Assessing Time-Varying Causal Effect Moderation in Mobile Health

Audrey Boruvka[1], Daniel Almirall[2], Katie Witkiewitz[3], and Susan A. Murphy[1,2]

[1]Department of Statistics, University of Michigan
[2]Institute for Social Research, University of Michigan
[3]Department of Psychology, University of New Mexico

Abstract

In mobile health interventions aimed at behavior change and maintenance, treatments are provided in real time to manage current or impending high risk situations or promote healthy behaviors in near real time. Currently there is great scientific interest in developing data analysis approaches to guide the development of mobile interventions. In particular data from mobile health studies might be used to examine effect moderators—individual characteristics, time-varying context or past treatment response that moderate the effect of current treatment on a subsequent response. This paper introduces a formal definition for moderated effects in terms of potential outcomes, a definition that is particularly suited to mobile interventions, where treatment occasions are numerous, individuals are not always available for treatment, and potential moderators might be influenced by past treatment. Methods for estimating moderated effects are developed and compared. The proposed approach is illustrated using BASICS-Mobile, a smartphone-based intervention designed to curb heavy drinking and smoking among college students.

Keywords: mHealth, structural nested mean model, effect modification

1 Introduction

Mobile health (mHealth) broadly refers to the practice of healthcare using mobile devices, such as smartphones and wearable sensors both to deliver treatment as well as to sense the current context of the individual. In mobile interventions for behavior maintenance or change, treatments are typically designed to help individuals manage high risk situations or promote healthy behaviors. Examples include medication reminders, motivational messages, physical activity suggestions, cognitive exercises to help manage stress or other risky situations, and prompts to facilitate activity in support networks.

There is intense interest in data analysis approaches to guide the development of mobile interventions (Free et al. 2013; Muessig et al. 2013) and to test the dynamic behavioral theories on which these interventions are based (Spring et al. 2013; Mohr et al. 2014). Micro-randomized trials (MRTs; Klasnja et al. 2015; Liao et al. 2015; Dempsey et al. 2015) provide data expressly for this purpose, with each participant in an MRT sequentially randomized to treatment numerous times, at possibly 100s to 1000s of occasions. In both MRTs and observational mHealth studies both treatment and measurement occur intensively over time. Measurements on individual characteristics, context and response to treatments are collected passively through sensors or actively by self-report.

One way in which these data may aid the design of a mobile intervention is through the examination of effect moderation; that is, inference about which factors strengthen or weaken the response to treatments. Consider, for example, an intervention for smoking cessation. Mindfulness-based treatments to help individuals manage their urge to smoke are presumably best delivered at times when there exists an inclination to smoke (e.g. Witkiewitz et al. 2014). However other factors might influence the effect of these treatments on subsequent smoking rate. For example it may be that the mindfulness-based approach reduces smoking only when stress levels or self-regulatory demands are low, and has little to no effect otherwise. In general knowledge about moderators can be used to deliver treatments only in settings where they have proven most efficacious or to identify alternative treatment strategies when the treatment shows little to no benefit. Treatment effects might also evolve over the course of the intervention, so functions of time could also be examined as possible moderators.

This paper provides two main contributions in the assessment of treatment effects from longitudinal data in which treatment, response, and potential moderators are time-varying. The first is a definition for treatment effects that is particularly suited for mHealth, where treatment occasions are numerous and potential moderators might be influenced by past treatment. These effects are a generalization of the treatment "blips" in the structural

nested mean model (SNMM; Robins 1989, 1994, 1997). The second is a weighted least squares method for estimating these treatment effects, conditional on a few select variables representing potential moderators of interest.

We begin by defining treatment effects in our setting. The aforementioned weighted least squares method and two alternatives are derived and assessed numerically using a variety of simulation scenarios. As an illustration, we apply the proposed approach to data from a study of BASICS-Mobile, a mobile intervention for heavy drinking and smoking among college students (Witkiewitz et al. 2014).

2 Proximal and Other Lagged Treatment Effects

2.1 Motivating Example

Our motivating example is drawn from BASICS-Mobile, a smartphone-based intervention designed to reduce heavy drinking and smoking among college students. Users are prompted three times per day (morning, afternoon and evening) to complete a self-report assessing a variety of individual and contextual factors including episodes of drinking or smoking, social settings, affect, and need to self-regulate thoughts. The afternoon and evening self-reports are possibly followed by a treatment module of three to four screens of information and at least one question to confirm that the module was received. Some of the treatment modules address smoking and heavy drinking using mindfulness messages (Bowen and Marlatt 2009). Other modules provide general (primarily health-related) information (Dimeff 1999). In an analysis of data arising from the implementation of BASICS-Mobile, it is natural to estimate the effect of the mindfulness messages (versus providing general health information) on a proximal response, such as the smoking rate between the current and following self-report, and to assess whether or not these effects differ according to the individual's context.

2.2 Notation and Data

For a given individual, let A_t denote the treatment at the tth treatment occasion and Y_{t+1} be the subsequent proximal response ($t = 1, \ldots, T$). Throughout we limit attention to the case where each A_t is binary and Y_{t+1} is continuous. Individual and contextual information at the tth treatment occasion is represented by X_t, which may contain summaries of previous measurements of context, treatment or response. For example, prior to each treatment occasion the individual might report their current mood. The vector X_t could then contain this

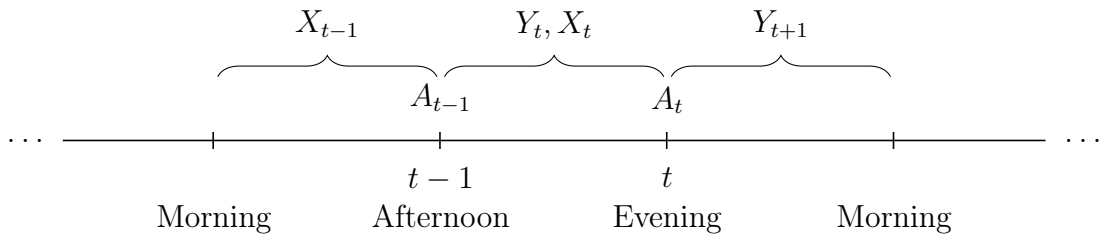

Figure 1: A BASICS-Mobile participant's data for two treatment occasions leading up to Y_{t+1}, depicted in chronological order. Information is primarily collected via self-reports three times per day—morning, afternoon and evening. Treatment occasions take place after the afternoon and evening self-reports.

measurement or, with previous measurements, variation or change in mood. Over the course of T treatment occasions, the resulting data from an individual ordered in time is $(X_1, A_1, Y_2, \ldots, X_T, A_T, Y_{T+1})$. The overbar is used to denote a sequence of random variables or realized values through a specific treatment occasion; for example $\bar{A}_t = (A_1, \ldots, A_t)$. Information accrued up to treatment occasion t is represented by the history $H_t = (\bar{X}_t, \bar{Y}_t, \bar{A}_{t-1})$.

In BASICS-Mobile (Fig. 1), $A_t = 1$ if a mindfulness message is provided at the tth treatment occasion and $A_t = 0$ otherwise, Y_{t+1} is the smoking rate between the occasion t self-report prompt and the following self-report prompt, $T = 28$ and X_t includes the time of day, number of reports recently completed, prior smoking rate, current need to self-regulate, and other summary variables formed from the reports up to and including the tth occasion. For example, from the self-reports at $t-1$ and t, we can examine the change in self-regulation needs and determine whether there was an increased need ($incr_t = 1$) or not ($incr_t = 0$).

To define treatment effects below, we adopt potential outcomes (Rubin 1974; Neyman 1990; Robins 1989) notation. However we will deviate slightly from this framework because, as will be seen below in (2), our estimands may involve the treatment distribution in the data. We represent random variables or vectors with uppercase letters; lowercase letters denote their realized values. In particular it will be useful to include in the set of potential outcomes, treatments expressed as potential outcomes of past treatment. That is, the potential outcomes are $\{Y_2(a_1), X_2(a_1), A_2(a_1)\}_{a_1 \in \{0,1\}}, \ldots, \{Y_T(\bar{a}_{T-1}), X_T(\bar{a}_{T-1}), A_T(\bar{a}_{T-1})\}_{\bar{a}_{T-1} \in \{0,1\}^{T-1}}$, $\{Y_{T+1}(\bar{a}_T)\}_{\bar{a}_T \in \{0,1\}^T}$. In BASICS-Mobile, for example, the smoking rate measured following the second treatment occasion has four potential outcomes: $Y_3(0,0)$, $Y_3(0,1)$, $Y_3(1,0)$, $Y_3(1,1)$. Here $Y_3(0,0)$ is the smoking rate that would arise for a given individual had they received no mindfulness treatments over the first two treatment occasions: $a_1 = a_2 = 0$. This idea can be similarly applied to the measurements X_t, since they might also be influenced by

past treatment; $X_{t+1}(\bar{a}_t)$ are the potential measurements had the sequence of treatments $\bar{a}_t$ been allocated. For brevity, we denote $A_2(A_1)$ by A_2 and so on with $A_t(\bar{A}_{t-1})$ denoted by A_t. Then $H_t(\bar{A}_{t-1}) = (X_1, A_1, Y_2(A_1), X_2(A_1), A_2, Y_3(\bar{A}_2), X_3(\bar{A}_2), A_3, \ldots, Y_t(\bar{A}_{t-1}), X_t(\bar{A}_{t-1}))$.

2.3 Treatment Effects

Many treatments are designed to influence an individual in the short term or proximally in time (Heron and Smyth 2010). For example, instruction in the mindfulness intervention used in BASICS-Mobile, called urge surfing, aims to help the individual to "ride out" urges, by recognizing the urge as it arises and allowing the urge to pass on its own. Questions related to these effects concern the proximal effect of treatment on the response defined by

$$\mathrm{E}\big[Y_{t+1}(\bar{A}_{t-1}, 1) - Y_{t+1}(\bar{A}_{t-1}, 0) \mid S_{1t}(\bar{A}_{t-1})\big], \tag{1}$$

where $S_{1t}(\bar{A}_{t-1})$ is a vector of summary variables chosen from $H_t(\bar{A}_{t-1})$. The difference in (1) represents the effect of $A_t = 1$ versus $A_t = 0$ on the response at $t+1$, given $S_{1t}(\bar{A}_{t-1})$. In conditioning only on $S_{1t}(\bar{A}_{t-1})$ as opposed to $H_t(\bar{A}_{t-1})$, the effect (1) is marginalized over variables in $H_t(\bar{A}_{t-1})$ that are not in $S_{1t}(\bar{A}_{t-1})$. Different choices of variables in S_{1t} address a variety of scientific questions, each of which is useful for understanding the effect of $A_t = 1$ versus $A_t = 0$ on the response Y_{t+1}. For example, a first analysis may focus on the proximal effect that is marginal over all variables in $H_t(\bar{A}_{t-1})$ (so that $S_{1t} = 1$), whereas a second analysis may focus on assessing this effect conditional on particular variables from $H_t(\bar{A}_{t-1})$.

Note that, for any A_u not contained in $S_{1t}(\bar{A}_{t-1})$, the expectation in (1) depends on distribution of A_u. This is a departure from the causal inference literature, where estimands do not depend on the treatment distribution in the data at hand. Nonetheless, for all choices of variables in $S_{1t}(\bar{A}_{t-1})$, the proximal treatment effect is causal, since (1) is the conditional mean of the contrast between the potential proximal response had an individual received ($a_t = 1$) versus not received ($a_t = 0$) treatment at occasion t. Considering the dependence of the proximal effect on the distribution of the treatments, it is best to always present this distribution along with the estimated treatment effect. For further discussion concerning including the treatment distribution as part of the estimand, see Section 8.

Many treatments may have delayed effects. For example, mindfulness messages have a delayed effect when individuals recall and employ mindfulness exercises provided prior to the most recent treatment occasion. In BASICS-Mobile, treatments suggesting alternative activities to smoking and drinking may achieve little to no immediate impact in the afternoon,

but the individual might follow these suggestions later on in the evening. So in general both proximal and other lagged effects of treatments on the response variable may be of interest. To define these lagged effects, we denote $A_{t+1}(\bar{A}_{t-1}, a)$ by $A_{t+1}^{a_t=a}$, $A_{t+2}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a})$ by $A_{t+2}^{a_t=a}$ and so on, with $A_{t+k-1}(\bar{A}_{t-1}, a, A_{t+1}^{a_t}, \cdots, A_{t+k-2}^{a_t=a})$ by $A_{t+k-1}^{a_t=a}$. We define the lag k effect of treatment on the response k treatment occasions into the future Y_{t+k} by

$$\mathrm{E}\Big[Y_{t+k}(\bar{A}_{t-1}, 1, A_{t+1}^{a_t=1}, \ldots, A_{t+k-1}^{a_t=1}) - Y_{t+k}(\bar{A}_{t-1}, 0, A_{t+1}^{a_t=0}, \ldots, A_{t+k-1}^{a_t=0}) \,\Big|\, S_{kt}(\bar{A}_{t-1})\Big], \qquad (2)$$

where k ranges from 1 up to the number of lags of scientific interest. So the proximal effect (1) corresponds to the lag $k = 1$ treatment effect. Note the dependence of both future actions as well as Y_{t+1} on the occasion t action as emphasized by the superscripts, $a_t = 1$ and $a_t = 0$. As with (1), $S_{kt}(\bar{A}_{t-1})$ is a vector of variables from the history $H_t(\bar{A}_{t-1})$. The lagged effect is also similarly averaged over the conditional distribution of variables in the history $H_t(\bar{A}_{t-1})$ not represented in $S_{kt}(\bar{A}_{t-1})$, which might include past treatment or underlying moderators. In addition, (2) is averaged over the distribution of treatments after occasion t but before response Y_{t+k}—namely $A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}$ for either $a = 1$ or $a = 0$.

The causal effect in (2) is a generalization of the treatment "blip" in the SNMM. In SNMMs, the tth treatment blip or intermediate effect on Y_{t+k} is usually defined with $S_{kt}(\bar{A}_{t-1}) = H_{kt}(\bar{A}_{t-1})$ and with respect to a prespecified future (after time t) "reference" treatment regime that defines the distribution for $A_{t+1}, \ldots, A_{t+k-1}$. For example, if we were studying treatment discontinuation, we might have chosen the reference regime $A_u = 0$ for $u > t$, with probability one (cf. Robins 1994, Section 3a). In this case the lag k treatment effect (2) represents the impact of one last additional treatment on the proximal response k time units later. The reference treatment regime reflected in (2), however, assigns treatment with probabilities between zero and one and corresponds to the distribution of treatments in the data we have at hand. For further discussion of the connection between the causal effects defined here and the SNMM, see Supplement A.1.

We now express the proximal and other lagged effects in terms of the observed data. For this we herein assume positivity, consistency and sequential ignorability (Robins 1994, 1997):

- Consistency: The observed data $(Y_2, X_2, A_2, \ldots, Y_T, X_T, A_T, Y_{T+1})$ are equal to the potential outcomes as follows: $Y_2 = Y_2(A_1), X_2 = X_2(A_1), A_2 = A_2(A_1)$ and for each subsequent $t \leq T$, $Y_t = Y_t(\bar{A}_{t-1}), X_t = X_t(\bar{A}_{t-1}), A_t = A_t(\bar{A}_{t-1})$ and lastly $Y_{T+1} = Y_{T+1}(\bar{A}_T)$.

- Positivity: If the joint density of $\{H_t = h_t, A_t = a_t\}$ is greater than zero, then $\Pr(A_t = a_t \mid H_t = h_t) > 0$, almost everywhere.

