Statistics
in Medicine

Research Article

Received 9 October 2012, Accepted 3 June 2013 Published online 19 July 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5892

Time-varying effect moderation using the structural nested mean model: estimation using inverse-weighted regression with residuals

Daniel Almirall,[a*†] Beth Ann Griffin,[b] Daniel F. McCaffrey,[b] Rajeev Ramchand,[b] Robert A. Yuen[c] and Susan A. Murphy[a,c,d]

This article considers the problem of examining time-varying causal effect moderation using observational, longitudinal data in which treatment, candidate moderators, and possible confounders are time varying. The structural nested mean model (SNMM) is used to specify the moderated time-varying causal effects of interest in a conditional mean model for a continuous response given time-varying treatments and moderators. We present an easy-to-use estimator of the SNMM that combines an existing regression-with-residuals (RR) approach with an inverse-probability-of-treatment weighting (IPTW) strategy. The RR approach has been shown to identify the moderated time-varying causal effects if the time-varying moderators are also the sole time-varying confounders. The proposed IPTW+RR approach provides estimators of the moderated time-varying causal effects in the SNMM in the presence of an additional, auxiliary set of known and measured time-varying confounders. We use a small simulation experiment to compare IPTW+RR versus the traditional regression approach and to compare small and large sample properties of asymptotic versus bootstrap estimators of the standard errors for the IPTW+RR approach. This article clarifies the distinction between time-varying moderators and time-varying confounders. We illustrate the methodology in a case study to assess if time-varying substance use moderates treatment effects on future substance use.

Keywords: effect modification; time-varying covariates; time-varying treatment; time-varying exposure; time-varying confounding; inverse-probability-of-treatment weighting

1. Introduction

Across a wide spectrum of the behavioral, medical, and social sciences, there is considerable interest in examining research questions assessing the impact of time-varying treatments (or exposures) using longitudinal data. The methodology we discuss in this manuscript focuses on *time-varying causal effect moderation* [1,2], known as *time-varying causal effect modification* in the epidemiology literature [3–5]. In interventions research developing time-varying treatments or dynamic treatment regimes [6,7], examining time-varying moderators is valuable because it can be used to shed light on conceptual models, or to generate hypotheses, about tailoring variables used to guide the timing, sequencing, and duration of treatment over time.

In point-treatment studies, moderator variables specify for whom (or under what conditions) treatment is more or less effective [8,9]. In the study of time-varying treatments, time-varying moderators are variables that specify for whom (or under what conditions) both the initial treatment and the next step in treatment (e.g., treatment switch, augmentation, or dis/continuation) is more or less effective.

[a]*Institute for Social Research, University of Michigan, Ann Arbor, MI 48104, U.S.A.*
[b]*RAND Corporation, Pittsburgh, PA, U.S.A.*
[c]*Department of Statistics, University of Michigan, Ann Arbor, MI 48104, U.S.A.*
[d]*Department of Psychiatry, University of Michigan, Ann Arbor, MI 48104, U.S.A.*
**Correspondence to: Daniel Almirall, Institute for Social Research, University of Michigan, Suite 214NU, 426 Thompson Street, Ann Arbor, MI 48104, U.S.A.*
[†]*E-mail: dalmiral@umich.edu*

A key distinction between point-treatment moderators and time-varying moderators is that time-varying moderators may be measured during, or in response to, prior treatment.

To illustrate time-varying causal effect moderation, consider a simplified version of our motivating example. The aim is to examine the effect of time-varying sequences (A_1, A_2) of adolescent substance use treatment (A_1 = yes(1)/no(0) initial treatment; A_2 = yes(1)/no(0) later treatment) on post-treatment substance use frequency (Y). (For simplicity, in this manuscript, we consider a non-time-varying end-of-treatment outcome Y.) One set of possible questions compares the population mean of Y under different sequences of treatment, such as 'What is the average effect of always receiving treatment $(A_1, A_2) = (1, 1)$ versus receiving only initial treatment $(A_1, A_2) = (1, 0)$?' These are called marginal time-varying treatment effects, which have received considerable methodological attention by means of the marginal structural model (MSM) [10–14]. In this manuscript, we are interested in asking more detailed questions, that is, concerning the moderated (or conditional) effects of time-varying treatment. Examples are 'How does the average effect of always receiving treatment $(1, 1)$ versus receiving only initial treatment $(1, 0)$ differ as a function of the evolving frequency of substance use prior to (S_0) and during (S_1) initial treatment?' and 'How does the average effect of receiving only initial treatment $(1, 0)$ versus not receiving treatment $(0, 0)$ differ as a function of the frequency of use prior to (S_0) treatment?' In these examples, substance use (S_0, S_1) is a candidate moderator of the impact of treatment (A_1, A_2) on Y. Understanding these effects is interesting for clinical practice, for example, because they provide information about the value (or need) for additional substance use treatment conditional on how the adolescent has responded to prior treatment. The marginal effects, on the other hand, provide information about additional treatment on average for the entire population.

An important challenge in the estimation of time-varying causal effects is that adjusting naively for other time-varying covariates may result in bias if the covariates are themselves impacted by prior treatment [15–18]. In observational studies examining time-varying effect moderation using traditional regression techniques, this problem arises from adjusting for two types of time-varying covariates: First, these analyses require adjusting for time-varying covariates, which are candidate moderators, because by definition, the aim is to understand the impact of time-varying treatments *conditional* on (i.e., as a function of) candidate time-varying moderators. Second, in observational studies examining time-varying effect moderation, data analysts often adjust for time-varying covariates, which may be directly related to both subsequent treatment and outcome to reduce or eliminate time-varying confounding bias. However, in either case—that is, whether adjusting for a time-varying covariate because it is a candidate moderator or whether adjusting for a time-varying covariate to eliminate bias due to possible time-varying confounding—the time-varying covariate may itself be impacted by prior treatment, possibly leading to bias in the estimated time-varying effects of interest.

To better appreciate the problems with adjusting for time-varying covariates and to set the context for the proposed methodology, consider a naive extension of the standard treatment-moderator interactions approach [8] for studying effect moderation in which a regression model such as the following one is used:

$$\begin{aligned} E(Y \mid X_0, S_0, A_1, X_1, S_1, A_2) = \beta_0 &+ \eta_1 X_0 + \eta_2 S_0 + \beta_{1,1} A_1 + \beta_{1,2} A_1 S_0 + \eta_3 A_1 X_0 \\ &+ \eta_4 X_1 + \eta_5 S_1 + \beta_{2,1} A_2 + \beta_{2,2} A_2 S_0 + \beta_{2,3} A_2 S_1 \\ &+ \eta_6 A_2 X_0 + \eta_7 A_2 X_1. \end{aligned} \tag{1}$$

where (X_0, X_1) are time-varying confounders. In this traditional regression approach, the analyst adjusts for (S_0, S_1) (e.g., substance use) because, as a candidate time-varying moderator, it is of particular scientific interest, whereas the analyst adjusts for (X_0, X_1) (e.g., social support) because it is a candidate time-varying confounder possibly associated with both subsequent treatment and Y.

Unfortunately, using this type of regression creates at least three problems for making causal inferences about the moderated time-varying effects of interest, in particular with the effects of A_1 given S_0 (the parameters $\beta_{1,1}$ and $\beta_{1,2}$). First, conditioning on S_1 and X_1 cuts off any portion of the effect of A_1 on Y that occurs via S_1 or X_1 (including moderated effects). Second, there are likely common, possibly unknown, causes of (S_1, X_1) and Y which, by conditioning on (S_1, X_1) (possible outcomes of treatment A_1), may introduce bias in the coefficients of the A_1 terms. The result is that the moderated effects of A_1 may appear to be (un)correlated with Y simply because A_1 impacts (S_1, X_1) and because both (S_1, X_1) and Y are affected by a common cause. This problem, known as collider bias [19], is particularly subtle; intuitive discussions of it are given in [20] and [1]. The third problem is that a regression approach such as (1) forces the analyst to consider time-varying effect moderation by (X_0, X_1) (consider (η_3, η_6, η_7))

even though it is not of scientific interest! This is because failure to model effect moderation by (X_0, X_1), should it be present, leads to mis-specification of the regression model; this, in turn, leads to bias in the moderated effects of (A_1, A_2) on Y because, by definition as candidate time-varying confounders, the X_t's are correlated with the A_t's. The practical implication of this is that the meaning of the parameters describing the effect of treatment conditional on S_t may change; the parameters will describe, instead, the effect of time-varying treatment conditional on both X_t's and S_t's. The third problem is especially problematic in most observational study settings, such as ours, where the list of observed time-varying confounders, the X_t's, is significantly larger than the list of time-varying moderators. The data analyst interested in moderated time-varying effects would benefit from an alternative to model (1) that gets around these barriers to causal inference.

Importantly, the three previous problems are not the result of unknown or unmeasured time-varying confounders (i.e., bias may occur even when (X_0, S_0, X_1, S_1) are the only time-varying covariates associated with treatment and outcome); indeed, these three problems can occur even when A_1 and/or A_2 are randomized such as in a sequential multiple assignment randomized trial (SMART) [21]. Further, the first two problems are not due to model mis-specification (e.g., bias may occur even in correctly specified models for the conditional mean of Y; for discussion, see [1]).

The structural nested mean model (SNMM) [17] provides a principled alternative to model (1). The SNMM specifies the moderated time-varying causal effects of interest in a conditional mean model for a continuous response given time-varying treatments and candidate moderators. The structure of the SNMM provides a clue for how to condition on (S_0, S_1) to avoid the first two problems mentioned previously. With regard to the third problem, one can use the SNMM to specify a model for only the time-varying moderated effects of interest (in our case, time-varying effects conditional on S_t); this is a model that averages over all other time-varying covariates, including X_t. Therefore, the SNMM does not require adjusting for the X_t's in the conditional mean model itself. Rather, the candidate time-varying confounders can be dealt with in the *estimation* of the causal parameters (via weighting, see later) of the SNMM but not as part of the conditional mean model itself *defining* the causal effects of interest.

This article contributes to the methodological literature by extending and illustrating the use of an estimator of the SNMM that combines an existing, easy-to-use regression-with-residuals (RR) approach together with an inverse-probability-of-treatment weighting (IPTW) strategy. In the previous work [1, 2, 22–24], the RR approach (when used without IPTW) has been used to estimate the moderated time-varying causal effects in the SNMM, assuming the time-varying moderators of interest are also the only time-varying confounders. In this manuscript, we show how the proposed, combined RR+IPTW strategy identifies the moderated time-varying causal effects by S_t in the presence of an additional, auxiliary, larger set of known, measured, candidate time-varying confounders X_t. Following Robins and colleagues [25–27], such an estimator is particularly attractive in observational study settings in which the dimensionality of the auxiliary data X_t (used to control for time-varying confounding) is much larger than that of the candidate moderators S_t. Or, even when the dimensionality of X_t is not much larger, it is useful in settings in which the measures in X_t are too costly to consider as tailoring variables for the embedded regimes (or treatment sequences) in actual clinical practice. In these cases, the researchers are interested in using X_t to adjust for confounding, but they are not interested in it scientifically (i.e., as a moderator).

In Section 2, we define the moderated time-varying causal effects more formally using the SNMM [17]. In Section 3, we consider parametric models for the SNMM. In Section 4, we present the RR+IPTW estimator of the SNMM and discuss implementation issues. In Section 5, we carry out a small simulation study to illustrate the issues and to examine asymptotic versus bootstrap standard error (BOOT) estimates for the IPTW regression-with-residuals (IPTW+RR) estimator. In Section 6, we illustrate the methods by examining the moderated effects of additional adolescent substance abuse treatment. Section 7 offers a discussion, including limitations of the IPTW+RR estimator and directions for future work.

2. A model for time-varying causal effect moderation

2.1. Notation

Suppose there are K time intervals under study. Treatment at each time interval t is denoted by a_t $(t = 1, \ldots, K)$; a_t is not a random variable. For shorthand, denote the time-varying treatment history up to interval t by $\bar{a}_t = (a_1, \ldots, a_t)$, $t = 1, \ldots, K$. For simplicity, we consider binary time-varying treatments a_t, where $a_t = 1$ denotes treatment receipt and $a_t = 0$ denotes no treatment receipt in

time interval t. Let $\mathcal{A}_K$ be the countable collection of all possible treatment vectors (e.g., for $K = 2$, $\mathcal{A}_2 = \{(0,0),(0,1),(1,0),(1,1)\}$, whereas for $K = 3$, $\mathcal{A}_3$ is the set of $2^3 = 8$ triplets of 0 or 1).

We use the potential outcomes framework [15, 28, 29] to define the causal parameters of interest in Section 2.2. For each fixed value of the treatment vector, $\bar{a}_K$, we conceptualize potential, candidate time-varying moderators $\{S_1(a_1), ..., S_{K-1}(\bar{a}_{K-1})\}$ and a potential final response $Y(\bar{a}_K)$ for each individual in the study. Thus, $S_t(\bar{a}_t)$ is the vector of candidate time-varying moderators at the beginning of the tth interval had the individual followed the treatment pattern $\bar{a}_{t-1}$ through the end of the $t-1$ interval; similarly, $Y(\bar{a}_K)$ is the value of the response at the end of study had the individual followed the treatment vector $\bar{a}_K$. Baseline moderators (pre-$\bar{a}_K$) are denoted by the vector S_0. For shorthand, let $\bar{S}_t(\bar{a}_t) = \{S_0, S_1(a_1), \ldots, S_t(\bar{a}_t)\}$, the history of candidate moderators up to the start of the tth time interval. By indexing $S_t(\bar{a}_t)$ by treatment $\bar{a}_t$, we acknowledge the potential for the moderators to be impacted by prior treatment.

