or exploratory analysis is to assess whether day in study moderates the effect of delivering an activity suggestion. Two additional examples of moderation analyses are whether the effect of delivering an activity suggestion depends on the current location of an individual, and whether the effect of the activity suggestion depends on whether the individual was prompted to plan an activity for that day. All of these statements are imprecise with regards to what is meant by a causal effect. In the following section we make these more precise.

**The Causal Excursion Effect**

In this section, we define the causal effect of a mobile health intervention component using the potential outcomes framework (Liao, Klasnja, Tewari, & Murphy, 2016; Robins, 1986, 1987; Rubin, 1978). For expositional clarity, throughout the paper we consider the setting in which there are only two intervention options denoted by treatment 1 and treatment 0—in the setting of the HeartSteps MRT, this would be delivering activity suggestion (treatment 1) and not delivering activity suggestion (treatment 0). First, we briefly review the definition of a causal effect using a hypothetical setting with a single time point treatment. Then we provide the definition of the “causal excursion effect” of a time-varying intervention component on a time-varying outcome.

In the classical potential outcomes formulation where there is only a single time point for possible treatment [see review by Rubin (2005)], the ideal but usually unattainable goal is to determine the individual-level causal effect, or the difference between the outcome under treatment 1 [denoted by ] and the outcomeunder treatment 0 [denoted by ] at the same time for each individual. Consider the first decision point in the HeartSteps MRT. At this decision point individuals are randomly assigned to receive an activity suggestion or no suggestion. The step count in the 30-minute window following this decision point is the outcome. For each individual, the treatment effect at this decision point is the difference between (a) the 30-minute step count had treatment been assigned to the individual () and (b) the 30-minute step count had the treatment not been assigned to the individual (). and are called *potential outcomes*, because in reality we can observe only one of them for each individual, as both treatment and no treatment cannot be assigned to an individual at the same time—this is the “fundamental problem of causal inference” (Holland, 1986).

Let denote the treatment assignment ( if treatment 1; if treatment 0). We observe only .[[3]](#footnote-3) A widely-adopted solution to this problem is to estimate a marginal causal effect, or to estimate an effect closer to an individual causal effect, the causal effect conditional on a pre-treatment variable —i.e., the difference between the expected outcome had everyone with received the treatment () and the expected outcome had everyone with *not* received the treatment (). In the example for the first decision point in the HeartSteps MRT, might be the individual’s current location (home, work or other), current weather, gender, and baseline activity level. An interesting scientific question would be whether the value of modifies the treatment effect. Inference for the difference, , provides the answer to this question.

Randomization of allows us to connect the difference in potential outcomes with a quantity we can estimate using data () on a sample of individuals from the population. In particular, if treatment is randomly assigned, the causal effect, , is equal to (see Rubin (2005)).

To define the causal excursion effect of a time-varying intervention component on a time-varying outcome, we need notation to accommodate time. Consider the HeartSteps MRT and suppose we are interested in assessing the effect of the activity suggestion on the subsequent 30-minute step count. Recall that the HeartSteps MRT is a 42-day study and there were 5 decision points per day for the activity suggestion component; thus, there are decision points overall. Let represent the data available from decision point up to and including at decision point .[[4]](#footnote-4) contains the availability indicator, , with meaning that the individual is available at decision point *t* and otherwise*.* In the HeartSteps example, also includes current weather, location, time of day, 30-minute step count prior to decision point , and whether planning support was provided on the prior evening. Let represent the treatment indicator at decision point , where means treatment is delivered and means treatment is not delivered. Let denote a sequence of treatments up to and including at decision point , i.e., .[[5]](#footnote-5) Let represent the proximal outcome, here the number of steps in the 30 minutes after decision point . Denote by the individual’s history of data observed up to decision point , i.e., . includes potential moderators and control variables; in addition, summaries of may be used to form these variables. We denote potential moderators by . Recall that potential moderators of the effect of the activity suggestions in HeartSteps, , include number of days in treatment, the individual’s current location (home, work or other), current weather as well as gender and baseline activity level; another potential moderator might be whether planning was prompted the prior evening. As in the single time point setting, the inclusion of potential moderators, , means that the desired causal excursion effect is conditional on these variables. Further discussion regarding this point is included towards the end of this section after the formal definition of the causal excursion effect.

To define the causal excursion effect, we use an extension of the potential outcomes framework to the setting of intensive longitudinal data (Robins, 1986, 1987). The potential outcomes for are , respectively. For example, is the 30-minute step count outcome after decision point that would have been observed if the individual had been assigned treatment sequence . (For binary treatments, there could be different potential outcomes, ’s!) This notation encodes the reality that an individual’s 30-minute step count outcome after decision point may be impacted by all prior treatments as well as current treatment, . Note that unlike the potential proximal outcome , potential outcomes for , as well as availability, , are indexed *only* by treatments, , prior to decision point —namely, , , and . This is because , , and occur prior to treatment at

The *causal excursion effect* of activity suggestions at decision point on subsequent 30-minute step count for available individuals with is defined as (Boruvka et al., 2018; Liao et al., 2016)

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

This formula contains the following information.

1. The effect,, is *causal* as it is the expected value of the contrast in step counts in the 30 minutes following a decision point if the treatment were delivered at [potential outcome ] versus if treatment were not delivered at [potential outcome ]; that is, .
2. The effect, , is *conditional*. This effect is only among individuals who are available () and for whom their potential moderators take on the value of .
3. The effect, , is *marginal*. For example, if is empty, then this effect is marginal over the effects of prior as well as the other intervention components. In HeartSteps, the effect of the activity suggestions component at decision point is marginal over the effects of interventions from prior decision points and the planning support component.
4. The effect, , is an *excursion* from the “treatment schedule” prior to . The treatment schedule prior to is a set of stochastic decision rules for treatment assignment at all decision points from the beginning of the intervention up to the previous decision point; that is, for assignment of . In the case of an MRT, the treatment schedule will always involve some randomization, but may include non-random assignment as well. For example, in the HeartSteps MRT the treatment schedule included, at five decision points per day, the following: If available deliver an activity suggestion with probability .6 and no suggestion with probability .4; if not available do not deliver an activity suggestion. Suppose the HeartSteps intervention included a component that was not examined in the MRT—for example, a brief motivational video sent to all participants every Monday morning at 8 am. Although there is no experimentation on this component, it would be included in the treatment schedule.

The causal excursion effect concerns what would happen if we follow the current treatment schedule up to time and then deviate from the schedule to assign treatment 1 at decision point , versus deviate from the schedule to assign treatment 0 at decision point . In other words, the definition of implicitly depends on the schedule for assigning . Technically this excursion can be seen from (1) in that the expectation, , is marginal over all prior treatments not contained in . For example, if contains only current weather, then the excursion effect is marginal over the schedule for assigning all of the prior treatments, as well as all prior treatments for other components such as the planning component.

To understand the excursion effect better, consider the following two very different treatment schedules. In the first schedule, treatment is provided on average once every other day; in the second schedule, the treatment is provided on average 4 times per day. Excursions from these two rather different schedules could result in potentially very different effects, . Indeed, in the latter schedule individuals may experience a great deal of burden and disengage with the result that would be close to 0, whereas in the former schedule individuals may still be very engaged, resulting in a larger . In other words, the definition of the causal excursion effect, , is dependent on the schedule for treatment assignment; this is different from the types of effects typically discussed in the causal inference literature (e.g., Robins, 1994; Robins, Hernán, & Brumback, 2000).