- Sequential ignorability: For each $t \leq T$, the potential outcomes $\{Y_{t+1}(\bar{a}_t), X_{t+1}(\bar{a}_t), A_{t+1}(\bar{a}_t), \ldots, Y_{T+1}(\bar{a}_T)\}$ are independent of A_t conditional on H_t.

The consistency assumption connects the potential outcomes with the data. When the treatment allocated to one individual may influence the response of others, the observed response Y_{t+1} is generally consistent not with the potential response $Y_{t+1}(\bar{A}_t)$ as above, but possibly with some other group-based conceptualization (e.g. Hong and Raudenbush 2006; Vanderweele et al. 2013). In particular, for a mobile intervention with a social media component, it may be necessary to define the potential outcomes for a given individual as a function of the treatments that are provided to individuals in their social network.

In an MRT, treatment is sequentially randomized according to known treatment probabilities, say $\Pr(A_t = 1 \mid H_t) = \rho_t(1 \mid H_t)$, $t = 1, \ldots, T$, and thus sequential ignorability is ensured by design. In an observational study, where treatment status is observed rather than randomized, sequential ignorability is often assumed. Here the underlying treatment probabilities $\rho_t(1 \mid H_t)$, $t = 1, \ldots, T$, are unknown.

In Supplement A.2 we show that, under these assumptions, the lag k treatment effect can be expressed in terms of the observed data as

$$\begin{aligned}
&\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, 1, A_{t+1}^{a_t=1}, \ldots, A_{t+k-1}^{a_t=1}) - Y_{t+k}(\bar{A}_{t-1}, 0, A_{t+1}^{a_t=0}, \ldots, A_{t+k-1}^{a_t=0}) \,\big|\, S_{kt}(\bar{A}_{t-1})\big] \\
&\qquad = \mathrm{E}[\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t] \mid S_{kt}] \\
&\qquad = \mathrm{E}\left[\frac{1(A_t = 1)Y_{t+k}}{\rho_t(1 \mid H_t)} - \frac{1(A_t = 0)Y_{t+k}}{1 - \rho_t(1 \mid H_t)} \,\middle|\, S_{kt}\right], \quad (3)
\end{aligned}$$

for $t = 1, \ldots, T - k + 1$, respectively. Note that if $S_{kt} = H_t$, then the lag k effect simplifies to

$$\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]. \quad (4)$$

3 Estimation

In the following we assume a linear model for the treatment effects. Fortunately, models for the proximal and other lagged treatment effects can in fact be specified separately, since (2) for differing lags k do not constrain one another (Robins 1994, 1997; see Supplement B). Suppose that the following holds.

A1 Each lag k treatment effect of interest takes the form

$$\mathrm{E}[\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t] \mid S_{kt}] = S_{kt}^{\intercal}\beta_k, \tag{5}$$

for some finite-valued vector β_k.

The rest of this paper is devoted to inference on β_k. Throughout we denote the true value of β_k by β_k^*, n represents the number of individuals in the data and $\mathbb{P}_n f(Z) = \sum_{i=1}^n f(Z_i)/n$ for some function f of the random vector Z. For simplicity we initially limit attention to the case where treatment is sequentially randomized as with the MRT; in this case sequential ignorability is satisfied. In particular we assume:

A2 Treatment is sequentially randomized with randomization probability $\Pr(A_t = 1 \mid H_t) = \rho_t(1 \mid H_t)$, for each $t = 1, \ldots, T$.

Inference concerning β_k using data from observational studies in which the treatment is not sequentially randomized can be handled–if the assumption of sequential ignorability holds–by estimating the treatment probability; see Supplement C.

Three different estimation methods are considered: "routine" regression, centering treatment status, and weighting by the inverse probability of treatment. In each case we aim to provide a method for estimating the lag k effect that is robust to misspecification of a nuisance function, $\mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]$ (or in the case of centering, $\mathrm{E}[Y_{t+k} \mid H_t]$). This robustness property is desirable for two reasons. First, the history H_t may be high dimensional, making it very difficult to model these nuisance functions correctly. Second, even when H_t is not very large, it can be difficult or impossible to specify models that can be correct for both the nuisance function as well as for the delayed treatment effects at lags $j > k$ (see Supplement B).

In a routine regression, one might think that the effect of the tth treatment is provided by the regression coefficients for a main treatment effect term A_t and interaction terms with A_t. Here we show that this intuition holds under special conditions. Let $\tilde{S}_{kt}$ be a vector constructed from H_t that contains S_{kt} as a subentry. Consider a routine regression analysis with the least squares estimating function

$$U_{\mathrm{R}}(\alpha_k, \beta_k) = \sum_{t=1}^{T-k+1} \left(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - A_t S_{kt}^{\intercal}\beta_k\right) \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}. \tag{6}$$

Let $\dot{U}_{\mathrm{R}}$ be the derivative of U_{R} with respect to the row vector $(\alpha_k^\intercal, \beta_k^\intercal)$. A more general version of following result is provided in Supplement C.1.

Proposition 3.1. *Assume* A1, A2 *and* at least one *of the following:*

R1 $\Pr(A_t = 1 \mid H_t) = \rho \in (0,1)$, *for each* $t = 1, \ldots, T$.

R2 $\mathrm{E}[Y_{t+k} \mid A_t, H_t] = \tilde{S}_{kt}^\intercal \alpha_k + A_t S_{kt}^\intercal \beta_k$, *for* S_{kt} *in* (5), *for each* $t = 1, \ldots, T-k+1$ *and for some finite-valued vectors* α_k *and* β_k.

Then, under invertibility and moment conditions, the solution to the estimating equation $\mathbb{P}_n U_{\mathrm{R}}(\alpha_k, \beta_k) = 0$ *yields an estimator* $(\hat{\alpha}_k, \hat{\beta}_k)$ *for which* $\sqrt{n}(\hat{\beta}_k - \beta_k^*)$ *is asymptotically normal with mean zero and variance-covariance matrix consistently estimated by the lower block entry of* $(\mathbb{P}_n \dot{U}_{\mathrm{R}}(\hat{\alpha}_k, \hat{\beta}_k))^{-1} \mathbb{P}_n U_{\mathrm{R}}(\hat{\alpha}_k, \hat{\beta}_k)^{\otimes 2} (\mathbb{P}_n \dot{U}_{\mathrm{R}}(\hat{\alpha}_k, \hat{\beta}_k))^{-1^\intercal}$.

If the trial design ensures that R1 holds, then $\tilde{S}_{kt}^\intercal \alpha_k$ can be viewed as a working model for $\mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]$; under R1, this working model need not be correctly specified to achieve consistent estimation of the treatment effects and conduct inference about these effects. Assumption R2 implies that the working model must be correct. If the trial design does not ensure R1, assumption R2 can be used to conduct inference about the treatment effects. However R2 will likely hold in only very special circumstances. This is because first, $\mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]$ is difficult to correctly specify when H_t is large and second the specified model constrains the form of treatment effects at other lags, $\mathrm{E}[\mathrm{E}[Y_{t+j} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+j} \mid A_t = 0, H_t] \mid S_{jt}]$, $j > k$ (Supplement B). In addition, R2 implies that $\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t] = S_{kt}^\intercal \beta_k^*$; that is, S_{kt} contains all moderators in H_t.

The centering approach might be considered if the randomization probabilities are stratified or time-varying or condition R2 is deemed too restrictive. The centering approach is a straightforward extension of Liao et al.'s (2015) least squares criterion, which gives the estimating function

$$U_{\mathrm{C}}(\alpha_k, \beta_k) = \sum_{t=1}^{T-k+1} \left(Y_{t+k} - \tilde{S}_{kt}^\intercal \alpha_k - (A_t - \rho_t(1 \mid H_t)) S_{kt}^\intercal \beta_k \right) \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t)) S_{kt} \end{pmatrix}, \quad (7)$$

where $\tilde{S}_{kt}$ is a vector based on the history H_t. Let $\dot{U}_{\mathrm{C}}$ be the derivative of U_{C} with respect to the row vector $(\alpha_k^\intercal, \beta_k^\intercal)$. A proof of a more general version of the following result can be found in Supplement C.2.

Proposition 3.2. *Assume* A1, A2 *and* at least one *of* C1 *and* C2*:*

C1 $\Pr(A_t = 1 \mid H_t) = \rho_t(1 \mid S_{kt})$, *with* S_{kt} *as in* (5) *and for each* $t = 1, \ldots, T$.

C2 S_{kt} *in* (5) *contains all underlying lag* k *treatment effect moderators in* H_t*; that is,*

$$\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t] = S_{kt}^\intercal \beta_k,$$

for each $t = 1, \ldots, T - k + 1$.

Then, under invertibility and moment conditions, the solution to the estimating equation $\mathbb{P}_n U_{\mathrm{C}}(\alpha_k, \beta_k) = 0$ *yields an estimator* $\hat{\beta}_k$ *for which* $\sqrt{n}(\hat{\beta}_k - \beta_k^*)$ *is asymptotically normal with mean zero and variance-covariance consistently estimated by the lower block entry of* $(\mathbb{P}_n \dot{U}_{\mathrm{C}}(\hat{\alpha}_k, \hat{\beta}_k)^{-1} \mathbb{P}_n U_{\mathrm{C}}(\hat{\alpha}_k, \hat{\beta}_k^*)^{\otimes 2} \mathbb{P}_n \dot{U}_{\mathrm{C}}(\hat{\alpha}_k, \hat{\beta}_k)^{-1^\intercal}$.

Under either C1 or C2, $\tilde{S}_{kt}^\intercal \alpha_k$ is a working model for $\mathrm{E}[Y_{t+k} \mid H_t]$; it need not be correct in order for $\hat{\beta}_k$ to be consistent for β_k^* (Supplement C.2). In the special case where the treatment probabilities are fixed, $\rho_t(H_t) = \rho$, the centering method is equivalent to a routine regression analysis, since the term $\rho S_{kt}^\intercal \beta_k$ in (7) is absorbed into the working model. Centering has been previously employed by Brumback et al. (2003) and Goetgeluk and Vansteelandt (2008) for causal inference. For example Goetgeluk and Vansteelandt (2008) center exposure variables by their overall mean to protect against unmeasured baseline confounders. Brumback et al. (2003) center time-varying exposures by their conditional mean given the history, as we do; they consider treatment effects under a treatment discontinuation reference regime and limit attention to overall effects without interaction terms. In contrast to these papers, our use of centering is solely to provide robustness to the specification of a model for $\mathrm{E}[Y_{t+k} \mid H_t]$; centering is not used to adjust for confounding.

Lastly the weighting approach might be considered if the randomization probabilities are stratified and we wish to consider a set of candidate moderators S_{kt} that do not include all of the variables used to randomize treatment. Such a scenario arises when interest lies in the marginal treatment effect of A_t: $S_{kt} = 1$.

Let $\tilde{S}_{kt}$ be a vector containing summary variables constructed from the history H_t and that contains S_{kt} as a sub-vector. Consider the weighted least squares estimating function

$$U_{\mathrm{W}}(\alpha_k, \beta_k) = \sum_{t=1}^{T-k+1} \left(Y_{t+k} - \tilde{S}_{kt}^\intercal \alpha_k - A_t S_{kt}^\intercal \beta_k\right) w_t(A_t, H_t) \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}, \tag{8}$$

where $w_t(A_t, H_t) = \frac{\rho^{A_t}(1-\rho)^{1-A_t}}{\rho_t(A_t|H_t)}$ for some fixed $\rho \in (0,1)$. Let $\dot{U}_{\mathrm{W}}$ be the derivative of U_{W} with respect to the row vector $(\alpha_k^\intercal, \beta_k^\intercal)$. In Supplement C.3 we prove a more general version of the following result.

Proposition 3.3. *Assume A1 and A2, both defined above. Then, under invertibility and moment conditions, the solution to the estimating equation* $\mathbb{P}_n U_{\mathrm{W}}(\alpha_k, \beta_k) = 0$ *yields an estimator* $(\hat{\alpha}_k, \hat{\beta}_k)$ *for which* $\sqrt{n}(\hat{\beta}_k - \beta_k^*)$ *is asymptotically normal with mean zero and variance-covariance matrix consistently estimated by the lower block diagonal entry of the matrix* $(\mathbb{P}_n \dot{U}_{\mathrm{W}}(\hat{\alpha}_k, \hat{\beta}_k))^{-1} \mathbb{P}_n U_{\mathrm{W}}(\hat{\alpha}_k, \hat{\beta}_k)^{\otimes 2} (\mathbb{P}_n \dot{U}_{\mathrm{W}}(\hat{\alpha}_k, \hat{\beta}_k))^{-1^\intercal}$.

In this setting $\tilde{S}_{kt}^\intercal \alpha_k$ is a working model for $\mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]$; it need not be correct in order for $\hat{\beta}_k$ to be consistent for β_k^* (Supplement C.3). In the special case where the randomization probabilities are fixed, we choose $\rho = \rho_t(H_t)$ so that $w_t(A_t, H_t) = 1$ and weighted least squares is equivalent to routine regression. Unlike centering, weighting imposes no restrictions on how closely S_{kt} represents the history used to randomize treatment (as in C1) or the underlying moderators (as in C2); the estimated treatment effect is simply averaged over any underlying moderators omitted from S_{kt}, regardless of their relationship with variables used in the randomization.

The weight $w_t(A_t, H_t)$ is reminiscent of inverse probability of treatment weighting (IPTW) in causal inference (Robins 1998). However here weighting is simply used to make the weighted least squares estimator $\hat{\beta}_k$ robust against the (usual) case in which the working model $\tilde{S}_{kt}^\intercal \alpha_k$ misspecifies $\mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]$. Following Robins et al. (2000), we might attempt to increase precision in $\hat{\beta}_k$ through weight stabilization, wherein the constant ρ in the numerator of w_t is replaced by some function of the data. Unfortunately if this numerator depends on time or covariates that are not independent of all S_{kt}, then the resulting $\hat{\beta}_k$ is in general no longer consistent for β_k^*. See Section 6 for an example.

4 Availability

Up to this point we have implicitly presumed that at every possible occasion t, the participant is available to engage with the mobile intervention. Consideration of availability is critical in not only the analysis of MRT data (Liao et al. 2015), but also in the design of the intervention, since it might be unreasonable, counter-productive or even unethical to always presume availability. For example in HeartSteps (Klasnja et al. 2015), smartphone notifications are used to deliver suggestions to disrupt sedentary behavior. Here the participant is considered

unavailable when driving a vehicle (because the notification may be distracting) or walking (as treatment at this time is scientifically inappropriate). Detection of availability can be carried out through sensors (as in the case of HeartSteps) or recent interaction with the mobile device. BASICS-Mobile took the latter approach by presuming that participants were available to receive a treatment only after they fully completed a self-report.

Assume that the measurements X_t just prior to the tth treatment occasion contain the user's availability status, denoted by I_t, where $I_t = 1$ if the participant is available to engage with the treatment at occasion t and $I_t = 0$ otherwise. If the participant is unavailable, the treatment A_t is not delivered. To define the treatment effects under limited availability, we use potential outcome notation. As compared to Section 2.3, here the potential outcome notation is slightly more complicated because treatment can only be provided when an individual is available. The potential outcomes are indexed by decision rules that incorporate availability. In particular define $d(a, i)$ for $a \in \{0, 1\}$, $i \in \{0, 1\}$ by $d(a, 0) = 0$ and $d(a, 1) = a$. Then for each $a_1 \in \{0, 1\}$, define $D_1(a_1) = d(a_1, I_1)$. Then we denote, for example, the potential availability indicator at $t = 2$ by $\{I_2(D_1(1)), I_2(D_1(0))\}$. The potential outcomes for availability emphasizes the fact that previous exposure to treatment can influence subsequent availability. In BASICS-Mobile, for example, repeated provision of treatment might lead to lower engagement with the intervention, and therefore lower availability for further delivery of the treatment.