In our example in Section 6, $K = 3$: a_1 denotes substance use treatment in months 1–3, a_2 denotes substance use treatment in months 4–6, a_3 denotes substance use treatment in months 7–9, the vector S_0 includes frequency of substance use prior to treatment intake and other demographic characteristics such as age, $S_1(a_1)$ is frequency of substance use during months 1–3, $S_2(a_1,a_2)$ is frequency of substance use in months 4–6, and $Y(\bar{a}_3)$, our outcome of interest, is an end-of-study measure of substance use frequency during months 10–12. For simplicity in defining the causal effects of interest, presenting the RR+IPTW estimator, and giving intuition about the method, henceforth, we focus on $K = 2$. Thus, we work with the following: in temporal order, $\{S_0, a_1, S_1(a_1), a_2, Y(a_1,a_2)\} = \{\bar{S}_1(a_1), \bar{a}_2, Y(\bar{a}_2)\}$. We return to $K = 3$ in Section 6.

2.2. *Moderated time-varying causal effects*

This section introduces time-varying effect moderation notationally using two functions (μ_1 and μ_2), which are defined using the potential outcomes introduced previously. We assume a continuous response $Y(a_1,a_2)$. We focus on modeling the mean of the response $Y(\bar{a}_2)$ as a function of $\bar{a}_2$ and $\bar{S}_1(a_1)$.

The first causal effect function of interest is at $t = 1$. It is defined as

$$\mu_1(s_0,a_1) = E(Y(a_1,0) - Y(0,0) \mid S_0 = s_0) = a_1 \times E(Y(1,0) - Y(0,0) \mid S_0 = s_0). \tag{2}$$

This function defines the average causal effects of $(a_1,0)$ versus $(0,0)$ on the outcome conditional on S_0. In the context of our motivating example, $\mu_1(s_0,1)$ represents the causal effect of receiving only initial treatment $(1,0)$ versus not receiving treatment $(0,0)$ as a function of the frequency of use prior to (S_0) treatment. $\mu_1(s_0,1)$ is a comparison of substance use frequency at the end of the study had all individuals with a fixed value of $S_0 = s_0$ received an initial dose/duration of treatment versus had they not received any treatment at all.

The second causal effect function of interest is at $t = 2$. It is defined as

$$\begin{aligned}\mu_2(s_0,a_1,s_1,a_2) = \mu_2(\bar{s}_1,\bar{a}_2) &= E(Y(a_1,a_2) - Y(a_1,0) \mid S_0 = s_0, S_1(a_1) = s_1) \\ &= a_2 \times E(Y(a_1,1) - Y(a_1,0) \mid S_0 = s_0, S_1(a_1) = s_1).\end{aligned} \tag{3}$$

This function defines the average causal effects of (a_1,a_2) versus $(a_1,0)$ on the outcome, conditional on both S_0 and $S_1(a_1)$. In the context of our motivating example, $\mu_2(s_0,a_1,s_1,1)$ represents the causal effect of receiving treatment during months 4–6 as a function of S_0, a_1, and $S_1(a_1)$. For example, $\mu_2(s_0,1,s_1,1)$ is a comparison of substance use frequency at the end of study had all individuals with a fixed value of $S_0 = s_0$ who responded to initial treatment with a fixed value of $S_1(1) = s_1$ received additional treatment versus had they not; that is, $\mu_2(s_0,1,s_1,1)$ is the effect of additional substance use treatment given the baseline frequency of use and response to prior treatment.

μ_1 and μ_2 are causal effect functions because at each time point, they represent comparisons of the potential outcomes at two (possibly) different levels of treatment: μ_1 is a contrast of the potential outcomes for the treatment at time 1 using a_1 versus 0, whereas μ_2 is a contrast of the potential outcomes for the treatment at time 2 using a_2 versus 0. They represent *moderated* causal effects because by conditioning on covariates that occur prior to each treatment, μ_1 and μ_2 describe the heterogeneity of the effects of a_1 and a_2, respectively, as they depend on these covariates.

Note that μ_1 isolates the causal effect of treatment at time 1 by setting future treatment at its inactive level, that is, $a_2 = 0$. On the other hand, μ_2, which corresponds to the effect at the last time point, is defined exclusively as a contrast in a_2 where, in general, a_1 can take on any value in its domain. Robins

[17] refers to μ_1 and μ_2 as blip-to-zero functions. It is easy to extend these definitions to define μ_1 with future treatment set to a level other than zero, such as to the active level $a_2 = 1$, to the average future treatment level, or to the optimal decision rule (regime) at $t = 2$ (see the work of Robins [27] and Henderson *et al.* [23, 24]) However, given that our substantive interest is in examining the effects of an additional duration of substance use treatment as we move through time, setting future $a_2 = 0$ is a sensible choice dictated by our scientific interests. Moreover, setting future treatment to zero provides an easier starting point to illustrate the methodology.

μ_1 averages over $S_1(a_1)$. Therefore, it constitutes the total effect of a_1 (conditional on S_0 and setting $a_2 = 0$), including effects of a_1 on the outcome that may be mediated [30, 31] by $S_1(a_1)$ or any other covariates on the pathway between a_1 and the outcome. Similarly, μ_2 averages over any covariates on the pathway between a_2 and the outcome.

2.3. The structural nested mean model

For continuous $Y(a_1, a_2)$, the SNMM is an additive, telescoping, decomposition of the conditional mean of $Y(a_1, a_2)$ given $\bar{S}_1(a_1)$ that includes the causal terms μ_1 and μ_2 in the decomposition. Specifically, for $K = 2$, the SNMM is

$$E\left(Y(a_1, a_2) \mid \bar{S}_1(a_1) = \bar{s}_1\right) = \beta_0 + \epsilon_1(s_0) + \mu_1(s_0, a_1) + \epsilon_2(\bar{s}_1, a_1) + \mu_2(\bar{s}_1, \bar{a}_2), \tag{4}$$

where the intercept $\beta_0 = E(Y(0, 0))$ is the mean outcome for the population under no treatment and the functions $\epsilon_1(s_0)$ and $\epsilon_2(\bar{s}_1, a_1)$ are defined as $\epsilon_1(s_0) = E(Y(0, 0)|S_0 = s_0) - E(Y(0, 0))$ and $\epsilon_2(\bar{s}_1(a_1)) = E(Y(a_1, 0)|\bar{S}_1(a_1) = \bar{s}_1) - E(Y(a_1, 0)|S_0 = s_0)$.

Note that $\epsilon_1(s_0)$ and $\epsilon_2(\bar{s}_1, a_1)$ are defined just so the right-hand-side of (4) equals the conditional mean on the left-hand-side. Following [17], we label the functions ϵ_1 and ϵ_2 as 'nuisance functions' to distinguish them from the causal functions of interest μ_1 and μ_2. The nuisance functions connote both causal and non-causal relationships (associations) between the candidate time-varying moderators and the response. The nuisance functions exhibit a special property that forms the basis for how we model these quantities using the RR approach in Section 4. Namely, the nuisance functions are mean-zero functions conditional on the past; that is,

$$E(\epsilon_1(S_0)) = 0, \text{ and} \tag{5}$$

$$E(\epsilon_2(\bar{S}_1(a_1))|S_0) = 0, \tag{6}$$

where the first expectation is over the random variable(s) S_0 and the second expectation is over the random variable(s) $S_1(a_1)$ conditional on S_0. Following [2, 23, 24], these properties form the basis for how we model the conditional mean $E(Y(a_1, a_2) \mid \bar{S}_1(a_1))$ using the RR approach later.

3. Linear parametric models for the SNMM

In this section, we consider parametric linear models for the SNMM, which we will later estimate using the proposed RR+IPTW approach. We begin by defining the linear models for the causal functions, then introduce linear models for the nuisance functions, and end by combining the two sets of linear models to create a model for the SNMM.

3.1. Linear models for the causal functions

We consider parametric linear models for the μ_t's of the form:

$$\mu_t(\bar{s}_{t-1}, \bar{a}_t; \beta_t) = a_t\left(H_{t-1}\beta_t\right) \tag{7}$$

where β_t is an unknown q_t-dimensional column vector of parameters and H_{t-1} is a corresponding row vector that is a function of $(\bar{S}_{t-1}(\bar{a}_{t-1}), \bar{a}_{t-1}) = (\bar{s}_{t-1}, \bar{a}_{t-1})$. H_{t-1} stands for *H*istory up to and including time $t - 1$. This functional form for the causal functions is an extension of the standard treatment-moderator interaction framework [8] (i.e., covariate-by-treatment product terms) to the time-varying setting. For example, for $t = 2$, let $H_1 = (1, s_1, a_1)$ and $\beta_2 = (\beta_{2,0}, \beta_{2,1}, \beta_{2,2})^T$ (where v^T means transpose of v) so that

$$\mu_2(\bar{s}_1, \bar{a}_2; \beta_2) = a_2(\beta_{2,0} + \beta_{2,1}s_1 + \beta_{2,2}a_1) = \beta_{2,0}a_2 + \beta_{2,1}s_1a_2 + \beta_{2,2}a_1a_2. \tag{8}$$

In this example, the effects of additional substance abuse treatment depend on the previous treatment a_1 (according to $\beta_{2,2}$) and also vary linearly in $S_1(a_1)$ (with slope equal to $\beta_{2,1}$); here, $S_1(a_1)$ is a moderator if $\beta_{2,1} \neq 0$.

3.2. *Linear models for the nuisance functions*

We also consider parametric linear models for the ϵ_t's. We consider models such as the following:

$$\epsilon_t(\bar{s}_{t-1}, \bar{a}_{t-1}; \eta_t, \gamma_t) = \eta_t \delta_t(\bar{s}_{t-1}, \bar{a}_{t-1}; \gamma_t) \tag{9}$$

where η_t is an unknown scalar parameter and the unknown 'residual' δ_t is equal to $s_{t-1}(\bar{a}_{t-1}) - m_t(\bar{s}_{t-2}, \bar{a}_{t-1}; \gamma_t)$ where $m_t(\bar{s}_{t-2}, \bar{a}_{t-1}; \gamma_t) = g_t(F_t\gamma_t)$ is a generalized linear model (GLM [32]) for the conditional expectation $E(S_{t-1}(\bar{a}_{t-1}) \mid \bar{S}_{t-2}(\bar{a}_{t-2}) = \bar{s}_{t-2})$, with link function $g_t()$, unknown j_t-dimensional column vector of parameters γ_t, and F_t is a corresponding row vector that is a function of $\bar{S}_{t-2}(\bar{a}_{t-2}) = \bar{s}_{t-2}$. For instance, for binary $S_{t-1}(\bar{a}_{t-1})$, $g_t()$ can be the 'inverse logit' transform, or for continuous $S_{t-1}(\bar{a}_{t-1})$, $g_t()$ would be the identity function.

Consistent with properties (5) and (6), $E(\epsilon_t(\bar{S}_{t-1}(\bar{a}_{t-1}); \eta_t, \gamma_t) \mid \bar{S}_{t-2}(\bar{a}_{t-2})) = 0$ because the residuals δ_t average to zero conditional on $(\bar{S}_{t-2}(\bar{a}_{t-2}))$. (Note that this expectation is over the conditional distribution $[S_{t-1}(\bar{a}_{t-1}) \mid \bar{S}_{t-2}(\bar{a}_{t-2})]$.) Indeed, this is the motivation for calling the estimator 'regression with residuals'.

As an example, suppose $S_1(a_1)$ is a continuous measure (e.g., frequency of substance use in months 1–3 in our motivating example). Consistent with (9), an example model for ϵ_2 is

$$\epsilon_2(s_0, a_1, s_1; \eta_2, \gamma_2) = \eta_2 \left(s_1 - m_1(s_0, a_1; \gamma_2)\right), \quad \text{where} \tag{10}$$

$$m_1(s_0, a_1; \gamma_2) = \gamma_{2,0} + \gamma_{2,1}s_0 + \gamma_{2,2}a_1 + \gamma_{2,3}s_0a_1 \tag{11}$$

is a linear model for the conditional mean $E(S_1(a_1) \mid S_0 = s_0)$. In this example, note that $F_2 = (1, s_0, a_1, s_0a_1)$, $\gamma_2 = (\gamma_{2,0}, \gamma_{2,1}, \gamma_{2,2}, \gamma_{2,3})$, and $g_2()$ is the identity function.

The parametric form in (9) is for univariate $S_t(\bar{a}_t)$. For multivariate $S_t(\bar{a}_t)$ (say, $S_t(\bar{a}_t) = (S_{tk}(\bar{a}_t) : k = 1, \ldots, r_t)$, a vector of r_t candidate moderators at time t), we propose postulating models such as $\epsilon_{tk} = \eta_{tk}\delta_{tk}$, one for each S_{tk} as in (9), and then summing these models together to create an overall parametric model for tth time-point nuisance function: $\epsilon_t = \sum_k^{r_t} \epsilon_{tk}$ (see the appendix in [2]). Note that in the multivariate case, δ_t is an r_t-dimensional row vector and η_t the appropriate column vector whereas in the univariate case (one moderator per time point), η_t is scalar, so that $r_t = 1$ for all t.