A primary hypothesis test might focus on inference about the marginal excursion effect, that is, (1) with equal to an empty set. A secondary analysis might consider treatment effect moderation by including in potential moderators, such as location, or number of days in the intervention. Note that does not need to include all true moderators for (1) to be a scientifically meaningful causal effect; instead, it is appropriate to choose any (or no) , provided that (1) is interpreted as the causal excursion effect marginal over all variables in that are not included in .

Under the assumption that the treatment is sequentially randomized (which is guaranteed by MRT design) and that the treatment delivered to one individual does not affect another individual’s outcome[[6]](#footnote-6), the causal excursion effect in (1) can be written in terms of expectations over the distribution of the data as:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Thus we can make inference about the excursion effect using the observed MRT data, by specifying a regression model for .

**A Primary Analysis for HeartSteps**

As discussed above, a natural primary analysis for the HeartSteps MRT is to assess the marginal causal excursion effect of delivering an activity suggestion on the subsequent 30-minute step count of the user, compared to not delivering any message. This effect is marginal in the sense that it averages over all users and all decision points (at which the user is available). To operationalize this marginal causal excursion effect, let be an empty set in (2) so that

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

The outer expectation on the right-hand side in (3) represents an average across all possible values of across individuals. For example, is averaged over weather on that day and on previous days, over previous treatment assignment of the activity suggestions (i.e., for ), and also over previous assignment of the planning support prompts. To create a scalar summary, we average over with weights equal to to obtain the marginal excursion effect:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

In the section “Analysis Using Data from HeartSteps MRT,” we will conduct inference about a variety of excursion effects including the marginal excursion, , of the activity suggestions (Klasnja et al., 2018).

**A Selection of Secondary Analyses for HeartSteps**

Secondary analyses concern moderation of the causal excursion effect by a non-empty . For example, a secondary analysis may assess whether the causal excursion effect deteriorates with day under treatment, in which case would include , the number of days in treatment prior to the decision point . Suppose the data analyst assumes a linear model on the causal excursion effect:

Note for all decision points on the first day of treatment; represents the causal excursion effect on the first day and represents the change in the causal excursion effect with each additional day.

Other secondary analyses might concern effect moderation by other time-varying covariates such as the current location of the user, or by another intervention component being examined in the MRT. Consider the planning support component in the HeartSteps MRT. Let denote the indicator of whether a planning support prompt was delivered on the evening prior to decision point ( if delivered, if not)[[7]](#footnote-7). This secondary analysis of effect moderation by a planning support prompt on the previous evening can be expressed as

Here represents the causal excursion effect when the individual did not receive planning support on the prior evening, and represents the causal excursion effect when the individual received planning support on the prior evening. See the section “Analysis Using Data from HeartSteps MRT” for results of the secondary analyses described above.

**Methods for Estimating Causal Excursion Effects from MRT Data**

We present a weighted and centered least-squares (WCLS) estimator, developed by Boruvka et al. (2018), that provides an unbiased estimator for the causal excursion effect, . Here for clarity we provide an overview of the estimation method that can be used when the randomization probabilities are constant, as is the case in HeartSteps. Recall that in HeartSteps a primary analysis might be an assessment of the marginal, causal excursion effect of the activity suggestions on the subsequent 30-minute step count. Throughout, we use the term “unbiased/biased” to mean that the estimate becomes ever closer to the true parameter value as sample size increases; this roughly means that the estimator is unbiased if the number of participants is sufficiently large. We use the superscript to denote the transpose of a vector or a matrix.

Suppose we assume a linear model for the causal excursion effect: with defined in (2), and the goal is to make inference about . Note that the model for only characterizes the treatment effect (as a contrast between the two treatments with the proximal outcomes as dependent variables). We will also use a *working model* for the conditional mean of given no treatment at decision point and history [i.e., ]. One working model might be where is a vector of summaries of the observations made prior to decision point (i.e., summaries constructed from ). These summaries are often called control variables. For example, in HeartSteps a natural control variable is the step count in the 30 minutes prior to the decision point. This prior 30-minute step count variable is likely strongly correlated with the proximal outcome, the 30-minute step count following the decision point. As we will see shortly, *unbiasedness of the WCLS estimator for does not require to be a correct model for .* The role of is to reduce noise in the analysis: Inclusion of prognostic control variables (those variables in that are correlated with ) will usually reduce the variance of the estimator of . A simulation study that illustrates this point is presented in Appendix C. In summary, the inclusion of the step count in the 30 minutes prior to the decision point is to reduce variance and increase the power to detect a nonzero excursion effect, .

The WCLS estimator for is calculated as follows. Suppose is the value that solves the following estimating equation (Diggle et al., 2002)

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

where is the randomization probability and is the index for the -th individual. Note that the indicator is centered by subtracting *p.* Recall that in HeartSteps, at each decision point there is .6 probability to deliver an activity suggestion (if the individual is available), so . The that solves (5) is the WCLS estimator for .

**Remarks**.

1. WCLS does *not* assume a model for the proximal outcome such as . The primary assumption (Boruvka et al., 2018) that ensures that is unbiased is

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

that is, is a correct model for the causal excursion effect conditional on .

1. The centering of the treatment indicator, , in (5) creates orthogonality between the columns of the design matrix for the causal excursion effect (i.e., ) and the columns of the design matrix involved in the working model, , for (i.e., ). This centering provides robustness in the estimation of ; in particular robustness against a mis-specified working model for . In other words, the data analyst can use a possibly incorrect working model , and will still be an unbiased estimator of . In mobile health this robustness property is of practical importance, because vast amounts of data (i.e., high-dimensional ) on the individual has usually been collected prior to decision point . As a result, it is virtually impossible to correctly model . For example, in HeartSteps there are 210 decision points (210 = 42 days 5 times/day) for each individual; can include the outcome, treatment and covariates from all the past decision points, which means hundreds of variables at a later decision point . In addition, may depend on variables in in a nonlinear way, which adds to the difficulty of correctly modeling and thus makes the robustness property desirable.
2. While the choice of doesn’t affect the unbiasedness of , a better working model for has the potential to decrease the variance of . Because is usually high dimensional, choosing can be done by hand-picking a subset of (e.g., those covariates and outcomes at recent decision points). As discussed above, in HeartSteps a natural control variable is the step count in the 30 minutes prior to the decision point, as this variable is likely highly correlated with the proximal outcome of step count in the 30 minutes after the decision point. In Appendix C we illustrate through a simulation study how inclusion of control variables that are correlated with the proximal outcome in can reduce the variance of .
3. Although software based on the estimating equation (5) also outputs an estimator and its standard error, we recommend not interpreting them, unless it is safe to assume that is a correct model for .

For simplicity in presentation of the estimator, so far we have considered the setting where the randomization probability, , is constant. There are also practical settings where the randomization probability may change over time; for example, in the stratified micro-randomized trial, different micro-randomization probabilities are used depending on a time-varying variable such a prediction of risk. For example, a higher randomization probability may be used when the individual is categorized as high-risk, and a lower randomization probability may be used when the individual is categorized as low-risk. The rationale for such risk stratification is to ensure that there are adequate numbers of treatments delivered both at risk times and at non-risk times. See Dempsey, Liao, Kumar, & Murphy (2019) for details. The WCLS estimator presented above can be generalized to this setting (Boruvka et al., 2018); see Appendix B for a general WCLS estimator that allows randomization probability to depend on the individual’s history, .