For each $\bar{a}_2 = (a_1, a_2)$ with $a_1, a_2 \in \{0, 1\}$, define $D_2(\bar{a}_2) = d(a_2, I_2(D_1(a_1)))$ and $\overline{D_2(\bar{a}_2)} = (D_1(a_1), D_2(\bar{a}_2))$. A potential proximal response following occasion $t = 2$ and corresponding to $\bar{a}_2$ is $Y_3(\overline{D_2(\bar{a}_2)})$ and a potential availability indicator at occasion $t = 3$ is $I_3(\overline{D_2(\bar{a}_2)})$. Similarly, for each $\bar{a}_t = (a_1, \ldots, a_t) \in \{0, 1\}^t$, define $D_t(\bar{a}_t) = d(a_t, I_t(\overline{D_{t-1}(\bar{a}_{t-1})}))$ and $\overline{D_t(\bar{a}_t)} = (D_1(a_1), \ldots, D_t(\bar{a}_t))$. For each $\bar{a}_t = (a_1, \ldots, a_t) \in \{0, 1\}^t$, the potential proximal response is $Y_t(\overline{D_{t-1}(\bar{a}_{t-1})})$ (following occasion $t - 1$) and potential availability indicator is $I_t(\overline{D_{t-1}(\bar{a}_{t-1})})$ at occasion t.

We now incorporate availability into the definition of the proximal treatment effect; first recall the notation from Section 2.3: denote $A_2(A_1)$ by A_2 and so on with $A_t(\bar{A}_{t-1})$ denoted by A_t. The proximal treatment effect is

$$\mathrm{E}\Big[Y_{t+1}\Big(\overline{D_t(\bar{A}_{t-1}, 1)}\Big) - Y_{t+1}\Big(\overline{D_t(\bar{A}_{t-1}, 0)}\Big) \mid I_t\Big(\overline{D_{t-1}(\bar{A}_{t-1})}\Big) = 1, S_{1t}\Big(\overline{D_{t-1}(\bar{A}_{t-1})}\Big)\Big].$$

Unlike (1), this effect is defined for only individuals available for treatment at time t, that is, $I_t\Big(\overline{D_{t-1}(\bar{A}_{t-1})}\Big) = 1$. This subpopulation is not static; at a given treatment occasion t only certain types of individuals might tend to be available and availability for any given

individual may change with t. Conditioning on availability is related to the concept of viable or feasible dynamic treatment regimes (Wang et al. 2012; Robins 2004), in which one assesses only the causal effect of treatments that can actually be provided.

To incorporate availability into the definition of the lagged effects, first recall the notation from Section 2.3: denote $A_{t+1}(\bar{A}_{t-1}, a)$ by $A_{t+1}^{a_t=a}$, $A_{t+2}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a})$ by $A_{t+2}^{a_t=a}$, and so on, with $A_{t+k-1}(\bar{A}_{t-1}, a, A_{t+1}^{a_t}, \cdots, A_{t+k-2}^{a_t=a})$ by $A_{t+k-1}^{a_t=a}$. The lag k effect of treatment on the response k treatment occasions into the future Y_{t+k} is defined by

$$\mathrm{E}\Big[Y_{t+k}\Big(\overline{D_t(\bar{A}_{t-1}, 1, A_{t+1}^{a_t=1}, \ldots, A_{t+k-1}^{a_t=1})}\Big) - Y_{t+k}\Big(\overline{D_t(\bar{A}_{t-1}, 0, A_{t+1}^{a_t=0}, \ldots, A_{t+k-1}^{a_t=0})}\Big) \mid S_{kt}\Big(\overline{D_{t-1}(\bar{A}_{t-1})}\Big)\Big].$$

Assuming consistency, positivity and sequential ignorability, the lag k treatment effect under limited availability can be expressed in terms of the data as

$$\mathrm{E}[\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] \mid I_t = 1, S_{kt}] = \mathrm{E}\left[\frac{1(A_t = 1)Y_{t+1}}{\rho_t(1 \mid H_t)} - \frac{1(A_t = 0)Y_{t+1}}{1 - \rho_t(1 \mid H_t)} \,\middle|\, I_t = 1, S_{kt}\right],$$

where $\rho_t(1 \mid H_t)$ is now $\Pr(A_t = 1 \mid I_t = 1, H_t)$. Modeling and estimation proceeds following the same approach as with the always-available setting. In particular for the lag k treatment effect, we consider the linear model

$$\mathrm{E}[\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] \mid I_t = 1, S_{kt}] = S_{kt}^{\intercal}\beta_k^*.$$

In applying the routine, centering or weighting estimation methods, we simply factor the tth contribution to the corresponding estimating function by I_t. The resulting estimating equations are provided in displays (13), (17) and (18) of the Supplement. For each, the relevant working and treatment probability models are now conditional on $I_t = 1$. Proofs can be found in Supplements C.1, C.2 and C.3.

5 Implementation

All of the proposed estimators can be implemented using standard software for generalized estimating equations (GEE, Liang and Zeger 1986), provided that we: (i) incorporate availability I_t or (in the case of weighted least squares) $I_t w_t$ as "prior weights" (McCullagh and

Nelder 1989, Section 2.2) and (ii) employ the independence working correlation structure.

Non-independence working correlation structures such as exchangeable or AR(1) are often adopted in the analysis of longitudinal data to improve precision (e.g. Schafer 2006). One might wish to use these structures in our setting for the same reason, but this strategy will generally introduce bias. Such a result is unsurprising given the bias that arises under non-independence structures in IPTW (Vansteelandt 2007; Tchetgen Tchetgen et al. 2012) or in GEEs where a time-varying response is modelled by time-varying covariates (Pepe and Anderson 1994; Schildcrout and Heagerty 2005). In Section 6 we provide an example of how bias in the proximal treatment effect can arise through use of AR(1) structure. A working independence structure with variance models that depend on at most the variables used in the treatment effect, however, does not result in large sample bias (see Supplement C). For simplicity, in Sections 6 and 7 below, we do not employ a working variance model.

The standard errors expressed in Propositions 3.1, 3.2 and 3.3 directly correspond to the sandwich variance-covariance estimator provided by GEE software. From existing work on GEEs, it is well understood that the sandwich estimator is non-conservative in small samples. To address this, whenever $n \leq 50$, we apply Mancl and DeRouen's (2001) small sample correction by inverse-scaling each residual in U_{R}, U_{C}, or U_{W} by one minus its leverage. When a working variance model is employed or at least one of ρ and $\rho_t(1 \mid H_t)$ is estimated, the sandwich variance-covariance estimator must be adjusted to account for this additional sampling error (see Supplement C). See Supplement E to obtain code that calculates standard errors using the R package geepack (Højsgaard et al. 2006; R Core Team 2015).

6 Simulation Study

For each estimation method, we have discussed various properties—conditions used to achieve consistency or implementations that can lead to bias. Here we illustrate some of these results with simulated data. For simplicity, we consider data arising from the MRT, where $\rho_t(1|H_t)$ is known. The generative model for the response is $Y_{t+1} = \theta(S_t - \mathrm{E}[S_t \mid A_{t-1}, H_{t-1}]) + (A_t - \rho_t(1 \mid H_t))(\beta^*_{10} + \beta^*_{11} S_t) + \epsilon_{t+1}$, where $\rho_t(1 \mid H_t) = \mathrm{expit}(\eta_1 A_{t-1} + \eta_2 S_t)$, $S_t \in \{-1, 1\}$ with $\Pr(S_t = 1 \mid A_{t-1}, H_{t-1}) = \mathrm{expit}(\xi A_{t-1})$, and $\epsilon_t \sim \mathrm{N}(0, 1)$ with $\mathrm{Corr}(\epsilon_u, \epsilon_t) = 0.5^{|u-t|}$. The proximal effect conditional on S_t is given by $\mathrm{E}[\mathrm{E}[Y_{t+1} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+1} \mid A_t = 0, H_t] \mid S_t] = \beta^*_{10} + \beta^*_{11} S_t$.

In the simulation scenarios below, we fix $\beta^*_{10} = -0.8$ and vary $(\theta, \beta^*_{11}, \eta_1, \eta_2, \xi)$. Note that if $\beta^*_{11} = 0$ or $\xi = 0$, then the marginal proximal treatment effect, $\mathrm{E}[\mathrm{E}[Y_{t+1} \mid A_t = 1, H_t] - \mathrm{E}[Y_{t+1} \mid A_t = 0, H_t]] = \beta^*_{10}$ is equal to -0.8. Throughout each subject is available at

every treatment occasion: $I_t = 1$ $(t = 1, \ldots, T)$.

In the analysis of the simulated data, we fit an intercept-only proximal treatment effect model; this model does not include covariates in the proximal treatment effect. So the primary interest is in the marginal proximal treatment effect.

We report average point estimates and coverage probabilities over 1000 Monte Carlo replicates with $n = 50$ or 100 and $T = 100$. Confidence intervals are evaluated using standard errors corrected for any estimation of weights, treatment probabilities or (for $n = 50$) small samples (see Section 5). The tables below omit the average estimated standard errors; these are provided in Supplement D and closely correspond to the standard deviations of the point estimates. Alternative values for n and T were examined, but results were similar.

This first simulation scenario concerns the omission of an important moderator and illustrates that, when primary interest is in the marginal proximal treatment effect, weighting is preferable over both routine regression and centering. In the data generative model, we set $\theta = 0.8$, $\eta_1 = -0.8$, $\eta_2 = 0.8$ and $\xi = 0$ (recall $\xi = 0$ implies that the true marginal proximal treatment effect is 0.8). Different scenarios were devised by setting β_{11}^* to one of 0.2, 0.5, 0.8, giving respectively a small, medium, or large degree of moderation by S_t. Since η_1 and η_2 are nonzero, the treatment A_t is assigned with a probability depending on both S_t and past treatment A_{t-1}, for each t.

For routine regression, the working data analysis model for $\mathrm{E}[Y_{t+1} \mid H_t, A_t = 0]$ is $\alpha_0 + \alpha_1 S_t$, to give a conditional mean data analysis model $\mathrm{E}[Y_{t+1} \mid A_t, H_t] = \alpha_0 + \alpha_1 S_t + \beta_{10} A_t$. For weighted least squares, the same working model is used; the weights are set to $w_t(A_t, H_t) = \hat{\rho}^{A_t}(1 - \hat{\rho})^{1-A_t}/\rho_t(A_t \mid H_t)$ with $\hat{\rho} = \mathbb{P}_n \sum_{t=1}^{T} A_t/T$. A similar conditional mean model is adopted for centered least squares, but we replace the term $\beta_{10} A_t$ with $\beta_{10}(A_t - \rho_t(1 \mid H_t))$. In each of the three methods $\hat{\beta}_{10}$ is the estimator of marginal treatment effect.

With $\eta_2 = 0.8$, the randomization probability $\rho_t(1 \mid H_t)$ depends on the underlying moderator S_t, which is omitted from the treatment effect model under all three methods. So neither R1 nor R2 hold for the data analysis, and we therefore anticipate the $\hat{\beta}_{10}$ from routine regression to be a biased estimator of the marginal treatment effect of -0.8. Moreover both of C1 and C2 do not hold, so we expect centering to also result in bias. All of the requirements needed to achieve consistency in weighted least squares are satisfied. Hence the $\hat{\beta}_{10}$ from the weighting method should be unbiased, regardless of the value for β_{11}^*. These conjectures are supported by Table 1, where only in the weighting method is the nominal 95% coverage achieved for each value of β_{11}^*. Bias in routine regression and centering $\hat{\beta}_{10}$'s increases with β_{11}^*, as the degree of omitted moderation becomes larger.

Table 1: Estimator, $\hat{\beta}_{01}$ of the marginal proximal effect; omitting an important moderator.

	$\beta^*_{11} = 0.2$		$\beta^*_{11} = 0.5$		$\beta^*_{11} = 0.8$	
Method	Mean	CP	Mean	CP	Mean	CP
Weighted	–0.80	0.95	–0.80	0.95	–0.80	0.95
Centered	–0.78	**0.82**	–0.74	**0.32**	–0.71	**0.04**
Routine	–0.75	**0.34**	–0.72	**0.04**	–0.68	**0.00**

Mean, average point estimate; CP, proportion of 95% confidence intervals that contained the truth, with boldface indicating a significant difference from 0.95 at the 5% level. The true marginal proximal effect is -0.8 (averaged over an underlying moderator with coefficient β^*_{11}). Results are based on 1000 replicates with $n = T = 100$.

The second simulation scenario illustrates that our ability to stabilize the weights is limited, since weighted least squares is prone to bias if, instead of the constant ρ in $w_t(A_t, H_t)$, the weight numerator varies with time. In the data generative model, we set $\theta = 0.8$, $\beta^*_{11} = 0$, $\eta_1 = -0.8$, $\eta_2 = 0.8$ and $\xi = 0$. Thus as above, the randomization probability for A_t depends on both S_t and past treatment A_{t-1} ($t = 1, \ldots, T = 100$). However no variables moderate the proximal effect, since $\beta^*_{11} = 0$.

In the data analysis our working model for $\mathrm{E}[Y_{t+1} \mid A_t = 0, H_t]$ uses only an intercept: α_0. The denominator of the weight $w_t(A_t, H_t)$ is the known $\rho_t(A_t \mid H_t)$. We compare two different numerators: (i) $\hat{\rho} = \mathbb{P}_n \sum_{t=1}^{T} A_t/T$ and (ii) $\hat{\rho}(1|S_t) = \mathrm{expit}(\hat{\phi}_0 + \hat{\phi}_1 S_t)$, where $(\hat{\phi}_0, \hat{\phi}_1)$ is the solution to $\mathbb{P}_n \sum_t \exp(\phi_0 + \phi_1 S_t)\{\mathrm{expit}(\phi_0 + \phi_1 S_t)(1 - \mathrm{expit}(\phi_0 + \phi_1 S_t))\}^{-1}(A_t - \mathrm{expit}(\phi_0 + \phi_1 S_t))(1, S_t)^\intercal = 0$. In (i) the numerator is identical for all w_t ($t = 1, \ldots, T = 100$). In (ii) the numerator depends on S_t. Hence, we anticipate bias in estimates under (ii), but not (i). This is indeed reflected in Table 2, where the time-varying weight stabilization of (ii) induces a large degree of bias, with the corresponding confidence interval rarely capturing the true value of -0.8.

Table 2: Weighted least squares estimation of the proximal effect; stabilized weights

Method	Mean	CP
Weighted (i)	–0.80	0.95
Weighted time–varying stabilizer (ii)	–0.11	**0.00**

Mean, average point estimate; CP, proportion of 95% confidence intervals that contained the truth, with boldface indicating a significant difference from 0.95 at the 5% level. The true proximal effect is -0.8. Results are based on 1000 replicates with $n = T = 100$.

The last simulation scenario illustrates that employing a non-independence working correlation structure in the estimation can result in bias. In the data generative model, we set $\theta = 0.8$, $\beta_{11}^* = 0$, $\eta_1 = \eta_2 = 0$ and $\xi = 0.8$. There is no moderation of the proximal effect, since $\beta_{11}^* = 0$. Unlike the above scenarios, here the predictor S_t is influenced by A_{t-1}, since $\xi = 0.8$. Treatment is randomized with fixed probability $\rho_t(1 \mid H_t) = 1/2$ for each $t = 1, \ldots, T = 100$.