3.3. *Combining the causal and nuisance parametric linear models*

Combining the linear parametric models for the causal (μ_t) and nuisance (ϵ_t) functions, we arrive at a linear parametric SNMM, denoted m_Y. For instance, assuming the candidate time-varying moderator $S_t(\bar{a}_t)$ is univariate continuous and using the previous example models, plus letting $H_0 = (1, s_0)$ and $\beta_1 = (\beta_{1,0}, \beta_{1,1})^T$ make up the model for μ_1 and letting $F_1 = (1)$ and $\gamma_1 = (\gamma_{1,0})$ make up the 'model' $m_1 = \gamma_1$ for $E(S_0)$, imply the following example linear SNMM:

$$m_Y(\bar{s}_1, \bar{a}_2; \beta, \eta, \gamma) = \beta_0 + \delta_1\eta_1 + \beta_{1,0}a_1 + \beta_{1,1}s_0a_1 + \delta_2\eta_2 + \beta_{2,0}a_2 + \beta_{2,1}s_1a_2 + \beta_{2,2}a_1a_2, \tag{12}$$

where $\beta = (\beta_0, \beta_1^T, \beta_2^T)^T$, $\eta = (\eta_1, \eta_2)^T$, $\gamma = (\gamma_1, \gamma_2^T)^T$, $\delta_1 = s_0 - m_1$, and, $\delta_2 = s_1 - m_2$.

It is noteworthy that this linear SNMM is very similar to the traditional regression analysis approach, Equation (1), except in two important ways: First, in Equation (12), the 'main associational effects' of the candidate time-varying moderators are conditional-mean centered. That is, the S_t's in Equation (1) are replaced by δ_t's in Equation (12). The intuition here is that by 'residualizing' the S_t's using a conditional model for S_t given the past—in particular, residualizing S_1—we avoid the potential problems described in the Introduction related to naively conditioning on candidate moderators impacted by prior treatment. Second, Equation (12) focuses solely on relating the outcome Y with time-varying treatments and candidate moderators. It does not adjust for candidate time-varying confounders X_t because they are not of particular scientific interest. This allows for a more parsimonious model which focuses on the science. The next two subsections describe how to estimate the parameters of the SNMM all the while adjusting for the candidate time-varying confounders X_t using a weighted least squares regression approach.

More generally, letting $D_\gamma = (1, \delta_1, a_1 H_0, \delta_2, a_2 H_1)$ denote the SNMM 'design' vector and letting $\theta = (\beta_0, \eta_1^T, \beta_1^T, \eta_2^T, \beta_2^T)^T$ denote the $(1 + \sum r_t + \sum q_t)$-dimensional vector of unknown SNMM parameters, we can write linear parametric models for the SNMM more succinctly as $m_Y = D_\gamma \theta$. We index the design matrix D by γ as a reminder that it is a function of unknown parameters γ used in the residuals, the δ_t's, which make up the models for the nuisance functions, the ϵ_t's.

Apart from the special case of fully saturated SNMMs ([1]), which by definition cannot be misspecified, the previous parametric models constitute modeling assumptions. This is the first of four assumptions made in this methodology. The other three are described later.

4. Estimation

Now, we turn to estimation of the SNMM. This section describes the observed data and additional assumptions. Further, we discuss the proposed RR+IPTW, and we provide steps for implementing it. Finally, we discuss approaches for obtaining standard errors.

4.1. Observed data and assumptions

In this subsection, we describe the observed data that are used to estimate the SNMM and additional assumptions. The observed data in temporal order is $O = \{V_0, A_1, V_1, A_2, \ldots, V_{K-1}, A_K, Y\}$, where $V_t = \{X_t, S_t\}$ includes candidate time-varying moderators S_t and auxiliary time-varying variables X_t used to control for confounding (which we define later). A_t is the observed value of treatment; unlike a_t, A_t is a random variable. We envision estimation of the causal functions in the SNMM in settings in which the dimensionality of X_t is large as compared with S_t. As before, $\bar{A}_t = (A_1, \ldots, A_t)$ for $t = 1, \ldots, K$, and similarly, $\bar{S}_t = (S_1, \ldots, S_t)$ for $t = 1, \ldots, K-1$.

There are three assumptions in addition to the parametric modeling assumptions described in Section 3. We invoke the Consistency Assumption [17] for the S_t's and Y to establish the link between the potential outcomes and O. The Consistency Assumption states that $Y = Y(\bar{A}_K)$, where the $Y(\bar{A}_K)$ denotes the potential outcome indexed by values of $\bar{a}_K$ equal to $\bar{A}_K$. This assumption says that the observed outcome Y for an individual that follows the trajectory of observed treatment values A_K agrees with the potential outcome indexed by the same trajectory of values. Similarly, we assume consistency for the candidate time-varying moderators S_K.

To identify the μ_t's using the observed data, we assume the No Unmeasured or Unknown Direct Confounders Assumption [17]: For every t $(t = 1, 2, \ldots, K)$, A_t is independent of the set $\{Y(\bar{a}_K) : \bar{a}_K \in \mathcal{A}_K\}$ conditional on $(\bar{V}_{t-1}, \bar{A}_{t-1})$. In a SMART [21], this assumption is satisfied by design, whereas in observational studies, it is not possible to know whether this assumption is satisfied. In observational studies, this assumption informally states (for every t) that aside from the history of candidate time-varying moderators, history of treatment, and auxiliary time-varying covariates measured up to time t, there exist no other pre-A_t variables (measured or unmeasured, known or unknown) that are directly related to both A_t and the potential outcomes.

The following Positivity Assumption is also made: For all V_t and every t,

$$0 < Pr\left(A_t = 1 \mid \bar{V}_{t-1}, \bar{A}_{t-1}\right) < 1. \tag{13}$$

This assumption states that every individual could potentially be assigned to any of the treatments (at each time t). This assumption ensures we do not have true weights (defined later) with infinite values.

4.2. Inverse-probability-of-treatment-weighted regression with residuals

We now turn to the proposed RR+IPTW estimator for the SNMM. The estimator is the solution $\theta = \hat{\theta}$ to the following set of $d = 1 + \sum r_t + \sum q_t$ weighted estimation equations:

$$0 = \mathbb{P}_n \ \psi_\theta(O; \theta, \gamma, \alpha, \pi) = \mathbb{P}_n \ W(\alpha, \pi)(Y - D_\gamma \theta) D_\gamma^T, \tag{14}$$

where n is the number of individuals in the data set and $\mathbb{P}_n v$ is shorthand for the average $1/n \sum_i^n v_i$. (θ is $d \times 1$ dimensional; D_γ is $1 \times d$ dimensional.) The inverse-probability-of-treatment weights $W(\alpha, \pi)$

are defined as

$$W\left(\bar{V}_{K-1}, \bar{A}_K; \alpha, \pi\right) = \prod_{t=1}^{K} W_t\left(\bar{V}_{t-1}, \bar{A}_t; \alpha_t, \pi_t\right), \tag{15}$$

where

$$W_t\left(\bar{V}_{t-1}, \bar{A}_t; \alpha_t, \pi_t\right) = A_t \frac{p_t^{num}(\pi)}{p_t^{den}(\alpha)} + (1 - A_t) \frac{(1 - p_t^{num}(\pi))}{(1 - p_t^{den}(\alpha))}; \tag{16}$$

the numerator propensity score $p_t^{num}(\pi)$ is a model (say, a logistic regression) for $Pr(A_t = 1 \mid \bar{S}_{t-1}, \bar{A}_{t-1})$, and the denominator propensity score $p_t^{den}(\alpha)$ is a model for $Pr(A_t = 1 \mid \bar{V}_{t-1}, \bar{A}_{t-1})$.

Estimator (14) is nothing more than a weighted least squares regression estimator: Y is regressed on D_γ, in a regression fit weighted by W. The regression focuses on obtaining estimates of the parameters of the SNMM (including the effect estimates β), whereas the weights focus on reducing or eliminating time-varying confounding bias. Importantly, note that the auxiliary candidate time-varying confounders X_t are not a part of the linear SNMM (D_γ) but are adjusted for via the weights (W). In a slightly different context, Murphy *et al.* [25], van der Laan *et al.* [26] (Section 6.5), and Robins [27] (see Section 7.3, pp. 78–80) provide the theory that shows that under the assumptions listed previously and known $W(\alpha, \pi)$, the previous weighting approach can be used to identify the β parameters.

At each time point, the purpose of W_t is to re-weight the data such that confounding due to $\bar{V}_{t-1}$ is eliminated (under the assumption of no unknown or unmeasured time-varying confounders, and hopefully greatly reduced even if those assumptions do not hold). The denominator in W_t adjusts for imbalances due to $\bar{V}_{t-1}$ in the types of individuals who are treated ($A_t = 1$) versus those who are untreated ($A_t = 0$). The denominator in W_t accomplishes this by up-weighting individuals who are unlikely to receive the treatment they received given $(\bar{A}_{t-1}, \bar{V}_{t-1})$ and by down-weighting individuals who are likely to have received the treatment they received given $(\bar{A}_{t-1}, \bar{V}_{t-1})$.

The numerator's role in W_t is not to adjust for confounding (the denominator does this on its own); p_t^{num} is not required for eliminating or reducing bias due to time-varying confounding. Indeed, p_t^{num} can be set to any function $p(\bar{S}_{t-1}, \bar{A}_{t-1}) \in (0, 1)$ including the constant function. To ensure that (14) is an asymptotically unbiased estimating equation, p_t^{num} should not depend on the variables in $\bar{X}_{t-1}$ or any variables after time $t - 1$.

Rather, the reason for using numerator probabilities is that it potentially improves the statistical efficiency in the estimates $\hat{\theta}$ by making the weights W_t less variable because $0 < p_t^{num}(\pi) < 1$. Robins *et al.* [12] have labeled weights with this property as 'stabilized weights'. The form of p_t^{num} differs from the form of the numerator used in weights to estimate MSMs. Specifically, because MSMs are conditional only on baseline covariates, the numerator of the weights is only a function of baseline covariates and prior treatment, whereas, because SNMMs are conditional on time-varying moderators $\bar{S}_{t-1}$, the numerator of W_t may be a function of both $\bar{S}_{t-1}$ and prior treatment. The weights used here are also discussed in Petersen *et al.* [3] and used by Rosthøj *et al.* [33] to estimate history-adjusted MSMs.

The use of p_t^{num} is also intuitive. For instance, the numerator is used to project the $1/p^{den}$-weighted sample back to the space of conditional distributions given $\bar{S}_{t-1}$. Given that we are interested in and we explicitly model the effect of A_t on Y given $\bar{S}_{t-1}$, projecting the sample back to 'within observed levels of $\bar{S}_{t-1}$' is sensible. Another intuitive way to think about this projection in the context of SMARTs—which are used to obtain high-quality randomized data specifically for the purpose of examining time-varying moderators—is that the numerator projects the sample back to the SMART design that is 'closest to' or 'implied by' the observational data.

4.3. Implementation steps: IPTW regression with residuals

In practice, additional steps are needed prior to solving Equation (14) because both W and γ are unknown. This suggests a three-step estimation procedure, where estimates of the W and γ are obtained first, prior to carrying out the weighted least squares regression. This subsection describes steps to implement estimator (14) by first obtaining estimates of the weights W and γ and then plugging these into $\psi(O; \theta, \hat{\gamma}, \hat{\alpha}, \hat{\pi})$ prior to solving for θ.

Step 1: Estimate the weights. As discussed previously, for each t, the numerator propensity score is a function of $(\bar{S}_{t-1}, \bar{A}_{t-1})$. The denominator propensity score, which is used to balance treated

($A_t = 1$) and untreated ($A_t = 0$) groups (i.e., used to reduce or eliminate confounding), is a function of $(\bar{V}_{t-1}, \bar{A}_{t-1})$.

1a: Estimate the numerator model (obtain $\hat{\pi}$). For each t, estimate p_t^{num} using a logistic regression model for $Pr(A_t = 1 \mid \bar{S}_{t-1}, \bar{A}_{t-1})$ with unknown parameters π_t. Calculate and save the $\hat{p}_t^{num}(\hat{\pi}_t)$'s.

1b: Estimate the denominator model (obtain $\hat{\alpha}$). For each t, estimate p_t^{den} using a logistic regression model for $Pr(A_t = 1 \mid \bar{V}_{t-1}, \bar{A}_{t-1})$ with unknown parameters α_t. Calculate and save the $\hat{p}_t^{den}(\hat{\alpha}_t)$'s.

1c: Calculate final weights (obtain $\hat{W}(\hat{\alpha}, \hat{\pi})$).

$$\hat{W}_t := \hat{W}_t\left(\bar{V}_{t-1}, \bar{A}_t; \hat{\alpha}_t, \hat{\pi}_t\right) = A_t \frac{\hat{p}_t^{num}(\hat{\pi})}{\hat{p}_t^{den}(\hat{\alpha})} + (1 - A_t) \frac{(1 - \hat{p}_t^{num}(\hat{\pi}))}{(1 - \hat{p}_t^{den}(\hat{\alpha}))}$$

at each time t, and then calculate the final combined weight $\hat{W}(\hat{\alpha}, \hat{\pi}) = \prod_{t=1}^{K} \hat{W}_t$.