**Estimating the WCLS using Standard Statistical Software**

When the randomization probabilities are constant, standard statistical software that implements GEE (Liang & Zeger, 1986) such as SAS (SAS Institute Inc., 2019), Stata (StataCorp, 2019), and SPSS (IBM Corp., 2019) can be “tricked” into providing the WCLS estimator as well as its standard error. Consider SAS PROC GENMOD and suppose the assumed causal excursion effect model is (6), the working model for is , then the WCLS estimator and its standard error can be obtained by the following steps: (i) incorporate as the “prior weights”, (ii) choose a working independence correlation structure, and (iii) fit GEE with dependent variable and independent variables and . Then the estimated coefficient for is the WCLS estimate . Note that in choosing the control variables, needs to contain at least .

Recall that we are not actually fitting a GEE model for the conditional mean of . We are just using the GEE software as a means to output the WCLS estimator. Technically, this can be done because the estimating equation of a GEE with the above specification is algebraically equivalent to (5), the estimating equation of WCLS.

To obtain appropriate standard errors for the estimator through the above GEE fit, one needs to use the robust standard error [SAS calls this the “empirical standard error” (SAS Institute Inc., 2019)]. The robust standard error accounts for the fact that the proximal outcomes, are correlated. When the sample size is small (e.g., ), we recommend use of further small sample corrections for both the standard error and the degrees of freedom in the critical value for constructing confidence intervals (Boruvka et al., 2018). R code (R Core Team, 2019) for the implementation with the small sample correction is available at <https://github.com/StatisticalReinforcementLearningLab/HeartstepsV1Code/blob/master/xgeepack.R>.

**Analysis Using Data from HeartSteps MRT**

Recall that HeartSteps is a 6-week MRT for optimizing JITAI components of a digital intervention to promote physical activity with 37 sedentary participants (Klasnja et al., 2018). In the illustrative analysis below, we focus on the activity suggestion component, which was randomized at 5 decision points each day. At each decision point, if the individual was available, an activity suggestion was delivered with randomization probability .6. We first assess the marginal excursion effect of an activity suggestion versus no suggestion using data from the HeartSteps MRT; this is a natural primary analysis. The primary analysis for HeartSteps is published in Klasnja et al., (2018); for completeness we include this analysis as well as results of additional secondary analyses. As discussed before, secondary analyses might include how the excursion effect changes over time, whether its effect is moderated by current location, and whether this effect is moderated by whether a planning support prompt is delivered on the evening prior to the decision point. All analyses are conducted using the R programming language (R Core Team, 2019). We use the following variables in the analysis:

* : log-transformed 30-minute step count following decision point . This is the proximal outcome of interest.
* : indicator of whether an activity suggestion is delivered at decision point . The randomization probability is .6 at available decision points.
* : log-transformed 30-minute step count preceding decision point . Because this variable is expected to be correlated with , we will include as a control variable in the analysis to reduce noise.
* : day in the study, coded as 0, 1, 2, …, 41.
* : individual’s location at decision point *t*; coded as 1 if at home or at work, and 0 if at any other location.
* : indicator of a planning support delivered on evening prior to decision point; coded as 1 if delivered and 0 if not.
* : availability status at decision point . Recall that randomization can only occur if an individual is available.

Step count data is highly skewed; the log-transformation is used to make its distribution more symmetric (and we added .5 to the step count before taking log to avoid log(0)). Although the unbiasedness of the WCLS estimator does not require to be symmetrically distributed, the symmetry improves the accuracy of the approximation to the distribution of the test statistic in small samples.

**Question 1: Is there an average effect of delivering an activity suggestion on subsequent 30-minute step count, compared to no suggestion?**

As discussed above a natural primary analysis concerns the excursion effect marginal over all decision points and all covariates; this is the focus of this question. We address this question using the WCLS estimator with equal to the empty set, , working model , and weight . Table 1 lists the results. The causal effect of delivering an activity suggestion versus no suggestion on the log-transformed subsequent 30-min step count, averaged over all decision points and all covariates, is (*p* = 0.060, 95% CI = -0.006 to 0.268). This corresponds roughly to a 14% () increase in the average 30-minute step count (on its original scale), comparing decision points when an activity suggestion was sent with decision points when an activity suggestion was not sent.

**Question 2: Does the effect of the activity suggestions change with each additional day in the study?**

This question is motivated by the expectation that the longer a person participates in the study, the more they may habituate to the suggestions or become overburdened, leading them to become less responsive. We address this question via the WCLS estimator with (day in study), , working model , and weight . is included in to assess the effect moderation by , day in the study. Because is coded to start from 0, represents the causal excursion effect on the first day. Table 2 lists the results. There is a significant interaction between the activity suggestion and day in the study: the causal effect of the activity suggestion changes by with each additional day in the study (*p* = 0.005, 95% CI = -0.031 to -0.006). Combining this with  , the data indicates that sending an activity suggestion results in about 66% () increase in the 30-minute step count on the first day of the study, about 16% () increase on the 21st day of the study, and about 21% () decrease on the 42nd day of the study.

The above analysis uses a linear model for the causal excursion effect. To assess sensitivity of the result to potential non-linearity, we fit a local 2-degree polynomial regression with smoothing span 2/3 and tricubic weighting to estimate the causal excursion effect over time (the default setting for many local regression software such as the lowess function in R (R Core Team, 2019)). The estimated effect from local regression is presented in Figure 1 (black curve). Comparing this estimated effect with the estimated effect based on the linear model (blue curve in Figure 1, with blue shaded area being the pointwise 95% confidence interval), we see that the two estimates are relatively close to each other, indicating that the linear model fits well.

**Question 3: Does the effect of delivering *each type of* activity suggestion versus no suggestion depend on the individual’s current location (home/work, or other)?**

The activity suggestion involves suggestions for new physical activities; therefore, it is of interest to examine if its effect depends on the location of the individual, which might be a proxy for interruptibility. If an activity suggestion is sent, then ½ of the time the suggestion is a walking suggestion (instructing a walking activity that took 2-5 minutes to complete) and the remaining ½ of the time the activity suggestion is an anti-sedentary suggestion (instructing brief movements such as stretching one’s arms). It is conjectured by the investigators that the effect moderation by location may differ for walking suggestions and anti-sedentary suggestions. Therefore, here we assess whether the effect of delivering each of the two types of activity suggestion versus no suggestion is modified by the individual’s current location (home/work, or other). We address this question by using the WCLS estimator with (indicator of being at home or work), working model , and weight . is included in to assess the effect moderation by , location of the individual. Because here we have two treatment indicators (indicator of whether a walking suggestion is delivered, and indicator of whether an anti-sedentary suggestion is delivered), the causal excursion effect for the walking suggestion is modeled as , and the causal excursion effect for the anti-sedentary suggestion is modeled as . Table 3 lists the result. The causal excursion effect moderation by location (home/work or other) is statistically significant for walking suggestions (, *p* = 0.049, 95% CI = 0.001 to 0.753). The effect moderation is not statistically significant for anti-sedentary suggestions (, *p* = 0.472, 95% CI = -0.540 to 0.256).

**Question 4: Does the effect of activity suggestions depend on whether planning support was delivered on the previous evening?**

Whether planning support was delivered on the previous evening may impact the effectiveness of the activity suggestion. To assess this moderation effect, we use the WCLS estimator with , , working model , and weight . is included in to assess the effect moderation by , whether the individual received planning support on the previous evening. Table 4 lists the results. There is no evidence of effect moderation by the planning support prompt on the previous day (, *p* = 0.734, 95% CI = -0.228 to 0.320).