In data analysis, suppose that S_t is unobserved and the working model for $\mathrm{E}[Y_{t+1} \mid A_t = 0, H_t]$ is set to $\alpha_0 + \alpha_1 Y_t$ (the working model is incorrect). For simplicity we carry out data analysis with only weighted least squares; we use the weight numerator $\rho = 1/2$ and known denominator $\rho_t(A_t \mid H_t) = 1/2$. Two variants for the working correlation structure are considered: (i) estimation employing an independence working correlation structure (adhering to condition A3 in Supplement C) and (ii) estimation adopting a working AR(1) structure assuming a residual correlation of $0.5^{|u-t|}$ between times u and t. While AR(1) might better represent the true correlation matrix than an independence structure, we expect this choice to induce bias because the working model is not correctly specified. Table 3 demonstrates this result, with the non-diagonal structure achieving a coverage probability less than 60%.

Table 3: Weighted least squares estimation of the proximal effect; correlation structures.

Method	Mean	CP
Weighted (i)	–0.80	0.96
Weighted fixed AR(1) (ii)	–0.74	**0.59**

Mean, average point estimate; CP, proportion of 95% confidence intervals that contained the truth, with boldface indicating a significant difference from 0.95 at the 5% level. The true proximal effect is -0.8. Results are based on 1000 replicates with $n = T = 100$.

7 Application

BASICS-Mobile is a pilot study, with $n = 28$, $T = 28$. The response Y_{t+1} is the smoking rate from the tth occasion to the next self-report, and participants are presumed available only if they completed the preceding self-report. So the availability I_t is the self-report completion status just prior to t and the treatment decision D_t is 1 only if a mindfulness message is provided at t. Otherwise, $D_t = 0$.

BASICS-Mobile was neither a sequentially randomized trial nor an observational study.

Treatment delivery at occasion t was based on a complex decision rule involving primarily a self-reported measure that the user had an urge or inclination to smoke at the preceding self-report ($urge_t$), an indicator for the first three treatment occasions ($1(t < 4)$), and a combination of other variables. For illustrative purposes we provide an analysis acting as though the study was observational and assuming sequential ignorability; we estimate the treatment probabilities $\rho_t(H_t)$ based on $(Y_t, urge_t, 1(t < 4))$ using

$$\rho_t(H_t; \hat{\eta}) = \text{expit}(0.69 + 0.02Y_t + 0.17urge_t - 0.28\,1(t < 4) + 0.70urge_t\,1(t < 4)).$$

We examine one candidate moderator for the proximal effect: $S_{1t} = (1, \text{incr}_t)^\intercal$, where incr_t indicates whether or not the user reported an increase in need to self-regulate thoughts over the two self-reports preceding t. For delayed effects, we consider only the marginal lag-2 effect: $S_{2t} = 1$. In the working model a variety of predictors are incorporated in $\tilde{S}_{kt}$ ($k = 1, 2$), including incr_t, current urge to smoke, Y_{t+1-k}, time of day, baseline smoking severity, baseline drinking level, age and gender.

Here we apply weighted least squares because the proximal effect model contains only one moderator and the lag-2 model contains no moderators. Using $\rho_t(H_t; \hat{\eta})$ as the denominator of the weights and $\hat{\rho} = \mathbb{P}_n \sum_t A_t/T$ in the numerator, the data analysis leads to several conclusions. First the mindfulness message achieved a reduction in the average next-reported smoking rate, but only when the user was experiencing either a stable or decreased need to self-regulate (95% CI -5 to -0.2 cigarettes per day; see Table 4). Otherwise no proximal treatment effect is apparent. Second, evidence to support the presence of an overall lag-2 effect is relatively weak, with a 95% CI of -3 to 0.2 cigarettes per day for the average reduction achieved by mindfulness treatment at the second-to-last treatment occasion. Estimated standard errors (SEs) take into account sampling error in estimated treatment probabilities (see (19) for the formula), and are corrected for small n (see Section 5 for details on the correction).

Table 4: Proximal and lag-2 treatment effects estimated from BASICS-Mobile data.

Treatment effect	Estimate	SE	95% CI	p-value
Proximal, increase in self-regulation	-0.5	0.81	$(-2.0, \quad 1.1)$	0.565
Proximal, no increase self-regulation	-2.5	1.20	$(-4.9, -0.2)$	0.045
Delayed	-1.5	0.84	$(-3.1, \quad 0.2)$	0.079

8 Discussion

In this paper we define treatment effects suited for mobile interventions that enable frequent measurements and frequent delivery of treatments. As we discussed, the effect definition as provided in (1) and (2) is atypical in the field of causal inference in that the underlying mechanism for the assigned treatment is part of the definition of the causal effect. This definition of the causal effects is consistent with the effects defined via most models for intensively collected longitudinal data (see Schafer 2006, Schwartz and Stone 2007 and, more recently, Bolger and Laurenceau 2013). Commonly the model for the conditional mean of a time-varying response given time-varying covariates is a linear model (possibly with the use of covariates defined by flexible basis functions). If indicators of treatment as well as interactions between the treatment indicators and time varying covariates are included in the linear model then the coefficients of these covariates coincide with the moderated proximal effect defined here. However estimation of these casual effects using most common approaches (Schafer 2006; Schwartz and Stone 2007; Bolger and Laurenceau 2013), that is, either GEE approaches or approaches that employ random effects, can cause bias. Indeed the large sample and simulation results provided here show that GEEs based on a non-independence working covariance structure is not guaranteed to consistently estimate β_k^*.

Since the estimating functions under a response model with random intercepts or random coefficients (e.g. Goldstein 2011) have a similar form as those in GEEs, the covariance structure induced by models with random effects may yield a biased estimator for β_k^*. This connection is important given the fact that, in the analysis of intensive longitudinal data, there is a preference for including random effects and, when GEE models are used, to use a non-independence working correlation structure (such as exchangeable, $\text{Corr}(Y_u, Y_t) = r$ $(u \neq t)$, or AR(1), $\text{Corr}(Y_u, Y_t) = r^{|u-t|}$) to improve precision (Schafer 2006, p. 58). Future work is needed on whether or how to incorporate random effects in the estimation of proximal and lagged treatment effects.

Although we consider three possible methods to estimating proximal and other lagged effects, the choice between them is typically clear. In particular, if the treatment probabilities are fixed, routine analysis with moderation quantified by coefficients of treatment interaction terms in a regression model is simple to use and does not result in bias. Otherwise, if treatment is randomized with probabilities that vary by time and/or the moderators included in the analysis model, (5), then using the centering method is preferable to the routine regression in order to prevent bias in estimation. In all other settings, the weighting method is generally preferable. The weighted method facilitates inference concerning a small number

of potential treatment effect moderators with the fewest assumptions.

Methods in this manuscript were developed primarily for use with data arising from sequentially randomized trials in mobile health, such as with MRTs. We are currently involved in two MRTs, the first of which is currently in the field. In the first MRT participants are randomized with 5 times per day to receive or not receive a treatment with randomization probability equal to 0.6. Since $\rho_t(H_t) = 0.6$ for all t, data from this trial can be analyzed using routine or centering estimation methods (weighting is not necessary regardless of the estimand). In the second MRT participants will be randomized up to 600 times per day with randomization probabilities depending on the current stress classification and to ensure that, on average each day 2 treatments are delivered while an individual is classified as stressed and 2 treatments are delivered while an individual is classified as not stressed. Here, we plan to use the weighted least squares method to estimate the marginal proximal and lagged effects.

There are a number of other directions for future work. Throughout we limited attention to a continuous response and binary treatment decisions. Lagged effects ($k > 1$) were defined similar to proximal effects ($k = 1$), but in future work one might rather be interested in a lagged effect that quantifies the accumulation of past treatment. Finally, since small to moderate treatment effects may be difficult to detect, yet potential response predictors that can be used in the working models to reduce error variance are numerous, future work could consider penalized methods for the working model in order to accommodate and select from the large number of predictors.

References

Bolger, N. and J.-P. Laurenceau (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience*. New York, NY: The Guilford Press.

Bowen, S. and A. Marlatt (2009). Surfing the urge: Brief mindfulness-based intervention for college student smokers. *Psychology of Addictive Behaviors 23*(4), 666–671.

Brumback, B., S. Greenland, M. Redman, N. Kiviat, and P. Diehr (2003). The intensity-score approach to adjusting for confounding. *Biometrics 59*(2), 274–285.

Dempsey, W., P. Liao, P. Klasnja, I. Nahum-Shani, and S. A. Murphy (2015). Randomized trials for the Fitbit generation. *Significance 12*(6), 20–23.

Dimeff, L. A. (1999). *Brief Alcohol Screening and Intervention for College Students (BASICS): A Harm Reduction Approach*. Guilford Press.

Free, C., G. Phillips, L. Galli, L. Watson, L. Felix, P. Edwards, V. Patel, and A. Haines (2013). The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: A systematic review. *PLoS Medicine 10*(1), e1001362.

Goetgeluk, S. and S. Vansteelandt (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics 64*(3), 772–780.

Goldstein, H. (2011). *Multilevel Statistical Models*. John Wiley & Sons, Ltd.

Heron, K. E. and J. M. Smyth (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology 15*(1), 1–39.

Højsgaard, S., U. Halekoh, and J. Yan (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software 15*(2), 1–11.

Hong, G. and S. W. Raudenbush (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association 101*(475), 901–910.

Klasnja, P., E. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari, and S. Murphy (2015). Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology 34 Suppl*, 1220–1228.

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

Liao, P., P. Klasnja, A. Tewari, and S. A. Murphy (2015). Micro-randomized trials in mHealth. *Statistics in Medicine*. in press.

Mancl, L. A. and T. A. DeRouen (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics 57*(1), 126–134.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). CRC Press.

Mohr, D. C., S. M. Schueller, E. Montague, M. N. Burns, and P. Rashidi (2014). The behavioral intervention technology model: An integrated conceptual and technological framework for ehealth and mhealth interventions. *Journal of Medical Internet Research 16*(6), e146.

Muessig, K. E., E. C. Pike, S. Legrand, and L. B. Hightow-Weidman (2013). Mobile phone applications for the care and prevention of HIV and other sexually transmitted diseases: A review. *Journal of Medical Internet Research 15*(1), e1.

Neyman, J. (1990). On the application of probability theory to agricultural experiments: Essay on principles. *Statistical Science 5*(4), 465–472. Translated and edited by D. M. Dabrowska and T. P. Speed.

Pepe, M. S. and G. L. Anderson (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.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robins, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, A. Mulley, and National Center for Health Services Research and Health Care Technology Assessment (Eds.), *Health Services Research Methodology: A Focus on AIDS*, Rockville, MD, pp. 113–159. National Center for Health Services Research and Health Care Technology Assessment, Public Health Service, U.S. Department of Health and Human Services.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods 23*(8), 2379–2412.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality*, Volume 120 of *Lecture Notes in Statistics*, pp. 69–117. New York: Springer.

Robins, J. M. (1998). Marginal structural models. In *1997 Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA, pp. 1–10. American Statistical Association.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In D. Y. Lin and P. J. Heagerty (Eds.), *Proceedings of the Second Seattle Symposium in Biostatistics*, Number 179 in Lecture Notes in Statistics, pp. 189–326. New York: Springer.

Robins, J. M., M. A. Hernán, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 1–94. Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Schafer, J. L. (2006). Marginal modeling of intensive longitudinal data by generalized estimating equations. In T. A. Walls and J. L. Schafer (Eds.), *Models for Intensive Longitudinal Data*, pp. 38–62. New York, NY: Oxford University Press.

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

Schwartz, J. E. and A. A. Stone (2007). The analysis of real-time momentary data: A practical guide. In A. A. Stone, S. Shiffman, A. A. A., and L. Nebeling (Eds.), *The Science of Real-Time Data Capture*, pp. 76–113. New York, NY: Oxford University Press.

Spring, B., M. Gotsis, A. Paiva, and D. Spruijt-Metz (2013). Healthy apps: Mobile devices for continuous monitoring and intervention. *IEEE Pulse 4*(6), 34–40.

Tchetgen Tchetgen, E. J., M. M. Glymour, J. Weuve, and J. Robins (2012). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. *Epidemiology 23*(4), 644–646. Letter to the Editor.

Vanderweele, T. J., G. Hong, S. M. Jones, and J. L. Brown (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4rs educational intervention. *Journal of the American Statistical Association 108*(502), 469–482.

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.

Wang, L., A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association 107*(498), 493–508.

Witkiewitz, K., S. A. Desai, S. Bowen, B. C. Leigh, M. Kirouac, and M. E. Larimer (2014). Development and evaluation of a mobile intervention for heavy drinking and smoking among college students. *Psychology of Addictive Behaviors 28*(3), 639–650.

Acknowledgements: Funding was provided by the National Institute on Drug Abuse (P50DA039838, R01DA039901, R01DA015697), National Institute on Alcohol Abuse and Alcoholism (R01AA023187), National Heart Lung and Blood Institute (R01HL125440), and National Institute of Biomedical Imaging and Bioengineering (U54EB020404).

Supplementary Material

A Lagged Treatment Effects

A.1 Connection to Treatment Blips in the Structural Nested Mean Model

This Supplement connects a generalization of the structural nested mean model (SNMM; Robins 1989, 1994) to the lag k treatment effect defined in Section 2.3. In particular, consider a causal effect or treatment "blip" as defined by the SNMM framework (Robins 1994, Section 3a), with a minor departure in choosing the random reference treatment regime. We show how these effects are additive on the conditional mean of the potential proximal response. We conclude by connecting this particular SNMM generalization to the lag k moderated effect (2) considered throughout the paper.

The typical reference treatment regime used to define the treatment "blip" functions under the SNMM framework, is a prespecified non-random reference regime; here instead our reference treatment regime is stochastic and will match the conditional distribution of the treatments given history in the data generating distribution. In particular suppose that in the data generating distribution $\Pr(A_t = 1 \mid \bar{Y}_t = \bar{y}_t, \bar{X}_t = \bar{x}_t, \bar{A}_{t-1} = \bar{a}_{t-1}] = \rho_t(1 \mid h_t)$ for each t and where $h_t = (\bar{y}_t, \bar{x}_t, \bar{a}_{t-1})$. Then the reference treatment regime for the potential treatment is given, for each t, by $\Pr(A_t(\bar{a}_{t-1}) = 1 \mid H_t(\bar{a}_{t-1}) = h_t] = \rho_t(1 \mid h_t)$ (recall $H_t(\bar{a}_{t-1}) = (\bar{Y}_t(\bar{a}_{t-1}), \bar{X}_t(\bar{a}_{t-1}), \bar{A}_{t-1}(\bar{a}_{t-2}))$).