Step 2: Residualize candidate moderators (obtain $\hat{\gamma}$). For each t, specify and estimate the appropriate weighted GLM, $g_t(F_t\gamma_t)$, for $E(S_{t-1} \mid \bar{S}_{t-2}, \bar{A}_{t-1})$ with design matrix F_t and unknown parameters γ_t. For the trivial $t = 0$ models for $E(S_0)$, the GLM is unweighted (or, equivalently, weighted with known weight $W_0 = 1$); for $t \geqslant 1$, use the estimated weights $\prod_{j=1}^{t} \hat{W}_j$. (For multivariate $S_{t-1} = (S_{t-1,1}, \ldots, S_{t-1,k}, \ldots, S_{t-1,r_t})$, specify and estimate weighted GLMs for each of the $S_{t-1,k}$'s given the past.) From each fitted GLM, calculate the estimated residual $\hat{\delta}_t(\hat{\gamma}_t)$. In Step 3, the $\hat{\delta}_t$'s will be used as covariates in the model for the SNMM.

Step 3: Estimate the SNMM using RR+IPTW (obtain $\widehat{\theta}$). Specify a model $D_{\hat{\gamma}}$ for the SNMM. Note that the models for the nuisance functions (e.g., main effects of the candidate time-varying moderators) in $D_{\hat{\gamma}}$ use the residuals $\hat{\delta}_t$'s from Step 2. To obtain the estimate $\hat{\theta}$, employ a weighted least squares regression of Y on $D_{\hat{\gamma}}$ with weights $\hat{W}$.

4.4. Standard errors

The nominal standard errors (i.e., those reported from standard regression procedures using over-the-counter statistical software packages such as SAS) for the weighted least squares regression estimates of θ (Step 3) are inappropriate because they assume that the residuals $\delta_t(\gamma_t)$ and the weights $W_t(\alpha_t, \pi_t)$ are known; that is, nominal standard errors do not take into account estimation of (γ, α, π) in the final estimates $\hat{\theta}(\hat{\gamma}, \hat{\alpha}, \hat{\pi})$ of the SNMM. Consequently, the use of nominal standard errors may result in p-values and confidence intervals for $\widehat{\theta}$ that are smaller than appropriate. Asymptotic standard errors (ASEs) obtained using the delta method (e.g., Taylor series arguments, see Appendix B), which take into account sampling error in the estimation of $(\hat{\gamma}, \hat{\alpha}, \hat{\pi})$ are used instead. However, because not all investigators have the resources to computer program the ASEs, we also compare results with bootstrap (BOOT) estimates for the standard error of $\hat{\theta}$, which are easier to calculate using most over-the-counter statistical software packages. To obtain the BOOT, we implement the RR+IPTW estimator on 500 data sets of size n sampled at random (with replacement) from the original data set of size n and take the standard deviation (SD) of the 500 estimates.

5. Simulations

We conducted two small simulation experiments: (1) to illustrate and compare IPTW+RR, RR, and the traditional regression estimators of the causal parameters in an SNMM for $E(Y(a_1, a_2) \mid S_0, S_1(a_1))$ and (2) to compare small and large sample properties of asymptotic versus bootstrap estimators of the standard errors for the IPTW+RR approach. We generated the data $\{U, V_0, A_1, V_1, A_2, Y\} = \{U, (X_0, S_0), A_1, (X_1, S_1), A_2, Y\}$ to mimic the adolescent substance use data with $K = 2$. We did this by ensuring that the marginal distributions (e.g., proportions, means, and SDs) of the generated data were similar to the adolescent substance use data.[‡] We give the details concerning the data generating models in Appendix A. Key features of the data generating model include the following:

[‡]*The conditional distributions we specified, however, differ from the results of the adolescent substance use data analysis described in the next section.*

(1) The generating model implies a linear SNMM for $E(Y(a_1, a_2) \mid S_0, S_1(a_1))$.
(2) S_t is a time-varying moderator of the effect of A_t on Y.
(3) Both S_1 and X_1 are affected by A_1, which, in turn, are associated with Y. Intuitively, S_1 and X_1 may be mediators of the effect of A_1 on Y.
(4) We generated a baseline variable U that affects both S_1 and Y. (U is not on the causal pathway between A_1 and Y, and it is neither a time-varying moderator nor a confounder.) This variable creates a spurious (non-causal) correlation between A_1 and Y when adjusting naively for S_1.
(5) In the first simulation, we vary whether there exists time-varying confounding by including or removing the association between $\bar{V}_{t-1}$ and A_t.

5.1. *IPTW+RR versus RR versus traditional regression*

The first simulation illustrates, in the context of a simple example, how the IPTW+RR estimator compares with RR and two versions of the traditional regression approach in terms of bias. We consider two versions of the data generating model in this experiment: with and without time-varying confounding. Both IPTW+RR and RR use the correct functional form for $E(Y(a_1, a_2) \mid S_0, S_1(a_1))$. The first traditional regression estimator (TRAD1) fits the same functional form as fitted with the IPTW+RR and RR estimators except that the δ_t are replaced with S_t. The second traditional regression estimator (TRAD2) in addition adjusts naively for X_t. Under each condition (with or without time-varying confounding), we generated 1000 data sets of size $n = 2870$. We give the details concerning the data analysis models in Appendix A.

Table I shows the results of the first simulation experiment. We report summaries of the bias and SD for the (β_1, β_2) parameters indexing (μ_1, μ_2) in the generative model, as well as for $\beta_0 =$ the marginal mean outcome under no treatment. We make the following general observations:

- When there is time-varying confounding, RR and TRAD1 were generally biased, whereas the IPTW+RR was unbiased. This is because RR and TRAD1 do not adjust for X_t in any way.
- Whether there is time-varying confounding, TRAD2 was biased for (β_0, β_1) but unbiased for β_2. TRAD2 is unbiased for β_2 because it adjusts for all measures associated with A_2 and Y. However, even though it also adjusts for all measures associated with A_1 and Y, it is biased for (β_0, β_1) because it adjusts for (S_1, X_1) naively. The bias we see in TRAD2 is due to both cutting off the

Table I. Results from a simulation experiment to illustrate and compare the bias and standard deviation between IPTW+RR, RR, and traditional regression (TRAD) approach.

		Bias (100 × SD)			
Generative model	Effect	IPTW+RR	RR	TRAD1	TRAD2
Time-varying confounding	β_0	0.0 (0.7)	**2.4** (0.7)	**31.0** (2.8)	**41.5** (2.5)
	$\beta_{1,0}$	0.0 (3.1)	**1.9** (3.0)	**5.1** (2.8)	**7.4** (2.5)
	$\beta_{1,1}$	0.1 (5.3)	0.1 (5.0)	**3.5** (4.7)	**2.3** (4.1)
	$\beta_{2,0}$	0.0 (1.6)	**1.4** (1.5)	**1.4** (1.5)	0.0 (1.4)
	$\beta_{2,1}$	0.1 (3.7)	0.1 (3.5)	0.1 (3.5)	0.0 (3.2)
No time-varying confounding	β_0	0.0 (0.3)	0.0 (0.4)	**29.9** (2.0)	**41.4** (1.9)
	$\beta_{1,0}$	0.0 (2.4)	0.0 (2.4)	**3.6** (2.3)	**7.4** (2.1)
	$\beta_{1,1}$	0.0 (3.9)	0.0 (3.9)	**3.5** (3.7)	**2.2** (3.2)
	$\beta_{2,0}$	0.1 (1.4)	0.1 (1.4)	0.1 (1.4)	0.0 (1.3)
	$\beta_{2,1}$	0.1 (2.9)	0.1 (2.9)	0.1 (2.9)	0.0 (2.5)

Appendix A describes the data generating and analysis models. IPTW+RR refers to the inverse-probability-of-treatment weighted regression with residuals estimator. RR refers to the regression with residuals estimator. TRAD1 refers to the traditional regression estimator that adjusts naively for S_t only (the candidate time-varying moderator). TRAD2 is the traditional regression estimator that adjusts naively for both S_t and X_t (a time-varying confounder). We used 1000 data sets of size $n = 2870$ in all simulation conditions. We define bias as 100 × |TRUE - EST|. SD is the empirical standard deviation of the 1000 estimates. Conditions for which BIAS > SD/2 (i.e., greater than a 'moderate' amount of bias using Cohen's [34] rules of thumb) are shown in bold.

effect of A_1 on Y via (S_1, X_1) and due to the non-causal association between A_1 and Y via U as a result of adjusting naively for S_1.

- When there is no time-varying confounding, both IPTW+RR and RR are unbiased. In this case, it is not necessary to use weighting: RR by itself is sufficient because the goal of IPTW is to adjust for time-varying confounders.
- When there is no time-varying confounding, TRAD1 and TRAD2 continue to be biased for the parameters in $(\beta_0, \beta_{1,0})$. This bias occurs because of the problems with traditional regression noted in the Introduction. The bias is generally greater in TRAD2 than with TRAD1 in these simulations (although this will not in general be the case).
- When there is no time-varying confounding, all four estimators are unbiased for $\beta_{2,0}$.
- Whether there was time-varying confounding, estimates of the parameters in μ_2 (and therefore bias) were identical for RR and TRAD1. This is as expected: Given the same model for μ_2, RR and TRAD1 will always yield identical estimates for μ_2. This is because the estimating equations for the parameters in μ_2 are identical for RR and TRAD; see [2] for details.

5.2. *Asymptotic versus bootstrap standard errors*

The second experiment focuses on comparing the large and small sample properties of the bootstrap versus asymptotic estimates of the standard errors of the IPTW+RR. For this experiment, we employed the data generating model with time-varying confounding used previously, but we varied the sample size: $n = 100$ (small), 250 (medium), 500 (large), 2870 (very large). In Table II, we report the SD of the IPTW+RR estimates, mean BOOT, mean ASEs, and coverage of the 95% confidence intervals for both the bootstrap (BOOT95) and asymptotic (ASE95) standard errors over the 1000 simulated data sets. At very large samples, such as with $n = 2870$ (the sample size of our substance use data set), the bootstrap and the ASE were nearly indistinguishable in our simulation experiments. However, in small samples sizes $n = 100$, the 95% confidence interval calculated using the ASE sometimes had lower than nominal coverage (0.932 for β_0) and much lower than nominal coverage (0.919 for $\beta_{1,0}$). In general, BOOT95 had closer to nominal coverages than ASE95. On the basis of these results, we use BOOT in the data analysis in Section 6.

Table II. Results from a simulation experiment to compare the large and small sample properties of the bootstrap and asymptotic estimates of the standard errors of the IPTW+RR.

Sample size	Effect	SD	BOOT	BOOT95	ASE	ASE95
$n = 100$	β_0	0.041	0.041	0.942	0.043	0.932
	$\beta_{1,0}$	0.185	0.204	0.953	0.177	0.919
	$\beta_{2,0}$	0.087	0.088	0.952	0.087	0.945
$n = 250$	β_0	0.025	0.024	0.947	0.027	0.965
	$\beta_{1,0}$	0.113	0.111	0.942	0.109	0.933
	$\beta_{2,0}$	0.053	0.053	0.944	0.054	0.942
$n = 500$	β_0	0.017	0.017	0.949	0.019	0.968
	$\beta_{1,0}$	0.078	0.077	0.941	0.077	0.934
	$\beta_{2,0}$	0.037	0.037	0.945	0.038	0.944
$n = 2870$	β_0	0.007	0.007	0.948	0.008	0.944
	$\beta_{1,0}$	0.031	0.032	0.961	0.032	0.964
	$\beta_{2,0}$	0.016	0.016	0.946	0.016	0.954

Appendix A describes the data generating models. We used 1000 data sets in all simulation conditions and report the empirical standard deviation (SD) of estimates. BOOT95 and ASE95 refer to the coverage probabilities (over the 1000 data sets) for the 95% confidence interval constructed using either the bootstrap standard error (SE) (BOOT) or the asymptotic SE (ASE), respectively.

6. Illustrative data example

6.1. Data

We illustrate the methodology by using data ($n = 2870$ individuals) pooled from a number of adolescent treatment studies funded by the Substance Abuse and Mental Health Services Administration's (SAMHSA's) Center for Substance Abuse Treatment (CSAT) involving adolescents entering community-based substance abuse treatment programs. We collected all data by using the Global Appraisal of Individual Needs (GAIN [35]), a structured clinical interview of individual characteristics and functioning administered at baseline/intake and at the end of 3, 6, 9, and 12 months for a total of five measurement occasions. At each measurement occasion, GAIN questions ask about constructs over the past 90 days (past 3 months). Later, subscript values $t = 0, 1, 2, 3$ denotes that the measurement was taken at baseline and the end of 3, 6, and 9 months, respectively.