**Discussion**

**FIGURE 2**

In this article we define the causal excursion effect of a digital intervention component in an MRT using the potential outcomes framework. We illustrate how primary and secondary analyses concerning causal excursion effects can be formulated for an MRT, using the HeartSteps MRT as an example. We introduce the weighted and centered least-squares (WCLS) as a data analysis method for MRTs that results in unbiased estimators for the causal excursion effect, and present how to obtain the WCLS estimator through standard statistical software. We illustrate WCLS by using it to analyze the marginal and moderated causal excursion effects using data from the HeartSteps MRT.

**Using moderation effect analysis to inform JITAI development**

Conducting moderation analyses as well as exit interviews with participants can be useful both in formulating decision rules as well as generating hypotheses to be tested in subsequent optimization trials. For example, in exit interviews we might learn that participants found that the activity suggestions begin to appear similar as the trial progressed. This combined with the evidence of a moderation by day in study might motivate the development of different types of activity suggestions that could be introduced after, say, intervention week 3. The moderating effect of location is an early indication that the decision rules might specify no delivery of activity suggestions when an individual is at the “other” location. In the case of HeartSteps, findings from analyses such as those above, along with other moderation analyses and exit interviews, were used to inform a second optimization trial currently underway in which a personalization algorithm is being used to reduce the probability of receiving an activity suggestion when there is evidence of a decreasing effect. The conjecture is that intervention effects will stop decaying if the probability of delivering an activity suggestion to a participant is decreased whenever this participant is showing evidence of a decreasing effect. This algorithm is also using location as a moderator.

**Internal and External Validity in MRTs**

Internal validity concerns the ability of the MRT to provide evidence for attributing the estimated effects to the manipulation of the intervention component and not some systematic error (Jüni, Altman, & Egger, 2001). It is well known that in a two-arm randomized controlled trial, internal validity is harmed if the randomization, by chance, did not achieve balance in *baseline* covariates between the two arms. One way to check for deviations that indicate a lack of internal validity is to check whether the distribution of the baseline variables is dissimilar across the two arms. For the MRT, because the randomization occurs sequentially over time, to check internal validity one can check for balance in any covariates occurring prior to each decision point. In the HeartSteps example, one can check if, among available users at decision point ­, whether the fraction of available users who are at home () among all available users randomized to an activity suggestion () is close to the fraction of available users who are at home () among all available users randomized to no activity suggestion (). Other time-varying variables (’s) observed prior to decision point besides location might be considered as well. Because the causal excursion effect is only defined for individuals that are currently available—that is, one aims to estimate causal effects only among those who are available at decision point —these checks concern only available individuals[[8]](#footnote-8).

External validity concerns the extent to which the estimated causal excursion effect in the MRT provides a basis for generalization to a target population (Jüni et al., 2001). As is well known, in randomized controlled trials we try to enhance external validity by striving to enroll participants that are representative of the target population. The same considerations hold in an MRT. One way to assess the lack of external validity (to a target population) is to check whether the distribution of the baseline variables is dissimilar to that in the target population. If some baseline variables are likely prognostic for the outcome or predictive for the causal excursion effect, then a distributional imbalance in these variables between the target population and the MRT sample raises concerns that the causal excursion effect estimated from the MRT might not generalize to the target population. If we do not detect such imbalances, then we have greater confidence in the generalizability of the estimated causal excursion effect. In addition to the proximal outcome, , an MRT can involve other outcomes such as availability, , and the potential moderators, . Therefore, any baseline variable that might be related to any of the outcomes should be considered in checking for the aforementioned imbalance.

The “excursion” aspect of the causal excursion effect is another important aspect when considering generalizability of the findings. The excursion aspect explicitly acknowledges that, prior to decision point , the individual was provided a particular treatment schedule as used in the MRT (rather than some other fixed treatment assignment); the interpretation of the causal excursion effect is *the causal effect concerning excursions from the existing treatment schedule*. In the case of the HeartSteps MRT, the existing treatment schedule is “deliver activity suggestion with probability 0.6, if user is available at the decision point” and the excursion effect is a contrast between sending activity message now vs. not sending activity message now, *assuming the user had experienced the existing treatment schedule up to now*. The excursion aspect makes it overt that the comparison of two excursions at time might depend on how treatments were assigned prior to that time, which in turn depends on the treatment schedule of the particular MRT. Therefore, the causal excursion effect estimated from an MRT with a particular treatment schedule may differ from the causal excursion effect estimated from an MRT with a different treatment schedule. That being said, recall that the main goal of an MRT is to inform intervention development by identifying ways to improve the *existing* treatment schedule (see the subsection “Using moderation effect analysis to inform intervention development” in Discussion section), and focusing on the causal excursion effect allows us to do exactly that.

**Connection to MOST**

As discussed in the companion paper (Walton, et al., submitted) the MRT fits naturally within the optimization phase of the multiphase optimization strategy (MOST). In this phase, the investigator conducts a randomized experiment aimed at gathering scientific information needed to construct a new optimized intervention or to optimize an existing intervention. Typically, this includes inference for causal effects of individual intervention components and well as moderation analyses. This paper provides the definition of as well as inferential methods for the causal effects for use of an MRT in the optimization phase.

Furthermore, inference concerning causal excursion effects fits naturally within the conceptual framework of MOST. In this framework optimization is an ongoing process of intervention improvement, in which each optimization trial provides information useful in generating hypotheses about how to improve the intervention. For example, the following question can be characterized by the causal excursion effect: If we were to alter the treatment schedule for the activity suggestions based on knowledge of the current location, would this improve subsequent 30-minute step count? In mobile health this inferential goal makes sense even in implementation as the team must continually monitor and update the mobile application software. Similarly, continually monitoring performance and assessing how to best improve the current schedule for assigning treatments is natural. The causal excursion effect is useful for this purpose.

**Sample size considerations**

Liao et al. (2016) provides theory for determining the sample size for the setting in which the moderator is exogenous (for example, time since the participant started the intervention). A web applet can be found at <https://methodologycenter.shinyapps.io/mrt_ss/>. The applet takes as input duration of the study, number of decision points per day, expected availability pattern, randomization probability (which can be time-varying), and target proximal treatment effect to be detected, and the desired power, and it outputs the required sample size (or vice versa; input sample size and output power). Alternately, R code is freely available at <https://cran.r-project.org/web/packages/MRTSampleSize/index.html>.

**Additional types of causal effects**

This paper focuses on the immediate causal excursion effect (“immediate” in the sense that there is no other treatment between the decision point and the proximal outcome ) of a time-varying digital intervention. One may also be interested in inference about a delayed causal excursion effect. For example, when assessing the effect of the planning support component, it may be of interest to assess the effect of a planning support prompt on the total step count over the next *x* days, as it would be desirable for the delivery of a planning prompt to have a longer-term effect such as forming a habit. The generalization of WCLS to assess such delayed effects is given in Boruvka et al. (2018). Here we note only that the interpretation of such delayed causal excursion effects averages over, in addition to the history information observed up to that decision point, future treatments and future covariates. Suppose we are interested in the effect of a planning support prompt on the total step count over the next *x* days. This effect would be marginal with respect to the schedule for delivering planning support during the subsequent *x* days.