The treatment blip of fixed $a_t \in \{0, 1\}$ versus stochastic treatment $A_t(\bar{a}_{t-1})$ on the proximal response Y_{t+1} is

$$\mu_{t,t+1}(h_t, \bar{a}_t) = \mathrm{E}[Y_{t+1}(\bar{a}_t) - Y_{t+1}(\bar{a}_{t-1}, A_t(\bar{a}_{t-1})) \mid H_t(\bar{a}_{t-1}) = h_t].$$

The treatment blip of fixed $a_{t-1} \in \{0, 1\}$ versus stochastic treatment $A_{t-1}(\bar{a}_{t-2})$ on the proximal response Y_{t+1} is

$$\begin{aligned}&\mu_{t-1,t+1}(h_{t-1}, \bar{a}_{t-1})\\ &= \mathrm{E}[Y_{t+1}(\bar{a}_{t-1}, A_t(\bar{a}_{t-1})) - Y_{t+1}(\bar{a}_{t-2}, A_{t-1}(\bar{a}_{t-2}), A_t(\bar{a}_{t-2}, A_{t-1}(\bar{a}_{t-2}))) \mid H_{t-1}(\bar{a}_{t-2}) = h_{t-1}].\end{aligned}$$

The treatment blip for general $u \leq t$ is defined similarly but with an increase in notation. However notice if we denote $A_2(A_1)$ by A_2 and so on with $A_t(\bar{A}_{t-1})$ denoted by

A_t, and we denote $A_{u+1}(\bar{A}_{u-1}, a)$ by $A_{u+1}^{a_u=a}$, $A_{u+2}(\bar{A}_{u-1}, a, A_{u+1}^{a_u=a})$ by $A_{u+2}^{a_u=a}$ and so on with $A_t(\bar{A}_{u-1}, a, A_{u+1}^{a_u=a}, \ldots, A_{t-1}^{a_u=a})$ by $A_t^{a_u=a}$ then we have the compact form

$$\mu_{u,t+1}(H_u(\bar{A}_{u-1}), \bar{A}_{u-1}, a) = \mathrm{E}\big[Y_{t+1}(\bar{A}_{u-1}, a_u, A_{u+1}^{a_u=a}, \ldots, A_t^{a_u=a}) - Y_{t+1}(\bar{A}_t) \mid \bar{H}_u(\bar{A}_{u-1})\big]. \tag{9}$$

Assume consistency and sequential ignorability. Then

$$\begin{aligned}
&\mathrm{E}\big[Y_{t+1}(\bar{A}_{u-1}, a, A_{u+1}^{a_u}, \ldots, A_t^{a_u=a}) \,\big|\, H_u(\bar{A}_{u-1})\big] \\
&= \mathrm{E}\big[Y_{t+1}(\bar{A}_{u-1}, a, A_{u+1}^{a_u=a}, \ldots, A_t^{a_u=a}) \,\big|\, H_u(\bar{A}_{u-1}), A_u = a_u\big] \\
&= \mathrm{E}\big[Y_{t+1}(\bar{A}_{u-1}, A_u, A_{u+1}^{a_u=A_u}, \ldots, A_t^{a_u=A_u}) \,\big|\, H_u(\bar{A}_{u-1}), A_u = a_u\big] \\
&= \mathrm{E}\big[Y_{t+1}(\bar{A}_t) \mid H_u(\bar{A}_{u-1}), A_u = a_u\big]
\end{aligned}$$

where the first equality follows from the consistency and sequential ignorability assumptions (recall that $H_u = H_u(\bar{A}_{u-1})$) and the last two equalities follow by the definitions of $A_j^{a_u}$ and A_j. Thus under sequential ignorability, the treatment blip satisfies

$$\mathrm{E}[\mu_{u,t+1}(H_u(\bar{A}_{u-1}), \bar{A}_u) \mid H_u(\bar{A}_{u-1})] = 0, \tag{10}$$

for each $u = 1, \ldots, t$ and $t = 1, \ldots, T$. The lag k treatment effect (2) can be expressed as the expected contrast of the treatment blips (9):

$$\begin{aligned}
&\mathrm{E}\big[\mu_{t,t+k}(H_t(\bar{A}_{t-1}), \bar{A}_{t-1}, 1) - \mu_{t,t+k}(H_t(\bar{A}_{t-1}), \bar{A}_{t-1}, 0) \mid S_{kt}(\bar{A}_{t-1})\big] \\
&= \mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, 1, A_{t+1}^{a_u=1}, \ldots, A_{t+k-1}^{a_u=1}) - Y_{t+k}(\bar{A}_{t-1}, 0, A_{t+1}^{a_u=0}, \ldots, A_{t+k-1}^{a_u=0}) \,\big|\, S_{kt}(\bar{A}_{t-1})\big],
\end{aligned} \tag{11}$$

given the candidate moderators $S_{kt}(A_{t-1})$.

As in (Robins 1989, 1994) the SNMM treatment blips are related to the conditional mean of $Y_{t+1}(\bar{a}_t)$ given $H_t(\bar{a}_{t-1})$ by way a telescoping sum. For clarity we first provide the sum for $t = 3$.

$$\begin{aligned}
&\mathrm{E}[Y_4(\bar{a}_3) \mid H_3(\bar{a}_2) = h_3] \\
&= \mathrm{E}[Y_4(\bar{a}_3) - Y_4(\bar{a}_2, A_3(\bar{a}_2)) \mid H_3(\bar{a}_2) = h_3] \\
&\quad + \mathrm{E}[Y_4(\bar{a}_2, A_3(\bar{a}_2)) \mid H_3(\bar{a}_2) = h_3] - \mathrm{E}[Y_4(\bar{a}_2, A_3(\bar{a}_2)) \mid H_2(\bar{a}_1) = h_2] \\
&\quad + \mathrm{E}[Y_4(\bar{a}_2, A_3(\bar{a}_2)) - Y_4(a_1, A_2(a_1), A_3(a_1, A_2(a_1))) \mid H_2(\bar{a}_1) = h_2] \\
&\quad + \mathrm{E}[Y_4(a_1, A_2(a_1), A_3(a_1, A_2(a_1))) \mid H_2(\bar{a}_1) = h_2] - \mathrm{E}[Y_4(a_1, A_2(a_1), A_3(a_1, A_2(a_1))) \mid H_1 = h_1]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathrm{E}\big[Y_4(a_1, A_2(a_1), A_3(a_1, A_2(a_1))) - Y_4(\bar{A}_3) \mid H_1 = h_1\big] \\
&+ \mathrm{E}\big[Y_4(\bar{A}_3) \mid H_1 = h_1\big] - \mathrm{E}\big[Y_4(\bar{A}_3)\big] \\
&+ \mathrm{E}\big[Y_4(\bar{A}_3)\big].
\end{aligned}
$$

Denote $A_{u+1}(\bar{a}_{u-1}, A_u(\bar{a}_{u-1}))$ by $A_{u+1}^{\bar{a}_{u-1}}$, $A_{u+2}(\bar{a}_{u-1}, A_u(\bar{a}_{u-1}), A_{u+1}^{\bar{a}_{u-1}})$ by $A_{u+2}^{\bar{a}_{u-1}}$ and so on with $A_t(\bar{a}_{u-1}, A_u(\bar{a}_{u-1}), A_{u+1}^{a_{u-1}}, \ldots, A_{t-1}^{\bar{a}_{u-1}})$ by $A_t^{a_{u-1}}$. Using this compact notation the treatment blips in (9) can be rewritten as

$$
\mu_{u,t+1}(h_u, \bar{a}_u) = \mathrm{E}\big[Y_{t+1}(\bar{a}_u, A_{u+1}^{\bar{a}_u}, \ldots, A_t^{\bar{a}_u}) - Y_{t+1}(\bar{a}_{u-1}, A_u^{\bar{a}_{u-1}}, \ldots, A_t^{\bar{a}_{u-1}}) \mid \bar{H}_u(\bar{a}_{u-1}) = h_u\big].
$$

The telescoping sum for general t using this compact notation is

$$
\begin{aligned}
&\mathrm{E}[Y_{t+1}(\bar{a}_t) \mid H_t(\bar{a}_{t-1}) = h_t] \\
=\ &\mathrm{E}\big[Y_{t+1}(\bar{a}_t) - Y_{t+1}(\bar{a}_{t-1}, A_t^{\bar{a}_{t-1}}) \mid H_t(\bar{a}_{t-1}) = h_t\big] \\
&+ \mathrm{E}[Y_{t+1}(\bar{a}_{t-1}, A_t^{\bar{a}_{t-1}}) \mid H_t(\bar{a}_{t-1}) = h_t] - \mathrm{E}[Y_{t+1}(\bar{a}_{t-1}, A_t^{\bar{a}_{t-1}}) \mid H_{t-1}(\bar{a}_{t-2}) = h_{t-1}] \\
&+ \mathrm{E}[Y_{t+1}(\bar{a}_{t-1}, A_t^{\bar{a}_{t-1}}) - Y_{t+1}(\bar{a}_{t-2}, A_{t-1}^{\bar{a}_{t-2}}, A_t^{\bar{a}_{t-2}}) \mid H_{t-1}(\bar{a}_{t-2}) = h_{t-1}] \\
&+ \mathrm{E}[Y_{t+1}(\bar{a}_{t-2}, A_{t-1}^{\bar{a}_{t-2}}, A_t^{\bar{a}_{t-2}}) \mid H_{t-1}(\bar{a}_{t-2}) = h_{t-1}] - \mathrm{E}[Y_{t+1}(\bar{a}_{t-2}, A_{t-1}^{\bar{a}_{t-2}}, A_t^{\bar{a}_{t-2}}) \mid H_{t-2}(\bar{a}_{t-3}) = h_{t-2}] \\
&\cdots \\
&+ \mathrm{E}[Y_{t+1}(a_1, A_2^{a_1}, \cdots, A_t^{a_1}) - Y_{t+1}(\bar{A}_t) \mid H_1 = h_1] \\
&+ \mathrm{E}[Y_{t+1}(\bar{A}_t) \mid H_1 = h_1] - \mathrm{E}[Y_{t+1}(\bar{A}_t)] \\
&+ \mathrm{E}[Y_{t+1}(\bar{A}_t)] \\
=\ &\mathrm{E}[Y_{t+1}(\bar{A}_t)] + \sum_{u=1}^{t} \mu_{u,t+1}(h_u, \bar{a}_u) + \sum_{u=1}^{t} \epsilon_{u,t+1}(h_u, \bar{a}_{u-1}), \qquad (12)
\end{aligned}
$$

where

$$
\begin{aligned}
\epsilon_{u,t+1}(h_u, \bar{a}_{u-1}) = \ &\mathrm{E}[Y_{t+1}(\bar{a}_{u-1}, A_u^{\bar{a}_{u-1}}, \ldots, A_t^{\bar{a}_{u-1}}) \mid H_u(\bar{a}_{u-1}) = h_u] \\
&- \mathrm{E}[Y_{t+1}(\bar{a}_{u-1}, A_u^{\bar{a}_{u-1}}, \ldots, A_t^{\bar{a}_{u-1}}) \mid H_{u-1}(\bar{a}_{u-2}) = h_{u-1}],
\end{aligned}
$$

are nuisance functions that satisfy the constraint $\mathrm{E}[\epsilon_{u,t+1}(H_u(\bar{a}_{u-1}), \bar{a}_{u-1}) \mid H_{u-1}(\bar{a}_{u-2})] = 0$, for each $\bar{a}_{u-1} \in \mathcal{A}_{u-1}$, $u = 1, \ldots, t$ and $t = 1, \ldots, T$.

A.2 Identification from Data

Here we derive the expression (3) of the lag k treatment effect (2). This is done under the consistency, positivity and sequential ignorability conditions described in Section 2.3.

To derive expression (3) for the lag k treatment effect (2), we show that

$$\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \,\big|\, S_{kt}(\bar{A}_{t-1})\big] = \mathrm{E}\big[\mathrm{E}[Y_{t+k} \mid A_t = a, H_t] \,\big|\, S_{kt}\big]$$

and

$$\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \mid S_{kt}(\bar{A}_{t-1})\big] = \mathrm{E}\left[\frac{1(A_t = a)}{\rho_t(a \mid H_t)} Y_{t+k} \,\middle|\, S_{kt}\right]$$

for $a \in \{0, 1\}$.

First recall that by consistency, $H_t = H_t(\bar{A}_{t-1})$ and $S_{kt} = S_{kt}(\bar{A}_{t-1})$. Second recall the definition of $A_{t+j}^{a_t=a}$, where in particular $A_{t+1}^{a_t=a}$ denotes $A_{t+1}(\bar{A}_{t-1}, a)$, $A_{t+2}^{a_t=a}$ denotes $A_{t+2}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a})$ and so on, with $A_{t+k-1}(\bar{A}_{t-1}, a, A_{t+1}^{a_t}, \ldots, A_{t+k-2}^{a_t=a})$ denoted by $A_{t+k-1}^{a_t=a}$). So for each $j = 1, \ldots, T - t + 1$, sequential ignorability implies that $A_{t+j}^{a_t=a}, a \in \{0, 1\}$ is independent of A_t given H_t. We have

$$\begin{aligned}
&\mathrm{E}[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \mid S_{kt}(\bar{A}_{t-1})] \\
&= \mathrm{E}\big[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \mid H_t(\bar{A}_{t-1})\big] \,\big|\, S_{kt}(\bar{A}_{t-1})\big] \\
&= \mathrm{E}\big[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \mid H_t\big] \,\big|\, S_{kt}\big] \\
&= \mathrm{E}\big[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \mid H_t, A_t = a\big] \,\big|\, S_{kt}\big] \\
&= \mathrm{E}\big[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, A_t, A_{t+1}^{a_t=A_t}, \ldots, A_{t+k-1}^{a_t=A_t}) \mid H_t, A_t = a\big] \,\big|\, S_{kt}\big] \\
&= \mathrm{E}\big[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, A_t, A_{t+1}, \ldots, A_{t+k-1}) \mid H_t, A_t = a\big] \,\big|\, S_{kt}\big] \\
&= \mathrm{E}\big[\mathrm{E}[Y_{t+k} \mid H_t, A_t = a] \,\big|\, S_{kt}\big],
\end{aligned}$$

where the second equality holds by consistency, the third by sequential ignorability and the fifth follows from the definition of $A_{t+j}^{a_t=a}$ implying that $A_{t+j}^{a_t=A_t} = A_{t+j}$.

Next note that, by sequential ignorability, $\mathrm{E}[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \mid H_t]\,\mathrm{E}[1(A_t = a) \mid H_t]$ is equal to $\mathrm{E}[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a})\, 1(A_t = a) \mid H_t]$. We have

$$\begin{aligned}
&\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \,\big|\, S_{kt}(\bar{A}_{t-1})\big] \\
&= \mathrm{E}\left[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \,\big|\, H_t\big] \,\middle|\, S_{kt}\right] \\
&= \mathrm{E}\left[\mathrm{E}\big[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=a}, \ldots, A_{t+k-1}^{a_t=a}) \,\big|\, H_t\big] \frac{\mathrm{E}[1(A_t = a) \mid H_t]}{\rho_t(a \mid H_t)} \,\middle|\, S_{kt}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\mathrm{E}\left[Y_{t+k}(\bar{A}_{t-1}, a, A_{t+1}^{a_t=A_t}, \ldots, A_{t+k-1}^{a_t=A_t})\frac{1(A_t=a)}{\rho_t(a \mid H_t)} \,\middle|\, H_t\right] \,\middle|\, S_{kt}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[Y_{t+k}\frac{1(A_t=a)}{\rho_t(a \mid H_t)} \,\middle|\, H_t\right] \,\middle|\, S_{kt}\right] \\
&= \mathrm{E}\left[Y_{t+k}\frac{1(A_t=a)}{\rho_t(a \mid H_t)} \,\middle|\, S_{kt}\right]
\end{aligned}
$$

B Model Specification

This supplement outlines why the treatment effect at a given lag can be modelled without consideration of treatment effect models at other lags. We also provide a simple example of how models for remaining components of the conditional mean response models (e.g., $\mathrm{E}[Y_{t+k} \mid H_t, A_t = 0]$ or $\mathrm{E}[Y_{t+k} \mid H_t]$ at different lags k) constrain one another and are constrained by and constrain the treatment effect models. These considerations lead us to avoid assumptions concerning the correctness of conditional mean response models as in R2.