Time-varying treatment $A_t = 1$ ($t = 1, 2, 3$) if an individual reports receiving any substance use treatment in the past 90 days, and $A_t = 0$ otherwise: 82%, 40%, and 26% reported $A_t = 1$, respectively. The primary time-varying moderator of interest is the Substance Frequency Scale (SFS) collected at baseline ($S_{0,1}$; Mean(SD) = 0.18(0.18)), and at the end of months 3 (S_1; Mean(SD) = 0.07(0.11)) and 6 (S_2; Mean(SD) = 0.08(0.13)). Higher scores indicate increased frequency of substance use in terms of days used, days staying high most of the day, and days causing problems. In addition to $S_{0,1}$, we also consider the following as candidate baseline moderator variables in S_0: $S_{0,2}$ is age (continuous; Mean(SD) = 15.98(1.4) years), and $S_{0,3}$ denotes whether the adolescent reports being in a controlled environment in 90 days prior to baseline assessment (binary; rate = 49%). Y is the SFS collected at the end of month 12 (Mean(SD) = 0.09(0.13)).

The analysis included a large number of auxiliary, candidate confounder variables in X_t (description of measures not shown due to space, but see [36]). In total, $V_0 = (X_0, S_0)$ included 46 pre-A_1 measures, (V_0, A_1, V_1) included 86 pre-A_2 measures, and $(V_0, A_1, V_1, A_2, V_2)$ included 128 pre-A_3 measures.

Of observations, 13.4% (across all individuals, measures, and time points) was missing. We generated ten data sets via multiple imputation to replace missing values using a sequential regression multivariate imputation algorithm [37, 38]. The imputation model was congenial with all analysis models. We calculated estimates and standard errors (SEs) reported later using rules [39, 40] for combining the results of identical analyses performed on each of the 10 imputed data sets.

6.2. Estimating the weights

We estimated $p_t^{num}(\pi_t)$ and $p_t^{den}(\alpha_t)$ (used in the weights) using logistic regression. The primary role of $p_t^{den}(\alpha_t)$ is to reduce or eliminate the imbalance between treated ($A_t = 1$) and untreated ($A_t = 0$) individuals on the basis of observed time-varying confounders up to $t - 1$. Therefore, following [41], we employed a strategy for selecting the denominator logistic regression models for $p_t^{den}(\alpha_t)$ that leads to improved balance between treated and untreated individuals at time t. At each time point t, we assessed balance before versus after IPTW (using only the denominator model) for each candidate time-varying confounder X_t. For each covariate, we measure balance as the standardized mean difference (effect size) between those who are treated versus untreated at time t. Figure 1 summarizes pictorially the balance before versus after weighting. The denominator-only weights were well behaved, with a maximum value of 37.93 (across all t); exhaustive details concerning weight diagnostics, as well as how the weights were chosen, are given elsewhere [36]. The maximum effect size in the denominator-only IPTW sample (across all t) was 0.16, reduced from as high as 1.15 in the unweighted sample. The average effect size (averaged over covariates) was reduced significantly from as high as 0.198 (unweighted at $t = 3$) to lower than 0.024 ($t = 2$). As shown in Figure 1, the IPTW sample resembles a sequentially randomized experiment (with respect to the observed data), whereas prior to weighting, there is clear potential for observed time-varying confounding.

6.3. Exploratory data analysis

Prior to estimation, we conducted an exploratory data analysis (EDA) to inform our choice of models for the SNMM. To keep the exploratory and illustrative analyses simple, we found it useful to dichotomize age at the median (also the mean) of 16 years old: Therefore, we set $S_{0,2}^* = 1$ if $S_{0,2}(\text{age}) \geqslant 16$ or

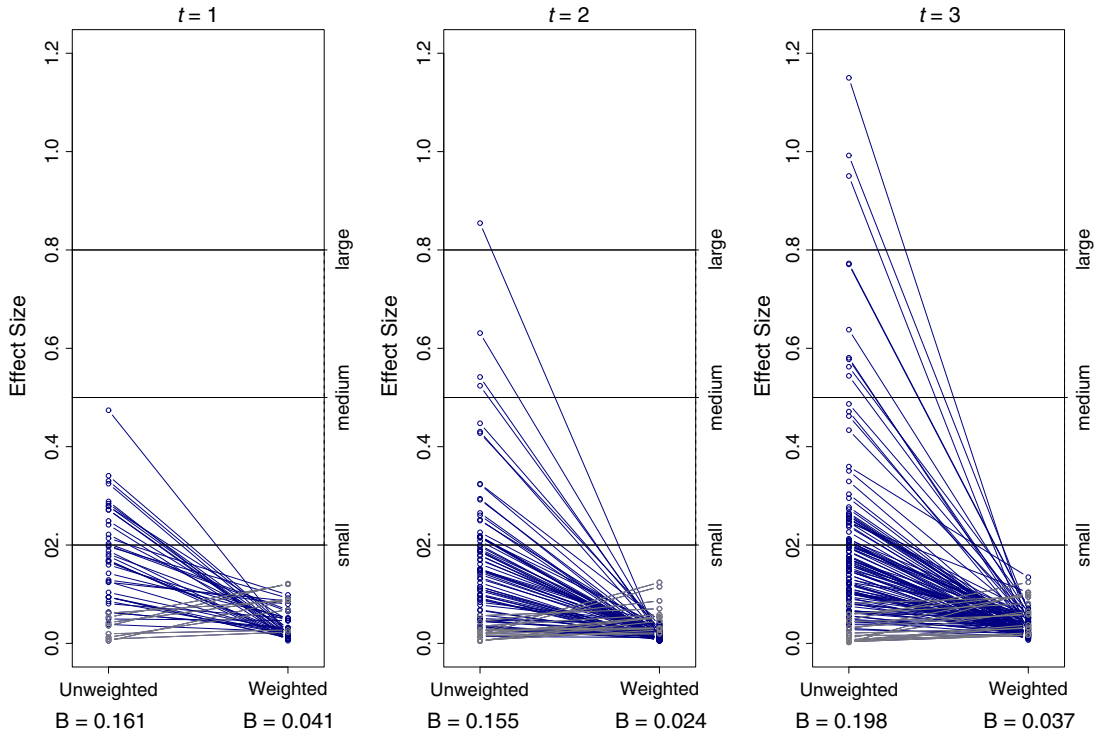

Figure 1. Balance before versus after denominator-only inverse-probability-of-treatment weighting $\left(W_t = A_t / p_t^{den} + (1 - A_t) / \left(1 - p_t^{den}\right)\right)$. Each line represents a covariate. For each covariate, balance is measured as the mean difference (weighted or unweighted, as appropriate) between treated versus untreated individuals divided by unweighted, pooled standard deviation. B is the average balance score over all covariates.

$S^*_{0,2} = 0$ if $S_{0,2} < 16$. The top, middle, and bottom four panels in Figure 2 compares data for adolescents treated with $(1, 0, 0)$ versus $(0, 0, 0)$, $(1, 1, 0)$ versus $(1, 0, 0)$, and $(1, 1, 1)$ versus $(1, 1, 0)$, respectively, to inform our choice of models for μ_1, μ_2, and μ_3. The y-axis in each panel is the outcome $Y =$ SFS in months 9–12; the x-axis for the top, middle, and bottom four panels are the time-varying candidate moderators S_0, S_1, and S_2, respectively. Each panel presents a scatter plot of Y versus S_t, with smoothing curves for each treatment trajectory for each of the four combinations of whether the adolescent had a history of controlled environment prior to intake $\times$ whether the adolescent is $\geqslant 16$ years of age. The points and the fitted smoothing splines were IPTW weighted to adjust for time-varying confounding in the EDA.[§] For μ_1, the top four panels of Figure 2 suggest that the distal effect of treatment is iatrogenic among adolescents $\geqslant 16$ years old and is more strongly iatrogenic among adolescents with higher severity at intake. For μ_2, the EDA suggests that among adolescents still using frequently at the end of 3 months despite treatment, the medial effect of treatment is beneficial to those who had been in a controlled environment prior to treatment yet iatrogenic among those who had no exposure to a controlled environment prior to treatment. For μ_3, there appears to be a beneficial proximal effect of treatment for adolescents who remain severe at the end of 6 months under treatment, regardless of age or history of controlled environment. An EDA for the non-monotonic comparisons of treated versus untreated individuals in μ_2 and μ_3 (EDA not shown) showed beneficial medial and proximal effects of treatment. For all μ_t, EDA suggested no effect of additional treatment among adolescents who are not severe at $t - 1$ regardless of age or history of controlled environment (i.e., it is difficult to see in Figure 2, but all the curves come together at $S_{t-1} = 0$).

6.4. Estimating the SNMM

On the basis of the EDA, we fitted the following SNMM to the adolescent substance use data using the IPTW+RR estimator:

[§] *For simplicity, Figure 2 presents EDA using only data from the first imputed data set; trends were similar (although magnitudes differed) for the other imputed data sets.*

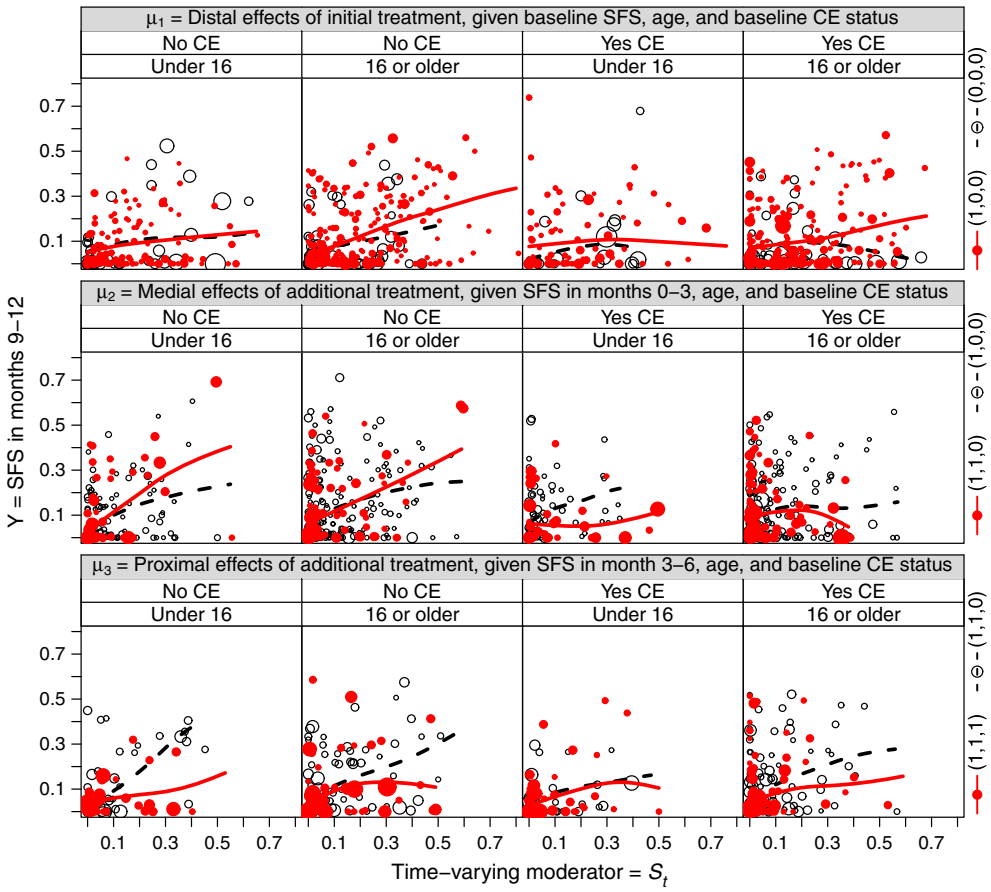

Figure 2. Exploratory data analysis of the distal, medial, and proximal effects of treatment on 12-month substance use frequency, conditional on baseline, 3-month, and 6-month substance use frequency, respectively, age (< 16 vs. $\geqslant 16$) and whether the adolescent was in a controlled environment (CE) prior to beginning any treatment. Treatment group-specific smoothing spline curves are weighted (denominator-only model); curves are plotted over the range of S_t in each treatment group. The key in the right margin indicates the treatment sequences (a_1, a_2, a_3) being compared in the panels for each row. The size of each data point is proportional to the weights.

$$
\begin{aligned}
m_Y = \beta_0 &+ \overbrace{\eta_{1,1}\delta_{1,1} + \eta_{1,2}\delta_{1,2} + \eta_{1,3}\delta_{1,3}}^{\epsilon_1} + \overbrace{a_1\left(\beta_{1,0} + \beta_{1,1}s_{0,1}s^*_{0,2}\right)}^{\mu_1} \\
&+ \overbrace{\eta_2\delta_2}^{\epsilon_2} + \overbrace{a_2a_1\left(\beta_{2,0} + \beta_{2,1}s_1s_{0,3} + \beta_{2,2}s_1(1-s_{0,3})\right) + a_2(1-a_1)\left(\beta_{2,3} + \beta_{2,4}s_1\right)}^{\mu_2} \\
&+ \overbrace{\eta_3\delta_3}^{\epsilon_3} + \overbrace{a_3a_2a_1\left(\beta_{3,0} + \beta_{3,1}s_2\right) + a_3(1-a_2a_1)\left(\beta_{3,2} + \beta_{3,3}s_2\right)}^{\mu_3}, \qquad (17)
\end{aligned}
$$

where $\delta_{1,j} = s_{0,j} - m_0(\gamma_{1,j}) = S_{0,j} - \gamma_{1,j}$ for $\forall\, j$; $\delta_2 = s_1 - m_1(s_0, a_1; \gamma_2)$, where $m_1(s_0, a_1; \gamma_2) = \gamma_{2,0}+\gamma_{2,1}s_{0,1}+\gamma_{2,2}s_{0,2}+\gamma_{2,3}s_{0,3}+\gamma_{2,4}a_1+\gamma_{2,5}a_1s_{0,1}+\gamma_{2,6}a_1s_{0,2}$; and $\delta_3 = s_2 - m_2(\bar{s}_1, \bar{a}_2; \gamma_3)$, where $m_2(\bar{s}_1, \bar{a}_2; \gamma_3) = \gamma_{3,0} + \gamma_{3,1}s_{0,1} + \gamma_{3,2}s_{0,2} + \gamma_{3,3}s_{0,3} + \gamma_{3,4}a_1 + \gamma_{3,5}s_1 + \gamma_{3,6}a_2 + \gamma_{3,7}a_2s_{0,1} + \gamma_{3,8}a_2s_1$.