Other more familiar causal effects might be estimated if one is comfortable with additional assumptions. For example, suppose we can safely assume that the treatments prior to the current decision point will not impact subsequent outcomes (i.e., these prior treatments do not have delayed positive or negative effects). Then the potential outcomes such as are actually and we might focus on inference for the effect, . In terms of the primary analysis of data from an MRT, we opt to make inference about causal excursion effects both due to interpretation in the above continual learning paradigm as well as the minimal causal inference assumptions required. Of course, in secondary and hypothesis generating analyses a variety of statistical assumptions would be made to conduct inference about other causal effects.

**Other types of analyses**

Generalized estimating equations (GEE; Liang & Zeger, 1986) and multi-level models (MLM; Laird & Ware, 1982; Raudenbush & Bryk, 2002) have been used with great success to analyze data from intensive longitudinal studies, and thus appear to be a natural choice for conducting primary and secondary data analysis for MRTs. However, these methods can result in biased causal effect estimates for MRTs when there are endogenous time-varying covariates, i.e., covariates that can depend on previous outcomes or previous treatments. For example, in HeartSteps, the prior 30-minute step count is likely impacted by prior treatment and is thus endogenous. We illustrate this bias in Appendix A.

**Limitations and Future Directions**

The WCLS method presented in this paper is applicable to the case where the proximal outcome is continuous. When the proximal outcome is binary, the WCLS provides an estimator of the causal effect on the risk difference scale. Qian, Yoo, Klasnja, Almirall, & Murphy (2019) provides an alternative method for analyzing MRT data with a binary outcome that yields causal effect estimates on the relative risk scale. Methods for other types of proximal outcomes (zero-inflated outcomes, categorical, ordinal, and longitudinal) require development. In particular, in HeartSteps, the 30-minute step count is 0 for about 30% of decision points, which means the distribution of the proximal outcome is zero-inflated. In this case it is natural to estimate the causal effect in other ways, in particular to estimate two sets of parameters—parameters that are related to the zero-inflation as well as parameters in the continuous or count part of the outcome distribution; work in this area is needed as well.

**Appendix A**

**GEE and MLM Can Be Biased When Estimating Causal Excursion Effects in MRTs**

MRTs produce intensive longitudinal data (Schafer, 2006), as individuals are randomized among intervention options repeatedly during the MRT, and outcomes and covariates are assessed in tandem with randomization. Repeated measurement of the same individuals over time means that the repeated observations are likely dependent. *Generalized estimating equations* (GEE; Liang & Zeger, 1986) and *multi-level models* (MLM; Laird & Ware, 1982; Raudenbush & Bryk, 2002), the latter also known as mixed models or random effects models, have had great success in analyzing longitudinal studies. However, as we illustrate below, inappropriate application of them to MRT data may result in biased estimates of the causal excursion effects when *endogenous time-varying covariates* are included in the model. A time-varying covariate is *endogenous* if it can depend on previous outcomes or previous treatments, which commonly occurs in MRTs. For example, in analyzing the effect of activity suggestion in the subsequent 30-minute step count in HeartSteps, one may want to control for the 30-minute step count prior to each decision point to reduce noise. Because the 30-minute step count prior to a decision point can be correlated with past step counts (i.e., past outcomes), it is endogenous. When a time-varying covariate is not endogenous, it is called *exogenous*. Examples of *exogenous time-varying covariates* include time, weather, and anything that cannot be impacted by previous treatments or previous outcomes.

**Inappropriate use of GEE and MLM can result in biased causal excursion effect estimates in the presence of endogenous time-varying covariates.** Pepe & Anderson (1994) demonstrated that in the presence of endogenous time-varying covariates, parameter estimates from GEE may be biased unless certain conditions, described below, are met. Such bias is also shown in subsequent research through simulation studies and analytic calculations (Diggle et al., 2002; Pan, Louis, & Connett, 2000; Schildcrout & Heagerty, 2005; Tchetgen, Glymour, Weuve, & Robins, 2012; Vansteelandt, 2007). For completeness we provide a brief explanation of the bias here. Consider a simplified version of HeartSteps MRT where there are two decision points for each individual and individuals are always available. Suppose the observed data for individual *i* is , where denotes the 30-minute step count prior to decision point (an endogenous time-varying covariate), is the indicator of whether an activity suggestion is delivered at decision point (so has .6 probability to be 1), and is the 30-minute step count following decision point *t*. We choose in equation (2); i.e., we want to assess whether the effect of the activity suggestion is moderated by the prior 30-minute step count. The data analyst may choose to impose the following linear model on the mean of the proximal outcome given the treatment and the covariate at decision point :

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

and use GEE to estimate the coefficients . Often a non-independent working correlation structure is used in GEE, aiming for efficiency gain (i.e., smaller standard error of the estimated coefficients compared to GEE with working independence correlation structure).

It is well known that GEE produces unbiased estimates regardless of the choice of the working correlation structure, as long as equation (7) holds; however, this is only true when all covariates are exogenous. In this above example with two decision points, Pepe & Anderson (1994) demonstrated that to guarantee the unbiasedness of the GEE estimates one of the following conditions needs to hold:

1. for ; or
2. a working independence correlation structure is used.

Condition (i) is usually violated when is endogenous: In this particular example, can be correlated with , so that . This means unless the independent working correlation structure is used, GEE can produce biased estimates even if equation (7) holds.

The same bias can occur when MLM is used instead of GEE. In general, for each MLM there is a corresponding GEE with a non-independent correlation structure that produces the same estimated coefficients. For example, an MLM resembling equation (7) is , where is a random intercept and is the error term. This corresponds to a GEE with compound symmetric (also called exchangeable) working correlation structure. Given this equivalency, MLM can produce biased estimates if the covariate is endogenous.

**A few scenarios where GEE or MLM provides unbiased causal excursion effect estimates from MRT data.** GEE builds upon a marginal mean model (i.e., the relationship between the mean of the proximal outcome and the covariates and the treatment assignments such as (7)). If no endogenous time-varying covariates are included in the model, the individuals are always available, and the randomization probability is constant, GEE with any working correlation structure gives unbiased estimates as long as the marginal mean model is correct. If there are endogenous time-varying covariates in the model, the individuals are always available, and the randomization probability is constant, GEE with independent working correlation structure still gives unbiased estimates as long as the marginal mean model is correct but GEE with other working correlation structure does not.

Because an MLM always correspond to a GEE with some non-independent working correlation structure, MLM provides unbiased causal excursion effect estimates if no endogenous time-varying covariates are included in the model, the individuals are always available, and the randomization probability is constant. However, although the estimated coefficients from an MLM will generally be biased for the causal excursion effect when there are endogenous time-varying covariates, those estimated coefficients can have a different, individual-specific, interpretation under a rather strong assumption. As shown in Qian, Klasnja, & Murphy (2019), if the endogenous time-varying covariates can be safely assumed to only depend on the random effect through the observed previous outcomes and previous covariates, then the fitted results from standard linear mixed models can be interpreted as a causal effect that is conditional on the random effect (i.e., individual-specific rather than population-average) and conditional on the entire history (rather than conditional only on ). An example where this strong assumption holds is when the endogenous time-varying covariates are previous proximal outcomes (e.g., the endogenous time-varying covariate at decision point is the proximal outcome at decision point ).

**A mathematical demonstration of the bias from inappropriate application of GEE when there are endogenous time-varying covariates.** For illustration clarity we consider the case where each participant is in the MRT for two decision points. The data for the *i*-th participant is , where is the covariate, is the treatment assignment, and is the continuous outcome. The covariate is endogenous time-varying, in the sense that it can depend on previous treatment and previous outcome.