From (11), we know that the lag k effect depends on only one of the SNMM treatment blips (9). From (12) these blips are in turn additive on the conditional mean of the potential response. Provided that this conditional mean is not *a priori* restricted to certain values in $(-\infty, \infty)$, the treatment blips do not constrain one another (Robins et al. 2000, Theorem 8.6). This implies the same result for the lag k effect; that is, the treatment effects at different lags can be specified separately, with each lag-specific model imposing no constraints on the models chosen for the treatment effects at the remaining lags.

As an example, here we provide an illustration of how a model chosen for the lag 1 conditional mean response $\mathrm{E}[Y_{t+1} \mid A_t, H_t]$ constrains the form of the treatment effects at lag 2. Consider the simple example in which the treatments are binary, randomized with probability .5. Suppose we model the conditional mean of the response, $\mathrm{E}[Y_{t+1} \mid A_t, H_t]$ by $\alpha_{10} + \alpha_{11}Z_t + \alpha_{12}A_t$, where Z_t is an binary variable influenced by A_{t-1}. Further suppose that we model the lag 2 treatment effect, $\mathrm{E}[Y_{t+1} \mid A_{t-1} = 1, H_{t-1}] - \mathrm{E}[Y_{t+1} \mid A_{t-1} = 0, H_{t-1}]$ by a linear model $H_{t-1}^{\intercal}\beta_2$. Unfortunately in general these two models are inconsistent; they cannot both be correct. To see this, suppose that unbeknownst to us, $\Pr[Z_t = 1 \mid H_{t-1}] = 1/(1 + \exp(Y_{t-1} + A_{t-1}))$. Now if the first model is correct then the true lag-2 treatment effect should satisfy

$$
\begin{aligned}
&\mathrm{E}[Y_{t+1} \mid A_{t-1} = 1, H_{t-1}] - \mathrm{E}[Y_{t+1} \mid A_{t-1} = 0, H_{t-1}] \\
= \; &\mathrm{E}[\mathrm{E}[Y_{t+1} \mid A_t, H_t] \mid A_{t-1} = 1, H_{t-1}] - \mathrm{E}[\mathrm{E}[Y_{t+1} \mid A_t, H_t] \mid A_{t-1} = 0, H_{t-1}]
\end{aligned}
$$

$$= \alpha_{11}\{\Pr[Z_t = 1 \mid A_{t-1} = 1, H_{t-1}] - \Pr(Z_t = 1 \mid A_{t-1} = 0, H_{t-1})\} + \alpha_{12}.5$$
$$= \alpha_{11}\left\{\frac{1}{1+e^{Y_{t-1}+1}} - \frac{1}{1+e^{Y_{t-1}}}\right\} + \alpha_{12}.5.$$

In general since the conditional probability of $Z_t = 1$ is constrained to $[0, 1]$, this expression will be non-linear in H_{t-1}. So these lag 2 treatment effect and the lag 1 conditional mean response models cannot both be true.

This example shows that both parsimony in the treatment effect models and correctness in the models for the conditional mean response is difficult to achieve in the presence of binary (or more generally non-continuous) response predictors. Two special scenarios in which models with main effect of the form $\tilde{S}_{kt}^{\intercal}\alpha_k$ might be coherent across different k arise when all variables in $\tilde{S}_{kt}$ are either (1) multivariate normal or (2) centered by their conditional mean so that $\tilde{S}_{kt} - \mathrm{E}[\tilde{S}_{kt} \mid H_{t-1}] = 0$. Both of these settings require strong restrictions or assumptions about the distribution of covariates. So in general we should prefer estimation methods where $\tilde{S}_{kt}^{\intercal}\alpha_k$ need only be a working model for $\mathrm{E}[Y_{t+k} \mid H_t]$ (or $\mathrm{E}[Y_{t+k} \mid A_t = 0, H_t]$, in the case of the regression method from Section 3).

C Large Sample Properties

In this supplement we derive the large sample properties stated in Section 3. Throughout we allow for the setting in which individuals are not always available as discussed in Section 4. In all three methods, we allow for the use of a working diagonal variance structure for $\mathrm{Cov}(Y_2, \ldots, Y_{T-k+1})$ in the estimating function given by $(v_{k2}(\tilde{S}_{k1};\gamma), \ldots, v_{k,T-k+1}(\tilde{S}_{k,T-k+1};\gamma))$. In all three methods we assume A1, as defined in (5) of Section 3. Recall that throughout we assume sequential ignorability. Other assumptions that may be used include the following conditions for the k's corresponding to the lags of interest.

A3 Working Diagonal Variance Structure: The functions $(v_{k2}(\tilde{S}_{k1};\gamma), \ldots, v_{k,T-k+1}(\tilde{S}_{k,T-k+1};\gamma))$ are functions of a vector parameter γ. Let $V_{kT}(\tilde{S}_{kT};\gamma)$ be a diagonal matrix with the v_{kj}'s on the diagonal. Suppose $\hat{\gamma}$ solves an estimating equation: $\mathbb{P}_n U_{\mathrm{V}}(\gamma) = 0$. Assume that, for a finite value of γ, say γ^*, there exists finite constants, $b_v > 0$ and B_v such that each $b_v < v_{kT}(\tilde{S}_{kT};\gamma^*) < B_v$ a.s. and $\sqrt{n}(\hat{\gamma} - \gamma^*) = \mathrm{E}[\dot{U}_{\mathrm{V}}(\gamma^*)]^{-1}\sqrt{n}(\mathbb{P}_n - P)U_{\mathrm{V}}(\gamma^*) + o_P(1)$ for a positive definite, finite, matrix, $\mathrm{E}[\dot{U}_{\mathrm{V}}(\gamma^*)]$. Assume $\sqrt{n}(\mathbb{P}_n - P)U_{\mathrm{V}}(\gamma^*)$ converges in distribution to a mean zero, Normal random vector with variance-covariance matrix given by $\mathrm{E}[U_{\mathrm{V}}(\gamma^*)^{\otimes 2}]$ which has finite entries. Assume that $\mathbb{P}_n \dot{U}_{\mathrm{V}}(\hat{\gamma})$ is a consistent

estimator of $\mathrm{E}[\dot{U}_{\mathrm{V}}(\gamma^*)]$.

A4 Restricted Working Variance: $v_{kt}(\tilde{S}_{kt};\gamma)$ is a function of $\tilde{S}_{kt}$ only through S_{kt}.

A5 Treatment Probability Model: Let $\rho_t(1 \mid H_t;\eta)$ be a correctly specified model for $\Pr(A_t = 1 \mid I_t = 1H_t)$. Let η^* be the true value of η; that is, $\Pr(A_t = 1 \mid I_t = 1, H_t) = \rho_t(1 \mid H_t;\eta^*)$. Assume that the estimator of η, say $\hat{\eta}$, satisfies $\mathbb{P}_n U_{\mathrm{D}}(\hat{\eta}) = 0$ and $\sqrt{n}(\hat{\eta} - \eta^*) = \mathrm{E}[\dot{U}_{\mathrm{D}}(\eta^*)]^{-1}\,\mathbb{P}_n U_{\mathrm{D}}(\eta^*) + o_P(1)$. Thus $\sqrt{n}(\hat{\eta} - \eta^*)$ converges in distribution to a mean zero, Normal random vector with variance-covariance matrix given by $\mathrm{E}[\dot{U}_{\mathrm{D}}(\eta^*)]^{-1}\,\mathrm{E}[U_{\mathrm{D}}(\eta^*)^{\otimes 2}](\mathrm{E}[\dot{U}_{\mathrm{D}}(\eta^*)]^{-1})^{\intercal}$ which has finite entries. Assume that $\mathbb{P}_n \dot{U}_{\mathrm{D}}(\hat{\eta})$ is a consistent estimator of $\mathrm{E}[\dot{U}_{\mathrm{D}}(\eta^*)]$. Assume there exists finite constants, $b_{\mathrm{D}} > 0$ and $B_{\mathrm{D}} < 1$ such that each $b_{\mathrm{D}} < \rho_t(1 \mid H_t;\eta^*) < B_{\mathrm{D}}$ a.s.

A6 Weight Stabilization Probability Model: Suppose the scalar, $\hat{\rho}$, solves an estimating equation: $\mathbb{P}_n U_{\mathrm{N}}(\rho) = 0$. Assume that, for a finite value of ρ, say ρ^* and $\sqrt{n}(\hat{\rho} - \rho^*) = \mathrm{E}[\dot{U}_{\mathrm{N}}(\rho^*)]^{-1}\sqrt{n}(\mathbb{P}_n - P)U_{\mathrm{N}}(\rho^*) + o_P(1)$ where the matrix, $\mathrm{E}[\dot{U}_{\mathrm{N}}(\rho^*)]$ is positive definite. Assume $\sqrt{n}(\mathbb{P}_n - P)U_{\mathrm{N}}(\rho^*)$ converges in distribution to a mean zero, Normal random vector with variance-covariance matrix given by $\mathrm{E}[U_{\mathrm{N}}(\rho^*)^{\otimes 2}]$ which has finite entries. Assume that $\mathbb{P}_n \dot{U}_{\mathrm{N}}(\hat{\rho})$ is a consistent estimator of $\mathrm{E}[\dot{U}_{\mathrm{N}}(\rho^*)]$. Assume $0 < \rho^* < 1$.

The rationale for excluding off-diagonal entries from the working variance-covariance matrix as in condition A3 is illustrated numerically in Section 6. Pepe and Anderson (1994) illustrate the bias that occurs when off-diagonal entries are included in the setting of generalized estimation equations with longitudinal data; see also Schildcrout and Heagerty (2005) for further discussion.

C.1 Proof of Proposition 3.1

Here we discuss the proof of Proposition 3.1 assuming A1, A3, A4 and R1; the proof assuming R2 instead of R1 is routine. Under A3 and limited availability, the routine regression estimating function based on (6) is

$$U_{\mathrm{R}}(\alpha_k, \beta_k; \hat{\gamma}) = \sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - A_t S_{kt}^{\intercal}\beta_k\Big) v_{kt}(\tilde{S}_{kt};\hat{\gamma})^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}, \tag{13}$$

for which the solution to $\mathbb{P}_n U_{\mathrm{R}}(\alpha_k, \beta_k; \hat{\gamma}) = 0$ gives the estimator

$$\begin{pmatrix} \hat{\alpha}_k \\ \hat{\beta}_k \end{pmatrix} = \left\{ \mathbb{P}_n \sum_t v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}^{\otimes 2} \right\}^{-1} \mathbb{P}_n \sum_t v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t Y_{t+k} \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}.$$

Assume the following for k lags of interest.

R3 All entries in $\{Y_{t+k}, \tilde{S}_{kt}\}_{t=1}^{t=T-k+1}$ have finite fourth moments.

R4 The matrices $\mathrm{E}[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt}^{\otimes 2}]$ and

$$\mathrm{E}\Big[\dot{U}_{\mathrm{R}}(\alpha_k, \beta_k; \gamma^*)\Big] = \mathrm{E}\left[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}^{\otimes 2}\right]$$

are both invertible.

Define

$$\begin{pmatrix} \alpha'_k \\ \beta'_k \end{pmatrix} = \Big\{ \mathrm{E}\Big[\dot{U}_{\mathrm{R}}(\alpha_k, \beta_k; \gamma^*)\Big] \Big\}^{-1} \mathrm{E}\left[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t Y_{t+k} \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}\right].$$

Then standard statistical arguments can be used to show that $\sqrt{n}(\hat{\alpha}_k - \alpha'_k, \hat{\beta}_k - \beta'_k)$ converges in distribution to a normal, mean zero, random vector with variance-covariance matrix given by

$$\Big\{ \mathrm{E}\Big[\dot{U}_{\mathrm{R}}(\alpha'_k, \beta'_k; \gamma^*)\Big] \Big\}^{-1} \Sigma_{\mathrm{R}}(\alpha'_k, \beta'_k; \gamma^*) \Big\{ \mathrm{E}\Big[\dot{U}_{\mathrm{R}}(\alpha'_k, \beta'_k; \gamma^*)\Big] \Big\}^{-1},$$

where

$$\Sigma_{\mathrm{R}}(\alpha_k, \beta_k; \gamma) = \mathrm{E}\left[\Big(U_{\mathrm{R}}(\alpha_k, \beta_k; \gamma) + \Sigma_{\mathrm{R,V}}(\alpha_k, \beta_k; \gamma)\, \mathrm{E}[\dot{U}_{\mathrm{V}}(\gamma)]^{-1} U_{\mathrm{V}}(\gamma)\Big)^{\otimes 2}\right]$$

and

$$\Sigma_{\mathrm{R,V}}(\alpha_k, \beta_k; \gamma) = \mathrm{E}\left[\sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - A_t S_{kt}^{\intercal}\beta_k\Big) I_t \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix} \frac{\dot{v}_{kt}(\tilde{S}_{kt}; \gamma)^{\intercal}}{v_{kt}(\tilde{S}_{kt}; \gamma)^2}\right]$$

and $\dot{v}_{kt}(\tilde{S}_{kt}; \gamma) = dv_{kt}(\tilde{S}_{kt}; \gamma)/d\gamma$. Note that if R2 were true then $\Sigma_{\mathrm{R,V}}$ would be a matrix of zeros greatly simplifying these displays. A consistent estimator of the variance-covariance matrix is given by

$$\Big\{ \mathbb{P}_n \dot{U}_{\mathrm{R}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}) \Big\}^{-1} \hat{\Sigma}_{\mathrm{R}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}) \Big\{ \mathbb{P}_n \dot{U}_{\mathrm{R}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}) \Big\}^{-1},$$

where

$$\hat{\Sigma}_{\mathrm{R}}(\alpha_k, \beta_k; \gamma) = \mathbb{P}_n \left[\Big(U_{\mathrm{R}}(\alpha_k, \beta_k; \gamma) + \hat{\Sigma}_{\mathrm{R,V}}(\alpha_k, \beta_k; \gamma) \{ \mathbb{P}_n \dot{U}_{\mathrm{V}}(\hat{\gamma}) \}^{-1} U_{\mathrm{V}}(\hat{\gamma}) \Big)^{\otimes 2} \right]$$

and

$$\hat{\Sigma}_{\mathrm{R,V}}(\alpha_k, \beta_k; \gamma) = \mathbb{P}_n \left[\sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \alpha_k - A_t S_{kt}^{\intercal} \beta_k \Big) I_t \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix} \frac{\dot{v}_{kt}(\tilde{S}_{kt}; \gamma)^{\intercal}}{v_{kt}(\tilde{S}_{kt}; \gamma)^2} \right].$$

The remaining issue is to show that $\beta_k' = \beta_k^*$. Since $\mathrm{E}[U_{\mathrm{R}}(\alpha_k', \beta_k'; \gamma^*)] = 0$,

$$\sum_{t=1}^{T-k+1} \mathrm{E} \left[\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \alpha_k' - A_t S_{kt}^{\intercal} \beta_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix} \right] = 0. \tag{14}$$

Using the fact that S_{kt} is a sub-vector of $\tilde{S}_{kt}$ we have

$$\begin{aligned}
0 = & \sum_{t=1}^{T-k+1} \mathrm{E} \Big[\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \alpha_k' - A_t S_{kt}^{\intercal} \beta_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt} \Big] \\
= & \sum_{t=1}^{T-k+1} \mathrm{E} \Big[\Big(\mathrm{E}[Y_{t+k} \mid A_t, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal} \alpha_k' - A_t S_{kt}^{\intercal} \beta_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt} \Big] \\
= & \sum_t \mathrm{E} \Big[\Big(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal} \alpha_k' - S_{kt}^{\intercal} \beta_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t \rho S_{kt} \Big] \\
& + \sum_t \mathrm{E} \Big[\Big(\mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal} \alpha_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} (1 - \rho) I_t S_{kt} \Big] \\
= & \sum_t \mathrm{E} \Big[(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - S_{kt}^{\intercal} \beta_k') v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} \rho I_t S_{kt} \Big] \\
& + \sum_t \mathrm{E} \Big[\Big(\mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal} \alpha_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt} \Big].
\end{aligned} \tag{15}$$