Estimates for $\gamma_{1,j}$ $\forall\, j$, and for γ_2 and γ_3 are given elsewhere [36]. Table III shows IPTW+RR estimates for the parameters of the SNMM in Equation 17.

To facilitate interpretation of the fitted SNMM, Table IV describes the meaning of some linear contrasts of interest and presents their estimates. Table IV also reports effect sizes (standardized estimates, see [34]) and confidence intervals for the effect sizes (using BOOT). Except for the medial effects of additional treatment, the results of the fitted model are consistent with the EDA: First, initial treatment alone may be iatrogenic for older kids with high severity at intake (small effect size, $ES = 0.240$, p-val $= 0.09$). One conjecture for this is that adolescents who are severe and only receive initial treatment will not only fail to benefit from treatment (because of insufficient time in treatment) but may also associate with other severe adolescents during treatment and, in turn, increase use in the long term. This

Table III. Estimates of the SNMM using the adolescent substance use data set.

Term	θ	Estimate (SE) IPTW+RR	RR	TRAD†
(Intercept)	β_0	0.094**	0.091**	0.13
		(0.008)	(0.008)	(0.033)
$\delta_{1,1}$	$\eta_{1,1}$	0.124**	0.121**	0.029
		(0.020)	(0.020)	(0.020)
$\delta_{1,2}$	$\eta_{1,2}$	0.004	#	0.002
		(0.003)	(0.002)	(0.002)
$\delta_{1,3}$	$\eta_{1,3}$	#	−0.003	0.003
		(0.008)	(0.005)	(0.005)
A_1	$\beta_{1,0}$	−0.001	#	−0.002
		(0.011)	(0.009)	(0.009)
$A_1\, S_{0,1}\, S^*_{0,2}$	$\beta_{1,1}$	0.065**	0.074**	0.080**
		(0.036)	(0.023)	(0.055)
δ_2	η_2	0.268**	0.333**	0.115**
		(0.047)	(0.028)	(0.030)
$A_2\, A_1$	$\beta_{2,0}$	−0.009	−0.005	#
		(0.010)	(0.007)	(0.007)
$A_2\, A_1\, S_1\, S_{0,3}$	$\beta_{2,1}$	−0.080	− 0.127**	−0.080
		(0.083)	(0.057)	(0.055)
$A_2\, A_1\, S_1\, (1 - S_{0,3})$	$\beta_{2,2}$	0.066	−0.054	−0.019
		(0.105)	(0.057)	(0.053)
$A_2\, (1 - A_1)$	$\beta_{2,3}$	−0.033	−0.002	0.004
		(0.024)	(0.019)	(0.019)
$A_2\, (1 - A_1)\, S_1$	$\beta_{2,4}$	0.294	−0.090	−0.048
		(0.263)	(0.161)	(0.162)
δ_3	η_3	0.483**	0.439**	
		(0.047)	(0.026)	
$A_3\, A_2\, A_1$	$\beta_{3,0}$	−0.006	−0.012	
		(0.010)	(0.008)	
$A_3\, A_2\, A_1\, S_2$	$\beta_{3,1}$	− 0.338**	− 0.194**	
		(0.079)	(0.061)	
$A_3\, (1 - A_2\, A_1)$	$\beta_{3,2}$	0.021	0.002	
		(0.017)	(0.011)	
$A_3\, (1 - A_2\, A_1)\, S_2$	$\beta_{3,3}$	− 0.359**	− 0.293**	
		(0.090)	(0.064)	

SNMM, structural nested mean model; SE, standard error; IPTW+RR, inverse-probability-of-treatment weighted regression with residuals estimator; RR, regression with residuals estimator; TRAD, traditional regression estimator.

†TRAD fits the same model fitted with IPTW+RR and RR, except that the δ_t are replaced with S_t.

#Indicates that $|\text{Estimate}| < 1 \times 10^{-3}$.

**Indicates that p-val < 0.10 for the hypothesis test that the parameter is zero.

is a hypothesis for which there is some support [42]. Second, receiving treatment during months 6–9 is most beneficial in terms of frequency of use at the end of the year among adolescents who are still severe at the end of month 6 (stressing perhaps the importance of treatment retention and engagement, as shown in [43, 44]). These effects were large and were slightly stronger among those who had consistent prior treatment in the last 6 months ($ES = -1.32$, p-val < 0.01) than among those who had intermittent treatment ($ES = -1.20$, p-val < 0.01).

Table IV. IPTW+RR estimated linear combinations of particular interest and their meaning.

Term	Description of the effect	EST	ES	90% CI	
	Mean distal effect of initial 0–3 month treatment alone, $Y(1,0,0)$ versus $Y(0,0,0)$:				
$\beta_{1,0}$	Among individuals who report not using 90 days prior to intake or younger than 16 years	−0.001	−0.006	(−0.132,	0.119)
$\beta_{1,0}+\beta_{1,1}/2$	Among individuals who are older than 16 years and with baseline substance frequency use of $s_{0,1}=1/2$	0.032	0.240	(0.028,	0.452)
	Mean medial effect of additional 4–6 month treatment, $Y(1,1,0)$ versus $Y(1,0,0)$:				
$\beta_{2,0}$	Among individuals who report not using in months 0–3	−0.009	−0.068	(−0.175,	0.038)
$\beta_{2,0}+\beta_{2,1}/2$	Among individuals who have been in a controlled environment in the 90 days prior to intake and with months 0–3 substance use frequency of $s_1=1/2$	−0.049	−0.371	(−0.803,	0.062)
$\beta_{2,0}+\beta_{2,2}/2$	Among individuals who were not in a controlled environment in the 90 days prior to intake and with months 0–3 substance use frequency of $s_1=1/2$	0.024	0.179	(−0.364,	0.722)
	Mean medial effect of 4–6 month treatment alone, $Y(0,1,0)$ versus $Y(0,0,0)$:				
$\beta_{2,3}$	Among individuals who report not using in months 0–3	−0.033	−0.248	(−0.516,	0.020)
$\beta_{2,3}+\beta_{2,4}/2$	Among individuals with months 0–3 substance use frequency of $s_1=1/2$	0.114	0.862	(−0.453,	2.177)
	Mean proximal effect of additional 7–9 month treatment, $Y(1,1,1)$ versus $Y(1,1,0)$:				
$\beta_{3,0}$	Among individuals who report not using in months 4–6	−0.006	−0.042	(−0.157,	0.072)
$\beta_{3,0}+\beta_{3,1}/2$	Among individuals with months 4–6 substance use frequency of $s_2=1/2$	−0.174	−1.32	(−1.719,	−0.913)
	Mean proximal effect of 7–9 month treatment & no or inconsistent past treatment, $Y(a_1,a_2,1)$ versus $Y(a_1,a_2,0)$ with $a_1a_2\neq 1$:				
$\beta_{3,2}$	Among individuals who report not using in months 4–6	0.021	0.158	(−0.028,	0.343)
$\beta_{3,2}+\beta_{3,3}/2$	Among individuals with months 4–6 substance use frequency of $s_2=1/2$	−0.159	−1.20	(−1.629,	−0.766)

EST = estimate; ES = effect size; CI = confidence interval

In addition to IPTW+RR estimates, Table III also shows estimates for the other two estimators RR and TRAD. As expected, RR and TRAD provide identical estimates of the β_3 parameters in μ_3. Comparing these estimates to those obtained using IPTW+RR, we find that the estimates of IPTW+RR are more negative (further away from zero). We conjecture that this is because the adolescents who were the worse off (that is, those with most severity up to the end of month 6) who were more likely to obtain full treatment $\bar{A}_3=(A_1,A_2,A_3)=(1,1,1)$, were also those with more substance use Y, leading to a positive confounding bias in the estimates which the IPTW+RR helps to reduce or eliminate. Estimates of the parameters in μ_1 and μ_2 were more similar across the three estimators, except in a few cases. Estimates of $\beta_{1,1}$ were slightly smaller under IPTW+RR than they were under RR and TRAD. We conjecture that this is due to a positive spurious bias that results from the TRAD estimator: For example, initial treatment may reduce severity at the end of month 3, but if there exist factors (such as social support at home) that are associated with lower use at the end of month 3 and subsequently, there will be a spurious positive

association between initial treatment and Y in TRAD (as a result of naively conditioning on month 3 use) that is reduced or eliminated by the IPTW+RR and RR estimators ([1]). For $\beta_{2,2}$, the results of the IPTW+RR were much more consistent with the EDA and differed substantially from the RR and TRAD estimates. Finally, the estimate of $\beta_{2,4}$ was large and positive for IPTW+RR and substantially different from the small and negative (close to zero) obtained under RR and TRAD. We do not have a reasonable explanation for this difference in estimates.

7. Discussion

This manuscript presents an application of the SNMM [17] for examining time-varying causal effect moderation. As stated in the Introduction, in interventions research, understanding time-varying moderation or time-varying treatments is valuable because it can be used to shed light on conceptual models, or to generate hypotheses, about tailoring variables used to guide the timing, sequencing, and duration of treatment (or treatment components) over time. For instance, in the context of our motivating example, time-varying covariates found to be moderators of the impact of additional treatment could be used in the design of a SMART [21] for further developing an individualized sequence of decision rules to guide the duration of adolescent substance use treatment. Such decision rules are also known as dynamic treatment regimes [6, 7].

This manuscript fits within the current statistical and epidemiological literature seeking to develop, evaluate, compare, and apply various methods for estimating the effects of time-varying treatments. First, we can use the proposed IPTW+RR estimator to obtain high-quality starting values for the G-Estimator [2, 17] of the SNMM. The G-Estimator has the advantage over the IPTW+RR of being doubly robust; that is, it does not require correct specification for the nuisance functions for unbiased estimation of the causal parameters. However, as suggested in the previous work [2], in certain situations, this may come at a cost in terms of statistical efficiency. The IPTW+RR may help improve the efficiency of the G-Estimator [2] by providing high-quality guesses for estimates of the nuisance functions. Second, in the context of the SNMM, the IPTW+RR shows one way to hybridize parametric estimators and IPTW estimators (often thought of as 'semi-parametric' in the time-varying confounders). Third, the methodology presented may serve as a useful starting point for applied statisticians and quantitative clinicians or behavioral scientists seeking to understand why and how to implement SNMM. For example, analysts may first understand how to fit the SNMM using the RR+IPTW estimator (which, as we show, resembles the traditional regression approach) prior to moving to more sophisticated estimators such as the G-estimator [17]. Finally, this methodology helps to further clarify the distinction between time-varying moderators and time-varying confounders. In particular, this article describes how to use IPTW methodology as a tool that allows scientists to deal with the nuisance of time-varying confounding bias, yet reserve the linear model for examining scientific questions of interest (in this case, moderated time-varying causal effects).

This methodology also helps clarify the distinction between the SNMM and the MSM [12], which is more commonly used in the epidemiological, behavioral, and medical sciences. Indeed, by averaging over the $\bar{S}_{t-1}(\bar{a}_{t-1})$ in the SNMM, we recover a model for the MSM. To appreciate this, consider the marginal time-varying causal effect $\Delta(a_1, a_2) = E(Y(a_1, a_2) - Y(0, 0))$, and note that $E(E(\mu_2(S_0, a_1, S_1(a_1), a_2) \mid S_0)) - E(\mu_1(S_0, a_1)) = \Delta(a_1, a_2)$. Similarly, if there is no effect moderation by $\bar{S}_1(\bar{a}_1)$ in μ_2 (e.g., $\beta_{2,1} = \beta_{2,2} = 0$ in Equation (3)), for example, then $\mu_2 = \beta_{2,0}$ represents the marginal time-varying effect of additional treatment at time t, $E(Y(1, 0) - Y(0, 0))$.

As is well known in the epidemiology literature [45], effect moderation depends on the scale of measurement [46]. For this reason, the more descriptive term 'effect-measure modification' is often considered more appropriate because it emphasizes that the presence or magnitude of a moderator on one scale may change when considered on a different scale.

This manuscript considers the SNMM for time-varying effect moderation for a continuous outcome. As is common for continuous outcomes, the effects of interest were defined on the linear scale (mean differences). The SNMM has been extended for binary outcomes under a log-linear scale [47–49]. As pointed out by a reviewer, it may be possible to consider the SNMM in Equation (4) with a binary outcome $Y(a_1, a_2)$. In this case, the causal effects would be defined on the risk difference scale; for example, $\mu_1 = Pr(Y(a_1, 0) = 1 \mid S_0) - Pr(Y(0, 0) = 1 \mid S_0)$. It may be possible to apply the IPTW+RR in this setting. A key challenge with 'linear probability models' is that such models do not require probabilities to stay within (0,1), so additional diagnostics related to modeling assumptions may be needed.