The model on the marginal mean of is . The corresponding GEE solves the following estimating equation:

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Here, denotes the number of participants, and is a working covariance matrix. Examples of include the following:

* Working independence:
* Compound symmetry:
* Autoregressive (in the special case of two decision points, autoregressive is the same as compound symmetry):

In this setting, the result in Pepe & Anderson (1994) implies that GEE is guaranteed to produce unbiased if either

1. for , or
2. a working independence correlation structure is used,

and they provided simulation results to show that GEE can produce biased estimates when neither conditions hold. In the following, we rephrase the intuitive argument given in Pepe and Anderson (1994) in this particular setting to show why GEE can be biased if neither condition holds.

We write , and write the residual . A summand (for fixed ) in equation (8) becomes

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

Because , we have

Therefore, all the terms with and (such as ; i.e., terms that are multiplied with the diagonal elements of ) in (9) have expectation zero, and what is left is the terms with and (i.e., terms that are multiplied with the off-diagonal elements of ). In other words, the expectation of (9) equals

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

Mathematical theory for GEE tells us that GEE outputs unbiased when (9) has expectation zero; i.e., when (10) equals zero.

If condition (i) holds, we have , and similarly . Therefore, (10) equals 0 with any choice of , and GEE estimators are unbiased.

If condition (ii) holds, we have . Hence (10) equals 0 and GEE estimators are unbiased.

When neither condition holds, it’s likely that (10) does not equal to zero. For example, suppose . Then the term equals

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

which is the residual multiplied with the outcome itself. Because the residual and the outcome at the same time point is correlated, (11) likely does not equal to zero. Therefore, (10) likely does not equal to zero. This means GEE can be biased when neither conditions hold, i.e., when endogenous time-varying covariates are included and non-independent working correlation structure is used.

**Appendix B**

**A general form of the WCLS estimator for the causal excursion effect that allows the randomization probability to vary over time**

We assume a linear model for the causal excursion effect: . Suppose is a working model for the conditional mean of given no treatment at decision point t and history , . Note that the unbiasedness of the estimator for does not require to be a correct model for . We use to denote the randomization probability at decision point , which may possibly depend on .

The WCLS estimator for is calculated as follows. Suppose is the value that solves the following estimating equation

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

then is the WCLS estimator for . is an arbitrary probability as long as it depends on through at most and it is bounded away from 0 and 1; is the index for the *i*th individual and is defined as

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

, the ratio of two probabilities, serves as a change of probability: It makes it as if the treatment is randomized with probability . It is used to marginalize the causal excursion effect over variables in but not in . As long as depends on through at most and it is bounded away from 0 and 1, the particular choice of doesn’t affect the unbiasedness of . For instance, one can set it to be 0.5 (or any constant between 0 and 1) for all individuals and all decision points, or set it to be the predicted value from a logistic regression fit of . If the true randomization probability depends at most on , then one can also set to be equal to the true randomization probability, in which case (12) is mathematically equivalent to (5). can impact the standard error of . In addition, when the causal excursion effect model is misspecified, impacts the limit of . See the Appendix of Boruvka et al. (2018) for more technical details on how the limit of is impacted by in this case.

Now we present a way to obtain the general WCLS estimator for time-varying randomization probability through standard statistical software that implements GEE. Suppose the assumed causal excursion effect model is (6), the working model for is , then the WCLS estimator and its standard error can be obtained by (i) incorporate as the “prior weights”, (ii) choose a working independence correlation structure, and (iii) fit GEE with dependent variable and independent variables and . Then the estimated coefficient for is the WCLS estimate .

**Standard error formula for WCLS.**

Below we provide the formula for the standard error of the WCLS estimator . For that solves estimating equation (12), their variance can be estimated by

where

and

Here, denotes sample average over *n* individuals. The standard error formula can be modified for the setting where the randomization probability is constant over time (i.e., the setting in the main paper) by letting and .

**Appendix C**

We conduct a simulation study to illustrate the claim that including variables that are correlated with in may reduce the variance of the WCLS estimator. The generative model mimics features of the HeartSteps data and is set up as follows. For simplicity we assume users are always available. At decision point , the covariate is drawn from the empirical distribution of the log-transformed 30-minute step count preceding a decision point in the HeartSteps data. For simplicity is generated independently of previous outcomes and treatments. The treatment is generated from a Bernoulli distribution with .6 success probability; this mimics the .6 randomization probability of activity suggestions in HeartSteps. The proximal outcome is generated from a Gaussian distribution with mean

and standard deviation 2.716. The coefficients in the above display are the estimated coefficients from a WCLS fit on the HeartSteps data with the same control variables and constant treatment effect model. The standard deviation is the empirical standard deviation of the residual in from the above WCLS fit. As in the HeartSteps data set, for each simulated trial we generate 37 individuals each with 210 decision points.

For each data set generated from the above generative model, we consider four WCLS fits for the true treatment effect 0.1229 and compare their performance. All four WCLS assumes the constant treatment effect model, and they differ in the choice of the working model. The first WCLS fit (WCLS-1) includes control variables ; the second WCLS fit (WCLS-2) includes control variables ; the third WCLS fit (WCLS-3) includes control variables ; and the fourth WCLS fit (WCLS-4) includes only the intercept. The bias, standard deviation (SD), and coverage probability of 95% confidence interval (CP) are listed in Supplementary Table 1. All four WCLS estimators are unbiased with nominal confidence interval coverage, because their assumed constant treatment effect model holds under this generative model. (This again illustrates the unbiasedness of the WCLS estimator does not require the control part of the model to be correct.) On the other hand, the choice of working model affects the efficiency of the estimator. In particular, WCLS-1 and WCLS-2 have smaller standard error than WCLS-3 and WCLS-4, because the former two includes , a covariate that is highly correlated with the proximal outcome .

**Reference**

Boruvka, A., Almirall, D., Witkiewitz, K., & Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, *113*(523), 1112–1121.

Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D., & Mamykina, L. (2015). No longer wearing: Investigating the abandonment of personal health-Tracking technologies on craigslist. *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 647–658. https://doi.org/10.1145/2750858.2807554

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of Experiments With Multiple Independent Variables: A Resource Management Perspective on Complete and Reduced Factorial Designs. *Psychological Methods*, *14*(3), 202–224. https://doi.org/10.1037/a0015826

Dempsey, W., Liao, P., Kumar, S., & Murphy, S. A. (2019). The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments. *ArXiv: 1711.03587*.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., & others. (2002). *Analysis of longitudinal data*. Oxford University Press.

Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, *7*(1), 1–11. https://doi.org/10.2196/jmir.7.1.e11

Ho, J., & Intille, S. S. (2005). Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. *CHI 2005: Technology, Safety, Community: Conference Proceedings - Conference on Human Factors in Computing Systems*, 909–918.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842. https://doi.org/10.1198/016214508000000292

IBM Corp. (2019). IBM SPSS Statistics for Windows.

Jüni, P., Altman, D., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *BMJ*, *323*(7303), 42–46.

Klasnja, P., Harrison, B. L., Legrand, L., Lamarca, A., Froehlich, J., & Hudson, S. E. (2008). Using wearable sensors and real time inference to understand human recall of routine activities. *UbiComp 2008 - Proceedings of the 10th International Conference on Ubiquitous Computing*, 154–163. https://doi.org/10.1145/1409635.1409656

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, *34*(S), 1220.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., … Murphy, S. A. (2018). Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine : A Publication of the Society of Behavioral Medicine*, *53*(6), 573–582. https://doi.org/10.1093/abm/kay067

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Liao, P., Klasnja, P., Tewari, A., & Murphy, S. A. (2016). Sample size calculations for micro-randomized trials in mHealth. *Statistics in Medicine*, *35*(12), 1944–1971.