Also from (14) we have

$$\begin{aligned}
0 = & \sum_{t=1}^{T-k+1} \mathrm{E} \Big[\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \alpha_k' - A_t S_{kt}^{\intercal} \beta_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t A_t S_{kt} \Big] \\
= & \sum_t \mathrm{E} \Big[\Big(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal} \alpha_k' - S_{kt}^{\intercal} \beta_k' \Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} \rho I_t S_{kt} \Big].
\end{aligned}$$

Insert these results into (15) to obtain

$$0 = \sum_t \mathrm{E}\Big[\Big(\mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - \tilde{S}_{kt}^\intercal \alpha_k'\Big) v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1}(1-\rho) I_t S_{kt}\Big].$$

Now insert the above into (15) (recall $\rho \in (0,1)$); we have

$$0 = \sum_t \mathrm{E}\Big[(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - S_{kt}^\intercal \beta_k') v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt}\Big]. \tag{16}$$

If $\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t, I_t = 1] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t, I_t = 1] = S_{kt}^\intercal \beta_k^*$ then (16) implies equal to

$$\begin{aligned} 0 &= \sum_t \mathrm{E}\Big[(S_{kt}^\intercal \beta_k^* - S_{kt}^\intercal \beta_k') v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt}\Big] \\ &= \mathrm{E}\left[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt}^{\otimes 2}\right](\beta_k^* - \beta_k') \end{aligned}$$

and thus $\beta_k' = \beta_k^*$. Otherwise A4 implies that $v_{kt}(\tilde{S}_{kt}; \gamma^*)$ is only a function of S_{kt} thus from (16) we have

$$\begin{aligned} 0 = \sum_t \mathrm{E}\Big[\Big(&\mathrm{E}\big[\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] \,\big|\, I_t = 1, S_{kt}\big] \\ &- S_{kt}^\intercal \beta_k'\Big) v_{kt}(S_{kt}; \gamma^*)^{-1} I_t S_{kt}\Big] \\ = \sum_t \mathrm{E}\big[(S_{kt}^\intercal \beta_k^* - S_{kt}^\intercal \beta_k') v_{kt}(S_{kt}; \gamma^*)^{-1} I_t S_{kt}\big] \\ = \mathrm{E}\left[\sum_t v_{kt}(S_{kt}; \gamma^*)^{-1} I_t S_{kt}^{\otimes 2}\right](\beta_k^* - \beta_k') \end{aligned}$$

and thus $\beta_k' = \beta_k^*$.

C.2 Proof of Proposition 3.2

Here we discuss the proof of Proposition 3.2 assuming A1, A2, using a working diagonal variance matrix satisfying A3, permitting limited availability and with the use of a treatment probability model assuming A5; if C2 does not hold then we will assume C1 and A4 (see the

end of this subsection). The centered least squares estimating function based on (7) is

$$
\begin{aligned}
&U_{\mathrm{C}}(\alpha_k, \beta_k; \hat{\gamma}, \hat{\eta}) \\
&= \sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - (A_t - \rho_t(1 \mid H_t; \hat{\eta}))S_{kt}^{\intercal}\beta_k\Big) v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \hat{\eta}))S_{kt} \end{pmatrix}
\end{aligned} \tag{17}
$$

for which the solution to $\mathbb{P}_n U_{\mathrm{C}}(\alpha_k, \beta_k; \hat{\gamma}, \hat{\eta}) = 0$ gives the estimator

$$
\begin{pmatrix} \hat{\alpha}_k \\ \hat{\beta}_k \end{pmatrix} = \Big\{\mathbb{P}_n \dot{U}_{\mathrm{C}}(\hat{\gamma}, \hat{\eta})\Big\}^{-1} \mathbb{P}_n \sum_t v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t Y_{t+k} \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \hat{\eta}))S_{kt} \end{pmatrix}
$$

where

$$
\mathbb{P}_n \dot{U}_{\mathrm{C}}(\hat{\gamma}, \hat{\eta}) = \mathbb{P}_n \sum_t v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \hat{\eta}))S_{kt} \end{pmatrix}^{\otimes 2}.
$$

Assume the following for the k lags of interest.

C3 All entries in $\{Y_{t+k}, \tilde{S}_{kt}\}_{t=1}^{t=T-k+1}$ have finite fourth moments.

C4 The matrices $\mathrm{E}[\dot{U}_{\mathrm{C}}(\gamma^*, \eta^*)]$ and $\mathrm{E}[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t (1 - \rho_t(1 \mid H_t; \eta^*))\rho_t(1 \mid H_t; \eta^*) S_{kt}^{\otimes 2}]$ are invertible.

Define

$$
\begin{pmatrix} \alpha_k' \\ \beta_k' \end{pmatrix} = \Big\{\mathrm{E}\Big[\dot{U}_{\mathrm{C}}(\gamma^*, \eta^*)\Big]\Big\}^{-1} \mathrm{E}\left[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t Y_{t+k} \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \eta^*))S_{kt} \end{pmatrix}\right].
$$

Then standard statistical arguments can be used to show that $\sqrt{n}(\hat{\alpha}_k - \alpha_k', \hat{\beta}_k - \beta_k')$ converges in distribution to a normal, mean zero, random vector with variance-covariance matrix given by

$$
\Big\{\mathrm{E}\Big[\dot{U}_{\mathrm{C}}(\gamma^*, \eta^*)\Big]\Big\}^{-1} \Sigma_{\mathrm{C}}(\alpha_k', \beta_k'; \gamma^*, \eta^*) \Big\{\mathrm{E}\Big[\dot{U}_{\mathrm{C}}(\gamma^*, \eta^*)\Big]\Big\}^{-1},
$$

where $\Sigma_{\mathrm{C}}(\alpha_k, \beta_k; \gamma, \eta)$ is

$$
\mathrm{E}\Big[\Big(U_{\mathrm{C}}(\alpha_k, \beta_k; \gamma, \eta) + \Sigma_{\mathrm{C,D}}(\alpha_k, \beta_k; \gamma, \eta)\{\mathrm{E}[\dot{U}_{\mathrm{D}}(\eta)]\}^{-1} U_{\mathrm{D}}(\eta) + \Sigma_{\mathrm{C,V}}(\alpha_k, \beta_k; \gamma, \eta)\{\mathrm{E}[\dot{U}_{\mathrm{V}}(\gamma)]\}^{-1} U_{\mathrm{V}}(\gamma)\Big)^{\otimes 2}\Big],
$$

with

$$
\begin{aligned}
&\Sigma_{\mathrm{C,V}}(\alpha_k, \beta_k; \gamma, \eta) \\
&= \mathrm{E}\left[\sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \alpha_k' - (A_t - \rho_t(1 \mid H_t; \eta)) S_{kt}^{\intercal} \beta_k' \Big) I_t \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \eta)) S_{kt} \end{pmatrix} \frac{\dot{v}_{kt}(\tilde{S}_{kt}; \gamma^*)^{\intercal}}{v_{kt}(\tilde{S}_{kt}; \gamma^*)^2} \right],
\end{aligned}
$$

$\dot{v}_{kt}(\tilde{S}_{kt}; \gamma) = dv_{kt}(\tilde{S}_{kt}; \gamma)/d\gamma$,

$$
\begin{aligned}
\Sigma_{\mathrm{C,D}}(\alpha_k, \beta_k; \gamma, \eta) = \; & \mathrm{E}\left[S_{kt}^{\intercal} \beta_k v_{kt}(\tilde{S}_{kt}; \gamma)^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \eta)) S_{kt} \end{pmatrix} \dot{\rho}_t(1 \mid H_t; \eta)^{\intercal} \right] \\
& - \mathrm{E}\left[\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \alpha_k - (A_t - \rho_t(1 \mid H_t; \eta)) S_{kt}^{\intercal} \beta_k \Big) v_{kt}(\tilde{S}_{kt}; \gamma)^{-1} I_t \begin{pmatrix} 0 \\ S_{kt} \end{pmatrix} \dot{\rho}_t(1 \mid H_t; \eta)^{\intercal} \right],
\end{aligned}
$$

and $\dot{\rho}_t(1 \mid H_t; \eta) = d\rho_t(1 \mid \eta)/d\eta$.

A consistent estimator of this variance-covariance matrix is given by

$$
\Big\{ \mathbb{P}_n \dot{U}_{\mathrm{C}}(\hat{\gamma}, \hat{\eta}) \Big\}^{-1} \hat{\Sigma}_{\mathrm{C}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) \Big\{ \mathbb{P}_n \dot{U}_{\mathrm{C}}(\hat{\gamma}, \hat{\eta}) \Big\}^{-1},
$$

where

$$
\begin{aligned}
\hat{\Sigma}_{\mathrm{C}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) = \mathbb{P}_n \Bigg[\Bigg(& U_{\mathrm{C}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) + \hat{\Sigma}_{\mathrm{C,D}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) \{ \mathbb{P}_n \dot{U}_{\mathrm{D}}(\hat{\eta}) \}^{-1} U_{\mathrm{D}}(\hat{\eta}) \\
& + \hat{\Sigma}_{\mathrm{C,V}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) \{ \mathbb{P}_n \dot{U}_{\mathrm{V}}(\hat{\gamma}) \}^{-1} U_{\mathrm{V}}(\hat{\gamma}) \Bigg)^{\otimes 2} \Bigg],
\end{aligned}
$$

with

$$
\begin{aligned}
\hat{\Sigma}_{\mathrm{C,D}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) = \; & \mathbb{P}_n \left[S_{kt}^{\intercal} \hat{\beta}_k v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \hat{\eta})) S_{kt} \end{pmatrix} \dot{\rho}_t(1 \mid H_t; \hat{\eta})^{\intercal} \right] \\
& - \mathbb{P}_n \left[\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \hat{\alpha}_k - (A_t - \rho_t(1 \mid H_t; \hat{\eta})) S_{kt}^{\intercal} \hat{\beta}_k \Big) v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t \begin{pmatrix} 0 \\ S_{kt} \end{pmatrix} \dot{\rho}_t(1 \mid H_t; \hat{\eta})^{\intercal} \right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\hat{\Sigma}_{\mathrm{C,V}}(\hat{\alpha}_k, \hat{\beta}_k; \hat{\gamma}, \hat{\eta}) \\
&= \mathbb{P}_n \left[\sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal} \hat{\alpha}_k' - (A_t - \rho_t(1 \mid H_t; \hat{\eta})) S_{kt}^{\intercal} \hat{\beta}_k' \Big) I_t \begin{pmatrix} \tilde{S}_{kt} \\ (A_t - \rho_t(1 \mid H_t; \hat{\eta})) S_{kt} \end{pmatrix} \frac{\dot{v}_{kt}(\tilde{S}_{kt}; \hat{\gamma}^*)^{\intercal}}{v_{kt}(\tilde{S}_{kt}; \hat{\gamma}^*)^2} \right].
\end{aligned}
$$

It remains to show that $\beta_k' = \beta_k^*$. Re-expressing $\mathrm{E}[U_{\mathrm{C}}(\alpha_k', \beta_k'; \gamma^*, \eta^*)]$ gives

$$
\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{T-k+1}\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k' - (A_t - \rho_t(1 \mid H_t;\eta^*))S_{kt}^{\intercal}\beta_k'\Big)v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1}I_t(A_t - \rho_t(1 \mid H_t;\eta^*))S_{kt}\right] \\
&= \mathrm{E}\left[\sum_{t=1}^{T-k+1}\Big(\mathrm{E}[Y_{t+k} \mid A_t, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal}\alpha_k' - (A_t - \rho_t(1 \mid H_t;\eta^*))S_{kt}^{\intercal}\beta_k'\Big)\frac{I_t(A_t - \rho_t(1 \mid H_t;\eta^*))}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}\right] \\
&= \mathrm{E}\left[\sum_{t=1}^{T-k+1}\Big(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal}\alpha_k' - \rho_t(0 \mid H_t;\eta^*)S_{kt}^{\intercal}\beta_k'\Big)\frac{I_t\rho_t(0 \mid H_t;\eta^*)}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}A_t\right] \\
&\quad - \mathrm{E}\left[\sum_{t=1}^{T-k+1}\Big(\mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal}\alpha_k' + \rho_t(1 \mid H_t;\eta^*)S_{kt}^{\intercal}\beta_k'\Big)\frac{I_t\rho_t(1 \mid H_t;\eta^*)}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}(1 - A_t)\right] \\
&= \mathrm{E}\left[\sum_{t=1}^{T-k+1}\Big(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal}\alpha_k' - \rho_t(0 \mid H_t;\eta^*)S_{kt}^{\intercal}\beta_k'\Big)\frac{I_t\rho_t(0 \mid H_t;\eta^*)\rho_t(1 \mid H_t;\eta^*)}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}\right] \\
&\quad - \mathrm{E}\left[\sum_{t=1}^{T-k+1}\Big(\mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - \tilde{S}_{kt}^{\intercal}\alpha_k' + \rho_t(1 \mid H_t;\eta^*)S_{kt}^{\intercal}\beta_k'\Big)\frac{I_t\rho_t(1 \mid H_t;\eta^*)\rho_t(0 \mid H_t;\eta^*)}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}\right].
\end{aligned}
$$

Thus

$$
\begin{aligned}
0 &= \mathrm{E}[U_{\mathrm{C}}(\alpha_k', \beta_k'; \gamma^*, \eta^*)] \\
&= \mathrm{E}\left[\sum_{t=1}^{T-k+1}(\mathrm{E}[Y_{t+k} \mid A_t = 1, I_t = 1, H_t] - \mathrm{E}[Y_{t+k} \mid A_t = 0, I_t = 1, H_t] - S_{kt}^{\intercal}\beta_k')\right. \\
&\qquad \left.\times \frac{I_t\rho_t(1 \mid H_t;\eta^*)(1 - \rho_t(1 \mid H_t;\eta^*))}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}\right].
\end{aligned}
$$

Assume C2; that is, $\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t, I_t = 1] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t, I_t = 1] = S_{kt}^{\intercal}\beta_k^*$ for each t. Then from the above display

$$
\begin{aligned}
0 &= \mathrm{E}\left[\sum_{t=1}^{T-k+1}(S_{kt}^{\intercal}\beta_k^* - S_{kt}^{\intercal}\beta_k')\frac{I_t\rho_t(1 \mid H_t;\eta^*)(1 - \rho_t(1 \mid H_t;\eta^*))}{v_{kt}(\tilde{S}_{kt};\gamma^*)}S_{kt}\right], \\
&= \sum_t \mathrm{E}\Big[(S_{kt}^{\intercal}\beta_k^* - S_{kt}^{\intercal}\beta_k')v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1}I_t(1 - \rho_t(1 \mid H_t;\eta^*))\rho_t(1 \mid H_t;\eta^*)S_{kt}\Big] \\
&= \mathrm{E}\left[\sum_t v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1}I_t(1 - \rho_t(1 \mid H_t;\eta^*))\rho_t(1 \mid H_t;\eta^*)S_{kt}^{\otimes 2}\right](\beta_k^* - \beta_k')
\end{aligned}
$$

and thus $\beta_k' = \beta_k^*$. Otherwise we use A4, that is, $v_{kt}(\tilde{S}_{kt};\gamma^*)$ is only a function of S_{kt} and we

use C1, that is, $\rho_t(1 \mid H_t; \eta^*)$ is a only a function of S_{kt}, thus the argument from the previous paragraph implies

$$\begin{aligned} 0 &= \sum_t \mathrm{E}\big[(S_{kt}^\intercal \beta_k^* - S_{kt}^\intercal \beta_k') v_{kt}(S_{kt}; \gamma^*)^{-1} I_t (1 - \rho_t(1 \mid S_{kt}; \eta^*)) \rho_t(1 \mid S_{kt}; \eta^*) S_{kt}\big] \\ &= \mathrm{E}\left[\sum_t v_{kt}(S_{kt}; \gamma^*)^{-1} I_t (1 - \rho_t(1 \mid S_{kt}; \eta^*)) \rho_t(1 \mid S_{kt}; \eta^*) S_{kt}^{\otimes 2}\right] (\beta_k^* - \beta_k') \end{aligned}$$

and thus $\beta_k' = \beta_k^*$.