We present simulation experiments in Section 5 to illustrate the methodology (vs. traditional regression, TRAD) under various scenarios. More careful simulation experiments could be conducted to quantify the biases incurred by the traditional regression estimator under other realistic scenarios, including under different assumptions about the extent of the time-varying confounding. For example, we conjecture that in scenarios where time-varying confounding bias is small to moderate (say, Cohen's [34] $d \approx 0.1$ at each time point), these small potentially inconsequential biases may amount to large cumulative ones, especially for the parameters associated with the effect at earlier time points.

The pattern of biases observed in the first simulation experiment will not always hold in practice. For example, our choice of data generative model led to bias under TRAD that was always greater than the bias under RR for the parameters in μ_1. This served the purpose of illustrating a scenario where TRAD was worse than both RR and IPTW+RR; however, there are scenarios in which RR may incur only one type of bias (that due to time-varying confounding), whereas TRAD may incur multiple types of bias (that due to time-varying confounding plus the other two problems with the traditional regression estimator discussed in the Introduction). In such cases, it is possible for these biases to have opposing signs and cancel each other out in such a way that TRAD may yield estimates closer to IPTW+RR than RR. We conjecture that this is what happened in the case study of the adolescent substance use data.

Our illustrative data analysis suggested interesting moderated effects of time-varying adolescent substance use treatment. For μ_1 and μ_3, we observed that the magnitude of treatment effect was larger for adolescents who were more severe prior to treatment. However, the direction of the effects were not consistent: The distal effects of treatment (μ_1) were not beneficial, whereas the proximal effects of additional treatment (μ_3) were beneficial. We offered a plausible scientific explanation for the iatrogenic effects of initial treatment, via association with more severe children. In the future work, it would be interesting to utilize and extend modern methods for causal mediation analysis [30] for unraveling the mechanisms by which initial treatment alone may have iatrogenic effects.

In the data analysis, it was also interesting to observe that the larger the t, the more potential there was for time-varying confounding. This was evident from the growing imbalance as measured by the maximum balance score for larger t (see largest value in left column of each of the panels in Figure 1). Although this may not be true of all applications, the pattern may not be surprising because as time progresses, many more covariates accumulate, leading to more chances for the analyst to see potential for confounding. Or this may be the result of stronger selection into treatment as time progresses: That is, over time, individuals may be more selective about staying in treatment, or, perhaps as treatment program managers get to know their clientele better, they may be more selective about who is encouraged to remain in treatment.

The methodology presented in this article does not explicitly aim to estimate the optimal dynamic treatment regime. However, we can use the methods presented as preliminary analyses for, or supplemented by, more sophisticated analyses or methodological development aimed explicitly at developing optimal dynamic treatment regimes [23, 27, 33, 50–53]. In particular, Henderson *et al.* [23, 24] use regret regression [50], which is an analogue of the RR estimator for estimating optimal dynamic treatment regimes. In future work, it is possible to extend regret regression with inverse weighting, analogous to how we extended RR with inverse weighting in this manuscript.

Appendix A. Data generating model for the simulation experiments

For the simulation experiments in Section 5, we generated data sets $\{U, S_0, X_0, A_1, S_1, X_1, A_2, Y\}$ of size n according to the following scheme ($\Lambda(v)$ that denotes the inverse-logit function $= \exp(v)/(1 + \exp(v))$):

$$
\begin{array}{ll}
U \sim N(0, \sigma), & \sigma = 0.1 \\
S_0 \sim N(\gamma_0, \sigma), & \gamma_0 = 0.4 \\
S_1(a_1)|S_0, U \sim N(F_1\gamma_1 + \gamma_U U, \sigma), & F_1 = (1, a_1, S_0, a_1 S_0), \\
 & \gamma_1 = (0.5, -0.5, 0.1, -0.1)^T, \\
 & \gamma_U = 0.1 \\
\mu_1(S_0, a_1) = \beta_{1,0}a_1 + \beta_{1,1}a_1 S_0, & \beta_{1,0} = -0.1, \beta_{1,1} = -0.1 \\
\mu_2\left(\bar{S}_1(a_1), \bar{a}_2\right) = \beta_{2,0}a_2 + \beta_{2,1}a_2 S_1(a_1), & \beta_{2,0} = -0.1, \beta_{2,1} = -0.1 \\
\epsilon_1(U, S_0) = \eta_1(S_0 - \gamma_0) + \gamma_U U, & \eta_1 = 0.15 \\
\epsilon_2(U, S_0, a_1, S_1(a_1)) = \eta_2(S_1 - F_1\gamma_1 - \gamma_U U), & \eta_2 = 0.30
\end{array}
$$

$$
\begin{array}{ll}
Y(a_1,a_2)|U,\bar{S}_1(a_1) \sim N(\beta_0+\epsilon_1+\mu_1+\epsilon_2+\mu_2,\sigma), & \beta_0=0.8 \\
X_1(a_1)|Y\ (\bar{a}_2) \sim N(\rho_1+\rho_2 a_1+\rho_3\textstyle\sum_{a_1,a_2\in(0,1)} Y(a_1,a_2),\sigma/2), & \rho_1=-0.65 \\
 & \rho_2=-0.25, \\
 & \rho_3=0.50 \\
X_0|Y\ (\bar{a}_2) \sim N(\rho_4\textstyle\sum_{a_1,a_2\in(0,1)} Y(a_1,a_2),\sigma), & \rho_4=0.25 \\
A_1|S_0,X_0 \sim \text{Bernoulli}\ (p=\Lambda(G_1\alpha_1)), & G_1=(1,S_0,X_0), \\
 & \alpha_1=(-1.3,1.5,2.5) \\
S_1=A_1S_1(1)+(1-A_1)S_1(0) & \\
A_2|S_1,X_1 \sim \text{Bernoulli}\ (p=\Lambda(G_2\alpha_2)), & G_2=(1,S_1,X_1), \\
 & \alpha_2=(-1.3,2.0,2.5) \\
Y=Y(A_1,A_2)=\textstyle\sum_{a_1,a_2\in(0,1)} I(A_1=a_1,A_2=a_2)Y(a_1,a_2) &
\end{array}
$$

We used the parameter values given previously for the first simulation in Section 5.1. For the second simulation in Section 5.2, we set $\alpha_1=\alpha_2=0$ (no confounding).

From the earlier equations, we derive the SNMM for the conditional mean of $Y(a_1,a_2)$ given $(S_0,S_1(a_1))$. In the earlier equations, Y is generated according to the following known SNMM for $Y(a_1,a_2)$ conditional on $(U,S_0,S_1(a_1))$

$$
\begin{aligned}
m_Y(U,S_0,S_1(a_1)) = E(Y(a_1,a_2)\mid U,S_0,S_1(a_1)) &= \beta_0 \\
&+\epsilon_1(U,S_0)+\mu_1(S_0,a_1)+\epsilon_2(U,S_0,a_1,S_1(a_1)) \\
&+\mu_2(\bar{S}_1(a_1),\bar{a}_2).
\end{aligned} \tag{A.1}
$$

By definition $\epsilon_1(U,S_0)=E(Y(0,0)\mid U,S_0)-E(Y(0,0)$ and it is generated using the linear form $\epsilon_1(U,S_0)=\eta_1(S_0-E(S_0))+\gamma_U(U-E(U))$ where $E(S_0)=\gamma_0$ and $E(U)=0$, whereas, by definition $\epsilon_2(U,S_0,a_1,S_1(a_1))=E(Y(a_1,0)\mid U,S_0,S_1(a_1))-E(Y(a_1,0)\mid U,S_0)$ and it is generated using $\epsilon_2(U,S_0,a_1,S_1(a_1))=S_1(a_1)-E(S_1(a_1)\mid U,S_0)$ where $\gamma_1F_1+\gamma_UU$ is the linear model for $E(S_1(a_1)\mid U,S_0)$.

In the simulation experiments, we are interested in estimating SNMMs for $Y(a_1,a_2)$ given $(S_0,S_1(a_1))$ (i.e., integrating over U). The baseline variable U is an unknown or unmeasured common cause of both $S_1(a_1)$ and Y (e.g., genetic make-up). It is used in the simulations to illustrate problems with the traditional regression estimator. Note that in this data generative model, U is neither a moderator (i.e., the μ_t's are independent of U) nor an unmeasured confounder (i.e., U is not used to generate A_t).

To obtain the SNMM of interest for $E(Y(a_1,a_2)\mid S_0,S_1(a_1))$, we integrate U out of $m_Y(U,S_0,S_1(a_1))$ by taking the following conditional expectation $E(m_Y(U,S_0,S_1(a_1))\mid S_0,S_1(a_1))$. Because only the ϵ_t's are a function of U, all we need is

$$
\begin{aligned}
&E\Big(\epsilon_1(U,S_0)+\epsilon_2(U,S_0,a_1,S_1(a_1))\Big|S_0,S_1(a_1)\Big) \\
&\quad=E\Big(\eta_1(S_0-E(S_0))+\gamma_UU\Big|S_0,S_1(a_1)\Big) \\
&\qquad+E\Big(\eta_2\big(S_1(a_1)-E\left(S_1(a_1)\mid U,S_0\right)\big)\Big|S_0,S_1(a_1)\Big) \\
&\quad=\eta_1(S_0-E(S_0))+\gamma_UE\Big(U\Big|S_0,S_1(a_1)\Big) \\
&\qquad+\eta_2\Big(S_1(a_1)-E\Big(E\left(S_1(a_1)\Big|U,S_0\right)\Big|S_0\Big)\Big), \\
&\quad=\eta_1(S_0-\gamma_0)+\gamma_UE\Big(U\Big|S_0,S_1(a_1)\Big) \\
&\qquad+\eta_2\Big(S_1(a_1)-\big(\gamma_1F_1+\gamma_UE\left(U\mid S_0\right)\big)\Big),
\end{aligned} \tag{A.2}
$$

which relies on knowing the conditional means $E(U\mid S_0)$ and $E(U\mid S_0,S_1(a_1))$. Because U is independent of S_0 and $E(U)=0$, then $E(U|S_0)=0$. By using standard normal theory, it can be shown that $E(U\mid S_0,S_1(a_1))=\frac{\gamma_U}{1+\gamma_U^2}(S_1(a_1)-\gamma_1F_1)$. It follows that

$$
E\Big(\epsilon_1(U,S_0)+\epsilon_2(U,S_0,a_1,S_1(a_1))\Big|S_0,S_1(a_1)\Big)=\eta_1(S_0-\gamma_0)+(\eta_2+\frac{\gamma_U^2}{1+\gamma_U^2})(S_1(a_1)-\gamma_1F_1).
$$

Therefore, we can use the following design matrix in an IPTW+RR fit of the SNMM for $E(Y(a_1,a_2) \mid S_0, S_1(a_1))$: $D_\gamma = (1, S_0 - \gamma_0, A_1 H_1, S_1(a_1) - \gamma_1 F_1, A_2 H_2)$ where $H_1 = (1, S_0)$ and $H_2 = (1, S_1)$.