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, *57*(1), 126–134. https://doi.org/10.1111/j.0006-341X.2001.00126.x

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2017). Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462.

Pan, W., Louis, T. A., & Connett, J. E. (2000). A Note on Marginal Linear Regression with Correlated Response Data. *American Statistician*, *54*(3), 191–195. https://doi.org/10.1080/00031305.2000.10474544

Pepe, M. S., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation*, *23*(4), 939–951.

Qian, T., Klasnja, P., & Murphy, S. A. (2019). Linear mixed models under endogeneity: modeling sequential treatment effects with application to a mobile health study. *Statistical Science*.

Qian, T., Yoo, H., Klasnja, P., Almirall, D., & Murphy, S. A. (2019). Estimating Time-Varying Causal Excursion Effect in Mobile Health with Binary Outcomes. *ArXiv: 1906.00528*.

R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., … Thompson, R. F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*, *92*(2), 135–138. https://doi.org/10.1016/j.nlm.2008.09.012

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9–12), 1393–1512.

Robins, J. M. (1987). Addendum to “a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect.” *Computers & Mathematics with Applications*, *14*(9–12), 923–945.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, *23*(8), 2417–2421. https://doi.org/10.1080/03610929408831394

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–58.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

SAS Institute Inc. (2019). SAS/STAT Software, Version 9.4. Cary, NC. Retrieved from http://www.sas.com/

Schafer, T. A. W. J. L. (2006). *Models for intensive longitudinal data*. Oxford University Press.

Schildcrout, J. S., & Heagerty, P. J. (2005). Regression analysis of longitudinal binary data with time-dependent environmental covariates: Bias and efficiency. *Biostatistics*, *6*(4), 633–652. https://doi.org/10.1093/biostatistics/kxi033

Shaw, R. J., Bosworth, H. B., Hess, J. C., Silva, S. G., Lipkus, I. M., Davis, L. L., & Johnson, C. M. (2013). Development of a theoretically driven mHealth text messaging application for sustaining recent weight loss. *Journal of Medical Internet Research*, *15*(5). https://doi.org/10.2196/mhealth.2343

StataCorp. (2019). Stata Statistical Software: Release 16.

Tchetgen, E. J. T., Glymour, M. M., Weuve, J., & Robins, J. (2012). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. *Epidemiology*, *23*(4), 644–646. https://doi.org/10.1097/EDE.0b013e31825727b5

Vansteelandt, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scandinavian Journal of Statistics*, *34*(3), 478–498. https://doi.org/10.1111/j.1467-9469.2006.00555.x

Walton, A., Collins, L. M., Corsby, L., Klasnja, P., Nahum-Shani, I., Rabbi, M., … Murphy, S. A. (2019). The Micro-Randomized Trial for Developing Mobile Health Interventions.

Yardley, L., Spring, B. J., Riper, H., Morrison, L. G., Crane, D. H., Curtis, K., … Blandford, A. (2016). Understanding and Promoting Effective Engagement With Digital Behavior Change Interventions. *American Journal of Preventive Medicine*, *51*(5), 833–842. https://doi.org/10.1016/j.amepre.2016.06.015

Table 1.

*Estimated main effect of activity suggestions on proximal outcome*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable |  | Estimate | 95% LCL | 95% UCL | SE | Hotelling *t* | *p* |
| Intercept |  | 1.783 | 1.537 | 2.029 | 0.121 | 217.3 | <0.001 |
| Past 30-min step count |  | 0.414 | 0.351 | 0.476 | 0.031 | 181.2 | <0.001 |
| Activity Suggestion |  | 0.131 | -0.006 | 0.268 | 0.067 | 3.79 | 0.060 |

*Note.* LCL (UCL) represents lower (upper) confidence limit. SE represents standard error. LCL, UCL, SE, and *p* are corrected for small sample size using method in (Liao et al., 2016; Mancl & DeRouen, 2001). The degrees-of-freedom for the Hotelling *t* test is (1, 34).

Table 2.

*Estimated effect of activity suggestion on proximal outcome as a linear function of time in study*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable |  | Estimate | 95% LCL | 95% UCL | SE | Hotelling *t* | *p* |
| Intercept |  | 2.003 | 1.765 | 2.240 | 0.117 | 294.7 | <0.001 |
| Past 30-minute step count |  | 0.412 | 0.351 | 0.473 | 0.030 | 189.6 | <0.001 |
| Time (in days) |  | -0.011 | -0.020 | -0.001 | 0.005 | 5.09 | 0.031 |
| Activity Suggestion |  | 0.507 | 0.201 | 0.814 | 0.151 | 11.37 | 0.002 |
| Activity Suggestion x Time (in days) |  | -0.018 | -0.031 | -0.006 | 0.006 | 9.19 | 0.005 |

*Note.* LCL (UCL) represents lower (upper) confidence limit. SE represents standard error. LCL, UCL, SE, and *p* are corrected for small sample size using method in (Liao et al., 2016; Mancl & DeRouen, 2001). The degrees-of-freedom for the Hotelling *t* test is (1, 32).

Table 3.

*Estimated effect of walking suggestion / anti-sedentary suggestion on proximal outcome, moderated by location (home/work or other) during decision point*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable |  | Estimate | 95% LCL | 95% UCL | SE | Hotelling *t* | *p* |
| Intercept |  | 1.715 | 1.461 | 1.968 | 0.124 | 191.3 | <0.001 |
| Past 30-minute step count |  | 0.414 | 0.351 | 0.477 | 0.031 | 182.0 | <0.001 |
| At home/work |  | 0.143 | -0.083 | 0.368 | 0.110 | 1.67 | 0.205 |
| Walking Suggestion |  | 0.050 | -0.167 | 0.267 | 0.106 | 0.22 | 0.640 |
| Walking Suggestion x At home/work |  | 0.377 | 0.001 | 0.753 | 0.184 | 4.18 | 0.049 |
| Anti-sedentary Suggestion |  | 0.092 | -0.166 | 0.351 | 0.127 | 0.53 | 0.472 |
| Anti-sedentary Suggestion x At home/work |  | -0.142 | -0.540 | 0.256 | 0.195 | 0.53 | 0.472 |

*Note.* LCL (UCL) represents lower (upper) confidence limit. SE represents standard error. LCL, UCL, SE, and *p* are corrected for small sample size using method in (Liao et al., 2016; Mancl & DeRouen, 2001). The degrees-of-freedom for the Hotelling *t* test is (1, 30).

Table 4.