C.3 Proof of Proposition 3.3

Here we discuss the proof of Proposition 3.3 assuming A1, A2, using a working diagonal variance matrix satisfying A3, permitting limited availability and with the use of models assuming A5 and A6. If the analog of C2 does not hold (e.g. $\mathrm{E}[Y_{t+k} \mid A_t = 1, H_t, I_t = 1] - \mathrm{E}[Y_{t+k} \mid A_t = 0, H_t, I_t = 1] \neq S_{kt}^\intercal \beta_k^*$ for some t) then we assume A4.

The estimating equation based on (8) is

$$U_{\mathrm{W}}(\alpha_k, \beta_k; \hat{\gamma}, \hat{\eta}, \hat{\rho}) = \sum_{t=1}^{T-k+1} \Big(Y_{t+k} - \tilde{S}_{kt}^\intercal \alpha_k - A_t S_{kt}^\intercal \beta_k\Big) v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t w_t(A_t, H_t; \hat{\eta}, \hat{\rho}) \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix} \tag{18}$$

where $w_t(A_t, H_t; \eta, \rho) = \rho^{A_t}(1-\rho)^{1-A_t} / \rho_t(A_t \mid H_t; \eta)$. The solution to $\mathbb{P}_n U_{\mathrm{W}}(\alpha_k, \beta_k; \hat{\gamma}, \hat{\eta}, \hat{\rho}) = 0$ gives the estimator

$$\begin{pmatrix} \hat{\alpha}_k \\ \hat{\beta}_k \end{pmatrix} = \Big\{\mathbb{P}_n \dot{U}_{\mathrm{W}}(\hat{\gamma}, \hat{\eta}, \hat{\rho})\Big\}^{-1} \mathbb{P}_n \sum_t v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t w_t(A_t, H_t; \hat{\eta}, \hat{\rho}) Y_{t+k} \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}$$

where

$$\mathbb{P}_n \dot{U}_{\mathrm{W}}(\hat{\gamma}, \hat{\eta}, \hat{\rho}) = \mathbb{P}_n \sum_t v_{kt}(\tilde{S}_{kt}; \hat{\gamma})^{-1} I_t w_t(A_t, H_t; \hat{\eta}, \hat{\rho}) \begin{pmatrix} \tilde{S}_{kt} \\ A_t S_{kt} \end{pmatrix}^{\otimes 2}.$$

Assume the following for the k lags of interest.

W1 All entries in $\{Y_{t+k}, \tilde{S}_{kt}\}_{t=1}^{t=T-k+1}$ have finite fourth moments.

W2 The matrices $\mathrm{E}\,\dot{U}_{\mathrm{W}}(\gamma^*, \eta^*, \rho^*)$ and $\mathrm{E}\Big[\sum_t v_{kt}(\tilde{S}_{kt}; \gamma^*)^{-1} I_t S_{kt}^{\otimes 2}\Big]$ are invertible.

Define

$$\begin{pmatrix}\alpha_k' \\ \beta_k'\end{pmatrix} = \left\{\mathrm{E}\left[\dot{U}_{\mathrm{W}}(\gamma^*,\eta^*,\rho^*)\right]\right\}^{-1}\mathrm{E}\left[\sum_t v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1}I_t w_t(A_t,H_t;\eta^*,\rho^*)Y_{t+k}\begin{pmatrix}\tilde{S}_{kt}\\ A_tS_{kt}\end{pmatrix}\right].$$

Then standard statistical arguments can be used to show that $\sqrt{n}(\hat{\alpha}_k-\alpha_k', \hat{\beta}_k-\beta_k')$ converges in distribution to a normal, mean zero, random vector with variance-covariance matrix given by

$$\left\{\mathrm{E}\left[\dot{U}_{\mathrm{W}}(\gamma^*,\eta^*,\rho^*)\right]\right\}^{-1}\Sigma_{\mathrm{W}}(\alpha_k',\beta_k';\gamma^*,\eta^*,\rho^*)\left\{\mathrm{E}\left[\dot{U}_{\mathrm{W}}(\gamma^*,\eta^*,\rho^*)\right]\right\}^{-1},$$

where

$$\begin{aligned}\Sigma_{\mathrm{W}}(\alpha_k,\beta_k;\gamma,\eta,\rho) = \mathrm{E}\Bigg[\Bigg(&U_{\mathrm{W}}(\alpha_k,\beta_k;\gamma,\eta,\rho) + \Sigma_{\mathrm{W,V}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\{\mathrm{E}[\dot{U}_{\mathrm{V}}(\gamma)]\}^{-1}U_{\mathrm{V}}(\gamma)\\ &+ \Sigma_{\mathrm{W,D}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\{\mathrm{E}[\dot{U}_{\mathrm{D}}(\eta)]\}^{-1}U_{\mathrm{D}}(\eta)\\ &+ \Sigma_{\mathrm{W,N}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\{\mathrm{E}[\dot{U}_{\mathrm{N}}(\rho)]\}^{-1}U_{\mathrm{N}}(\rho)\Bigg)^{\otimes 2}\Bigg],\end{aligned}$$

with

$$\begin{aligned}&\Sigma_{\mathrm{W,V}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\\ &= \mathrm{E}\left[\sum_{t=1}^{T-k+1}\left(Y_{t+k}-\tilde{S}_{kt}^{\intercal}\alpha_k - A_tS_{kt}^{\intercal}\beta_k\right)I_t v_{kt}(\tilde{S}_{kt};\gamma)^{-1}w_t(A_t,H_t;\eta,\rho)\begin{pmatrix}\tilde{S}_{kt}\\ A_tS_{kt}\end{pmatrix}\frac{\dot{v}_{kt}(\tilde{S}_{kt};\gamma)^{\intercal}}{v_{kt}(\tilde{S}_{kt};\gamma)}\right],\end{aligned}$$

$\dot{v}_{kt}(\tilde{S}_{kt};\gamma) = dv_{kt}(\tilde{S}_{kt};\gamma)/d\gamma,$

$$\begin{aligned}&\Sigma_{\mathrm{W,D}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\\ &= \mathrm{E}\left[\sum_{t=1}^{T-k+1}\left(Y_{t+k}-\tilde{S}_{kt}^{\intercal}\alpha_k - A_tS_{kt}^{\intercal}\beta_k\right)I_t v_{kt}(\tilde{S}_{kt};\gamma)^{-1}w_t(A_t,H_t;\eta,\rho)\begin{pmatrix}\tilde{S}_{kt}\\ A_tS_{kt}\end{pmatrix}\frac{\dot{\rho}_t(A_t\mid H_t;\eta)^{\intercal}}{\rho_t(A_t\mid H_t;\eta)}\right],\end{aligned}$$

$\dot{\rho}_t(1\mid H_t;\eta) = d\rho_t(1\mid H_t;\eta)/d\eta$ and

$$\Sigma_{\mathrm{W,N}}(\alpha_k,\beta_k;\gamma,\eta,\rho) = \mathrm{E}\left[\sum_{t=1}^{T-k+1}\left(Y_{t+k}-\tilde{S}_{kt}^{\intercal}\alpha_k - A_tS_{kt}^{\intercal}\beta_k\right)I_t v_{kt}(\tilde{S}_{kt};\gamma)^{-1}\begin{pmatrix}\tilde{S}_{kt}\\ A_tS_{kt}\end{pmatrix}\frac{2A_t-1}{\rho_t(A_t\mid H_t;\eta)}\right].$$

Note that if R2 were true then $\Sigma_{\mathrm{R,V}}$, $\Sigma_{\mathrm{W,N}}$ and $\Sigma_{\mathrm{W,D}}$ would be matrices of zeros greatly simplifying the variance-covariance matrix.

A consistent estimator of the variance-covariance matrix is given by

$$\Big\{\mathbb{P}_n \dot{U}_{\mathrm{W}}(\hat{\gamma},\hat{\eta},\hat{\rho})\Big\}^{-1}\hat{\Sigma}_{\mathrm{W}}(\hat{\alpha}_k,\hat{\beta}_k;\hat{\gamma},\hat{\eta},\hat{\rho})\Big\{\mathbb{P}_n \dot{U}_{\mathrm{W}}(\hat{\gamma},\hat{\eta},\hat{\rho})\Big\}^{-1}, \tag{19}$$

where

$$\begin{aligned}\hat{\Sigma}_{\mathrm{W}}(\alpha_k,\beta_k;\gamma,\eta,\rho) = \mathbb{P}_n\Bigg[\Bigg(&U_{\mathrm{W}}(\alpha_k,\beta_k;\gamma,\eta,\rho) + \hat{\Sigma}_{\mathrm{W,V}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\{\mathbb{P}_n \dot{U}_{\mathrm{V}}(\gamma)\}^{-1}U_{\mathrm{V}}(\gamma)\\ &+ \hat{\Sigma}_{\mathrm{W,D}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\{\mathbb{P}_n \dot{U}_{\mathrm{D}}(\eta)\}^{-1}U_{\mathrm{D}}(\eta)\\ &+ \hat{\Sigma}_{\mathrm{W,N}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\{\mathbb{P}_n \dot{U}_{\mathrm{N}}(\rho)\}^{-1}U_{\mathrm{N}}(\rho)\Bigg)^{\otimes 2}\Bigg],\end{aligned}$$

with

$$\begin{aligned}&\hat{\Sigma}_{\mathrm{W,V}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\\ &= \mathbb{P}_n\Bigg[\sum_{t=1}^{T-k+1}\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - A_t S_{kt}^{\intercal}\beta_k\Big) I_t v_{kt}(\tilde{S}_{kt};\gamma)^{-1} w_t(A_t,H_t;\eta,\rho)\begin{pmatrix}\tilde{S}_{kt}\\ A_t S_{kt}\end{pmatrix}\frac{\dot{v}_{kt}(\tilde{S}_{kt};\gamma)^{\intercal}}{v_{kt}(\tilde{S}_{kt};\gamma)}\Bigg],\end{aligned}$$

$$\begin{aligned}&\hat{\Sigma}_{\mathrm{W,D}}(\alpha_k,\beta_k;\gamma,\eta,\rho)\\ &= \mathbb{P}_n\Bigg[\sum_{t=1}^{T-k+1}\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - A_t S_{kt}^{\intercal}\beta_k\Big) I_t v_{kt}(\tilde{S}_{kt};\gamma)^{-1} w_t(A_t,H_t;\eta,\rho)\begin{pmatrix}\tilde{S}_{kt}\\ A_t S_{kt}\end{pmatrix}\frac{\dot{\rho}_t(A_t \mid H_t;\eta)^{\intercal}}{\rho_t(A_t \mid H_t;\eta)}\Bigg]\end{aligned}$$

and

$$\hat{\Sigma}_{\mathrm{W,N}}(\alpha_k,\beta_k;\gamma,\eta,\rho) = \mathbb{P}_n\Bigg[\sum_{t=1}^{T-k+1}\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k - A_t S_{kt}^{\intercal}\beta_k\Big) I_t v_{kt}(\tilde{S}_{kt};\gamma)^{-1}\begin{pmatrix}\tilde{S}_{kt}\\ A_t S_{kt}\end{pmatrix}\frac{2A_t - 1}{\rho_t(A_t \mid H_t;\eta)}\Bigg].$$

It remains to show that $\beta_k' = \beta_k^*$. Since $\mathrm{E}[U_{\mathrm{W}}(\alpha_k',\beta_k';\gamma^*,\eta^*,\rho^*)] = 0$,

$$\begin{aligned}0 &= \sum_{t=1}^{T-k+1}\Big(Y_{t+k} - \tilde{S}_{kt}^{\intercal}\alpha_k' - A_t S_{kt}^{\intercal}\beta_k'\Big) v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1} I_t w_t(A_t,H_t;\eta^*,\rho^*)\begin{pmatrix}\tilde{S}_{kt}\\ A_t S_{kt}\end{pmatrix}\\ &= \sum_{t=1}^{T-k+1}\Big(\mathrm{E}[Y_{t+k} \mid A_t,H_t,I_t=1] - \tilde{S}_{kt}^{\intercal}\alpha_k' - A_t S_{kt}^{\intercal}\beta_k'\Big) v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1} I_t w_t(A_t,H_t;\eta^*,\rho^*)\begin{pmatrix}\tilde{S}_{kt}\\ A_t S_{kt}\end{pmatrix}\\ &= \sum_{t=1}^{T-k+1}\sum_{a\in\{0,1\}}\Big(\mathrm{E}[Y_{t+k} \mid A_t = a,H_t,I_t=1] - \tilde{S}_{kt}^{\intercal}\alpha_k' - a S_{kt}^{\intercal}\beta_k'\Big) v_{kt}(\tilde{S}_{kt};\gamma^*)^{-1} I_t \rho^{*a}(1-\rho^*)^{1-a}\begin{pmatrix}\tilde{S}_{kt}\\ a S_{kt}\end{pmatrix}\end{aligned}$$

where the last equality averages out over A_t. Now follow the exact same steps as in Supplement C.1, starting with (14) but with $\rho = \rho^*$, to obtain the result.

D Additional simulation results

Section 6 provides simulation results that considers performance of the estimators in terms of bias. Here we give the Monte Carlo standard deviation of the point estimates and the average standard error estimates for the three settings of Section 6 in Tables 5, 6 and 7, respectively.

Table 5: Estimation of the marginal proximal effect -0.8, averaging over an underlying moderator with coefficient β_{11}. Results are based on 1000 replicates with $n = T = 100$.

	$\beta_{11}^* = 0.2$		$\beta_{11}^* = 0.5$		$\beta_{11}^* = 0.8$	
Method	SD	ASE	SD	ASE	SD	ASE
Weighted	0.023	0.023	0.023	0.023	0.024	0.025
Centered	0.023	0.022	0.023	0.023	0.024	0.024
Routine	0.020	0.020	0.021	0.021	0.022	0.022

SD, standard deviation of the point estimate; ASE, average standard error estimate.

Table 6: Weighted least squares estimation of the proximal effect -0.8 under different stabilization schemes. Results are based on 1000 replicates with $n = T = 100$.

Method	SD	ASE
Weighted (i)	0.032	0.031
Weighted time–varying stabilizer (ii)	0.042	0.041

Table 7: Weighted least squares estimation of the proximal effect -0.8 under independence (as proposed in Section 5) and fixed AR(1) working correlation structures. Results are based on 1000 replicates with $n = 50$ and $T = 100$.

Method	SD	ASE
Weighted (i)	0.033	0.034
Weighted fixed AR(1) (ii)	0.031	0.031

E Code to Generate Numerical Results

The code used to generate numerical results in this paper can be obtained from

https://github.com/dalmiral/mHealthModeration

This includes the additional calculations necessary to correct standard errors for estimated weights, treatment probabilities, or small samples.