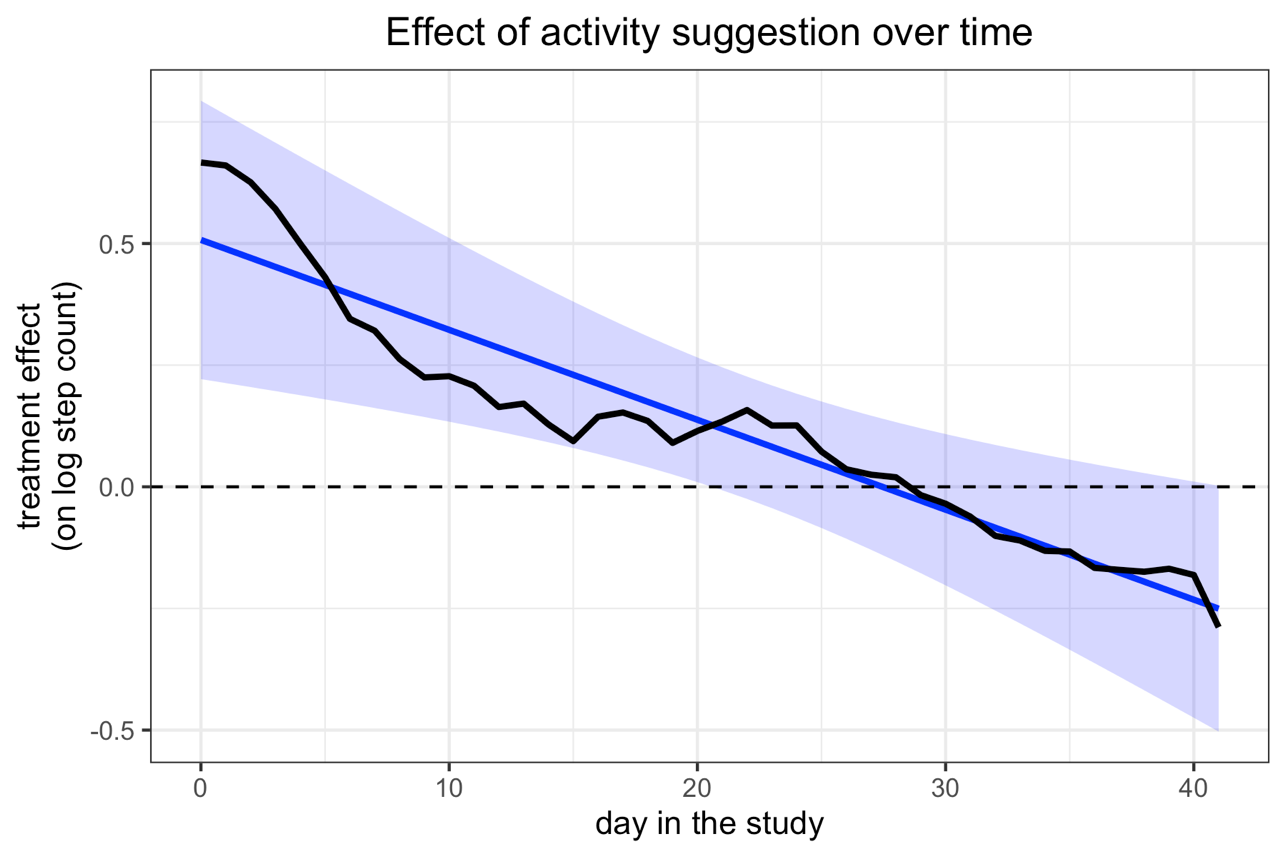
*Estimated effect of activity suggestion on proximal outcome, moderated by whether activity planning support was received on previous night*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable |  | Estimate | 95% LCL | 95% UCL | SE | Hotelling *t* | *p* |
| Intercept |  | 1.764 | 1.511 | 2.017 | 0.124 | 201.3 | <0.001 |
| Past 30-minute step count |  | 0.414 | 0.351 | 0.476 | 0.031 | 180.5 | <0.001 |
| Planning on previous day |  | 0.050 | -0.106 | 0.205 | 0.076 | 0.43 | 0.518 |
| Activity Suggestion |  | 0.113 | -0.035 | 0.261 | 0.073 | 2.43 | 0.129 |
| Activity Suggestion x Planning on previous day |  | 0.046 | -0.228 | 0.320 | 0.134 | 0.12 | 0.734 |

*Note.* LCL (UCL) represents lower (upper) confidence limit. SE represents standard error. LCL, UCL, SE, and *p* are corrected for small sample size using method in (Liao et al., 2016; Mancl & DeRouen, 2001). The degrees-of-freedom for the Hotelling *t* test is (1, 32).

Figure 1.

*Estimated effect of activity suggestion on proximal outcome as a linear function of days in study, and corresponding 95% pointwise confidence intervals*



*Note.* Figure for “Question 2: Does the effect of the activity suggestions change with each additional day in the study?” in section “Analysis Using Data from HeartSteps MRT.” The black curve is the estimated effect using local 2-degree polynomial regression with smoothing span 2/3 and tricubic weighting. The blue line represents the estimated causal excursion effect across the 42 study days assuming a linear time trend, and the shaded blue area is the pointwise 95% confidence interval.

Supplementary Table 1.

*Simulation results for Appendix C: Efficiency gain from including prognostic variable in the working model*

|  |  |  |  |
| --- | --- | --- | --- |
|  | bias | standard deviation | 95% coverage probability |
| WCLS-1 | -0.001 | 0.067 | 96.7% |
| WCLS-2 | -0.001 | 0.067 | 96.9% |
| WCLS-3 | -0.001 | 0.074 | 95.8% |
| WCLS-4 | -0.001 | 0.074 | 95.7% |

*Note.* All four WCLS assumes the constant treatment effect model, and they differ in the choice of the working model. WCLS-1 includes control variables ; WCLS-2 includes control variables ; WCLS-3 includes control variables ; WCLS-4 includes only the intercept.

1. The contextually tailored activity suggestion at each time is delivered in one of two forms (with equal probability), either a suggestion with a walking activity that took 2-5 minutes to complete, or an anti-sedentary suggestion that instructs brief movements such as to stand up and roll one’s arms; see Walton, et al. (submitted). For expositional simplicity we group all activity suggestions together for most parts of the paper. [↑](#footnote-ref-1)
2. An individual is sent a planning prompt with probability .5 every day. If a planning prompt is sent, it is delivered in one of two forms (with equal probability): structured planning (where the participant is prompted to select a plan from a list of their own past activity plans), or unstructured planning (where the participant is prompted to type their plan into a text box). For expositional simplicity we group both forms of planning together in this paper. [↑](#footnote-ref-2)
3. This equality holds under the consistency assumption often made in causal inference literature, which essentially requires that there are no two “versions” of the same treatment. In the example of activity suggestions, in order to properly define “delivering an activity suggestion” as treatment 1 and “not delivering an activity suggestion” as treatment 0, one would consider various framings and various contents of the suggestions as a “compound treatment”. However, if one wishes to distinguish between the effect of different versions of the suggestions in the analysis, then they would need to instead define the treatment to have multiple levels. [↑](#footnote-ref-3)
4. For simplicity, we omit the subscript *i* for the *i*th individual in and in all other variables unless necessary. [↑](#footnote-ref-4)
5. Note that the overbar in does not stands for the average; rather, it stands for the entire vector of treatment assignment and similarly for . This notation is common in causal inference literature (e.g., Robins, 1986, 1987). [↑](#footnote-ref-5)
6. This assumption is called “non-interference” in causal inference. If there are social network components in the digital intervention, this assumption may be violated and an extension of the potential outcomes framework to incorporate interference is needed (Hong & Raudenbush, 2006; Hudgens & Halloran, 2008). [↑](#footnote-ref-6)
7. Readers who are familiar with factorial experiments may wish to note that we are treating experimental treatments delivered prior to time as potential moderators. [↑](#footnote-ref-7)
8. Note that this only applies to MRTs with constant randomization probability (such as the HeartSteps MRT). For MRTs where the randomization probability may change depending on the individual’s history information, the aforementioned covariate imbalance may no longer be indicative of lack of internal validity. [↑](#footnote-ref-8)