Appendix B. Asymptotic standard errors of the IPTW+RR estimator

Asymptotic standard errors for the IPTW+RR estimates of θ of the SNMM should take into account sampling error in the estimation of (α_t, π_t) in the weights and the γ_t used in the residuals. The estimates $\widehat{\pi}_t$ are solutions for π_t to the b_t estimating equations $0 = \mathbb{P}_n \psi_{\pi_t} = \mathbb{P}_n \left(A_t - \Lambda\left(L_t \pi_t\right)\right) L_t^T$ where L_t is a $1 \times b_t$ model vector of the data $(\bar{S}_t, \bar{A}_{t-1})$ and $\Lambda(\cdot)$ is the inverse-logit function $\Lambda(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$; these are the logistic regressions to estimate the numerator probabilities. The estimates $\widehat{\alpha}_t$ are solutions for α_t to the c_t estimating equations $0 = \mathbb{P}_n \psi_{\alpha_t} = \mathbb{P}_n \left(A_t - \Lambda\left(G_t \alpha_t\right)\right) G_t^T$ where G_t is a $1 \times c_t$ model vector of the data $(\bar{V}_t, \bar{A}_{t-1})$; these are the logistic regressions to estimate the denominator probabilities (recall $V_t = (S_t, X_0)$). On the basis of standard Taylor series approximations, it follows that $\sqrt{n}\left(\hat{\alpha}_t - \alpha_t\right) = -\sqrt{n} J_{\alpha_t}^{-1} \mathbb{P}_n \psi_{\alpha_t} + o_P(1)$ and $\sqrt{n}\left(\hat{\pi}_t - \pi_t\right) = -\sqrt{n} J_{\pi_t}^{-1} \mathbb{P}_n \psi_{\pi_t} + o_P(1)$, where $J_{\alpha_t} = E \frac{\partial}{\partial \alpha_t} \psi_{\alpha_t}$ ($b_t \times b_t$ matrix) and $J_{\pi_t} = E \frac{\partial}{\partial \pi_t} \psi_{\pi_t}$ ($c_t \times c_t$ matrix). Next, the estimates $\widehat{\gamma}_t$ are solutions for γ_t to the k_t estimating equations $0 = \mathbb{P}_n \psi_{\gamma_t} = \mathbb{P}_n \tilde{W}_t \left(S_t - F_t \gamma_t\right) F_t^T$ where F_t is a $1 \times k_t$ model vector of the data $(\bar{S}_{t-1}, \bar{A}_t)$, and $\tilde{W}_t = \prod_{j=1}^t \hat{W}_j(\hat{\alpha}_j, \hat{\pi}_j)$ (see Step 1c in Section 4). On the basis of standard Taylor series approximations and taking into account the fact that the estimates $\widehat{\gamma}_t$ rely on the estimates of $\hat{\alpha}_j$ and $\hat{\pi}_j$ $(j = 1, \ldots, t)$, it follows that $\sqrt{n}\left(\hat{\gamma}_t - \gamma_t\right) = -\sqrt{n} J_{\gamma_t}^{-1} \mathbb{P}_n \left(\psi_{\gamma_t} - \sum_{j=1}^t J_{\gamma_t \alpha_j} J_{\alpha_j}^{-1} \psi_{\alpha_j} - \sum_{j=1}^t J_{\gamma_t \pi_j} J_{\pi_j}^{-1} \psi_{\pi_j}\right) + o_P(1)$, where $J_{\gamma_t} = E \frac{\partial}{\partial \gamma_t} \psi_{\gamma_t}(\bar{\alpha}_t, \bar{\pi}_t)$ ($k_t \times k_t$ matrix), $J_{\gamma_t \alpha_j} = E \frac{\partial}{\partial \alpha_j} \psi_{\gamma_t}(\bar{\alpha}_t, \bar{\pi}_t)$ ($k_t \times b_j$ matrix), and $J_{\gamma_t \pi_j} = E \frac{\partial}{\partial \pi_j} \psi_{\gamma_t}(\bar{\alpha}_t, \bar{\pi}_t)$ ($k_t \times c_j$ matrix) where $(\bar{\alpha}_t, \bar{\pi}_t) = (\alpha_1, \ldots, \alpha_t, \pi_1, \ldots, \pi_t)$. Finally, the estimates $\widehat{\theta}$ are solutions for θ to the $d = (1 + \sum r_t + \sum q_t)$ estimating equations $0 = \mathbb{P}_n \psi_\theta = \mathbb{P}_n \hat{W}(\hat{\alpha}, \hat{\pi}) \left(Y - D_{\hat{\gamma}} \theta\right) D_{\hat{\gamma}}^T$ where $\hat{W}(\hat{\alpha}, \hat{\pi}) = \prod_{t=1}^K \hat{W}_t(\hat{\alpha}_t, \hat{\pi}_t)$, $D_{\hat{\gamma}}$ is a $1 \times d$ model vector corresponding to the SNMM for the conditional mean of Y given $(\bar{V}_{K-1}, \bar{A}_K)$. For ease of notation in the next step, denote $\tilde{\psi}_{\gamma_t} = \psi_{\gamma_t} - \sum_{j=1}^t J_{\gamma_t \alpha_j} J_{\alpha_j}^{-1} \psi_{\alpha_j} - \sum_{j=1}^t J_{\gamma_t \pi_j} J_{\pi_j}^{-1} \psi_{\pi_j}$. On the basis of Taylor series approximations, it follows that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) = -\sqrt{n} J_\theta^{-1} \mathbb{P}_n \left(\psi_\theta - \sum_{t=1}^K J_{\theta\gamma_t} J_{\gamma_t}^{-1} \tilde{\psi}_{\gamma_t} - \sum_{t=1}^K J_{\theta\alpha_t} J_{\alpha_t}^{-1} \psi_{\alpha_t} - \sum_{t=1}^K J_{\theta\pi_t} J_{\pi_t}^{-1} \psi_{\pi_t}\right) + o_P(1),$$

where $J_\theta = E \frac{\partial}{\partial \theta} \psi_\theta$ ($d \times d$ matrix), $J_{\theta\gamma_t} = E \frac{\partial}{\partial \gamma_t} \psi_\theta$ ($d \times k_t$ matrix), $J_{\theta\alpha_t} = E \frac{\partial}{\partial \alpha_t} \psi_\theta$ ($d \times b_t$ matrix), and $J_{\theta\pi_t} = E \frac{\partial}{\partial \pi_t} \psi_\theta$ ($d \times c_t$ matrix). For simplicity, let $\tilde{\psi}_\theta = \psi_\theta - \sum_{t=1}^K J_{\theta\gamma_t} J_{\gamma_t}^{-1} \tilde{\psi}_{\gamma_t} - \sum_{t=1}^K J_{\theta\alpha_t} J_{\alpha_t}^{-1} \psi_{\alpha_t} - \sum_{t=1}^K J_{\theta\pi_t} J_{\pi_t}^{-1} \psi_{\pi_t}$. Because $E(\tilde{\psi}_\theta) = 0$, then by the Central Limit Theorem, $\sqrt{n}\left(\hat{\theta} - \theta\right) \rightsquigarrow N\left(0, J_\theta^{-1} E\left(\tilde{\psi}_\theta \tilde{\psi}_\theta^T\right) J_\theta^{-T}\right)$. Therefore, $\hat{\theta}$ is unbiased in large samples, with variance–covariance matrix $\Sigma_\theta = n^{-1} J_\theta^{-1} E\left(\tilde{\psi}_\theta \tilde{\psi}_\theta^T\right) J_\theta^{-T}$. To estimate Σ_θ, we use a 'plug-in' estimator where we replace all $(\theta, \alpha, \pi, \gamma)$'s in Σ_θ by $(\hat{\theta}, \hat{\alpha}, \hat{\pi}, \hat{\gamma})$, and we replace all the matrices in Σ_θ that are defined using expectations with their corresponding empirical means (e.g., replace J_θ with $\hat{J}_\theta = \mathbb{P}_n \frac{\partial}{\partial \theta} \psi_\theta$).

Acknowledgements

The development of this article was funded by the following grants: R01DA015697 (McCaffrey, Griffin, Ramchand), R01MH080015 (Murphy), and P50DA010075 (Murphy, Almirall), R03MH09795401 (Almirall), and RC4MH092722 (Almirall). It was also supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) contract #270-07-0191 using data provided by the following grantees: Cannabis Youth Treatment (Study: CYT; CSAT/SAMHSA contracts #270-97-7011, #270-00-6500, #270-2003-00006 and grantees: TI-11317, TI-11321, TI-11323,TI-11324), Adolescent Treatment Model (Study: ATM: CSAT/SAMHSA contracts #270-98-7047, #270-97-7011, #277-00-6500, #270-2003-00006 and grantees: TI-11894, TI-11892, TI-11422, TI-11423, TI-11424, TI-11432), the Strengthening Communities-Youth (Study: SCY; CSAT/SAMHSA contracts #277-00-6500, #270-2003-00006 and grantees: TI-13344, TI-13354, TI-13356), and Targeted Capacity Expansion (Study: TCE; CSAT/SAMHSA contracts #270-2003-00006, #270-2007-00004C, and #277-00-6500 and grantee TI-16400). The authors thank these grantees and their participants for agreeing to share their data to support the secondary data analysis to illustrate the methodology.

The opinions about this data are those of the authors and do not reflect official positions of the government or individual grantees. Please direct correspondence to Daniel Almirall, PhD, dalmiral@umich.edu, 734-936-3077.

References

1. Almirall D, McCaffrey D, Ramchand R, Murphy S. Subgroups analysis when treatment and moderators are time-varying. *Prevention Science* 2011; **14**(2):169–178.
2. Almirall D, Ten Have T, Murphy S. Structural nested mean models for assessing time-varying effect moderation. *Biometrics* 2009; **66**(1):131–139.
3. Petersen M, Deeks S, Martin J, van der Laan M. History-adjusted marginal structural models to estimate time-varying effect modification. *American Journal of Epidemiology* 2007; **166**(9):985–93.
4. Robins JM, Hernán MA, Rotnitzky A. Invited commentary on 'history-adjusted marginal structural models to estimate time-varying effect modification'. *American Journal of Epidemiology* 2007; **166**(9):994–1002.
5. Petersen M, van der Laan M. Response to invited commentary on 'history-adjusted marginal structural models to estimate time-varying effect modification'. *American Journal of Epidemiology* 2007; **166**(9):1003–1004.
6. Murphy SA, van der Laan MJ, Robins JM, CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* 2001; **96**:1410–1423.
7. Murphy S, Almirall D. Dynamic treatment regimens. In *Encyclopedia of Medical Decision Making*, Kattan MW (ed.). Sage Publications: Thousand Oaks, CA, 2009; 419–422.
8. Baron R, Kenny D. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 1986; **51**:1173–1182.
9. Kraemer HC, Wilson G, Fairburn C. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry* 2002; **59**:877–883.
10. Robins JM. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics. Springer-Verlag: New York, 1997; 69–117.
11. Robins JM. Association, causation, and marginal structural models. *Synthese* 1999; **121**:151–179.
12. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
13. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561–570.
14. Cole SR, Hernán MA, Robins JM, Anastos K, Chmiel J, Detels R, Ervin C, Feldman J, Greenblatt R, Kingsley L, Lai S, Young M, Cohen M, Muñoz A. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology* 2003; **158**(7):687–694.
15. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease* 1987; **40**(Supplement 2):139s–161s.
16. Robins JM. The control of confounding by intermediate variables. *Statistics in Medicine* 1989; **8**:679–701.
17. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics, Theory and Methods* 1994; **23**:2379–2412.
18. Robins JM. Estimating causal effects of time-varying endogenous treatments by g-estimation of structural nested models. In *Latent Variable Modeling and Applications to Causality*, Berkane M (ed.), Lecture Notes in Statistics. Springer: New York, 1997; 69–117.
19. Pearl J. Graphs, causality, and structural equation models. *Sociological Methods and Research* 1998; **27**:226–284.
20. Cole S, Platt R, Schisterman E, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* 2010; **39**:417–420.
21. Murphy S. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 2005; **24**(10):1455–1481.
22. Almirall D, Coffman C, Yancy W, Murphy S. Maximum likelihood estimation of the structural nested mean model using SAS PROC NLP. In *Analysis of Observational Health-Care Data Using SAS*, Faries D, Leon A, Haro J, Obenchain B (eds). SAS Press: Cary, NC, 2010; 231–261.
23. Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. *Biometrics* 2010; **66**(4):1192–201.
24. Henderson R, Ansell P, Alshibani D. Optimal dynamic treatment methods. *REVSTAT Statistical Journal* 2011; **9**(1):19–36.
25. van der Laan MJ, Murphy SA, Robins JM. Analyzing dynamic regimes using structural nested mean models, 2002. Unpublished Manuscript.
26. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*, Series in Statistics. Springer-Verlag: New York, NY, 2003.
27. Robins JM. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, Lin D, Heagerty P (eds). Springer: New York, NY, 2004; 189–326.
28. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**(5):688–701.
29. Holland P. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–970.
30. VanderWeele TJ. Mediation and mechanism. *European Journal of Epidemiology* 2009; **24**(5):217–224.
31. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface* 2009; **2**:457–468.
32. McCullagh P, Nelder J. *Generalized Linear Models*, 2nd edn. Chapman and Hall/CRC: Boca Raton, FL, 1989.
33. Rosthøj S, Keiding N, Schmiegelow K. Estimation of dynamic treatment strategies for maintenance therapy of children with acute lymphoblastic leukaemia: an application of history-adjusted marginal structural models. *Statistics in Medicine* 2009; **31**(5):470–488.

34. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Earlbaum Associates: Hillsdale, New Jersey, 1988.
35. Dennis ML, Titus JC, White MK, Unsicker JI, Hodgkins D. Global Appraisal of Individual Needs (GAIN): Administration guide for the GAIN and related measures, Chestnut Health Systems, Bloomington, IL: Chestnut Health Systems, 2002. Available online at http://www.chestnut.org/li/gain.
36. Almirall D, McCaffrey DF, Griffin BA, Ramchand R, Yuen RA, Murphy SA. Examining moderated effects of additional adolescent substance use treatment: Structural nested mean model estimation using inverse-weighted regression with residuals. *Technical Report 12-121*, Penn State University, University Park, PA, 2012.
37. Raghunathan T, Lepkowski J, Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 2001; **27**:85–95.
38. Raghunathan T, Solenberger P, Hoewyk J. IVEware: Imputation and variance estimation software, Survey Research Center, Institute for Social Research, Ann Arbor, MI, 2002.
39. Rubin D. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, NY, 1987.
40. Schafer J. *Analysis of Incomplete Multivariate Data*. Chapman and Hall / CRC Press: London, 1997.
41. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2004; **9**(4):403–425.
42. Dishion T, McCord J, Poulin F. When interventions harm: peer groups and problem behavior. *American Psychologist* 1999; **54**(9):755–764.
43. Garnick D, Lee M, O'Brien P, Panas L, Ritter G, Acevedo A, Garner B, Funk R, Godley M. The washington circle engagement performance measures' association with adolescent treatment outcomes. *Drug and Alcohol Dependence* 2012; **124**(3):250–258.
44. Williams R, Chang S. A comprehensive and comparative review of adolescent substance treatment outcome. *Clinical Psychology: Science and Practice* 2000; **7**(2):138–166.
45. Rothman K, Greenland S. *Modern Epidemiology*, 2nd edn. Lippincott-Raven: Philadelphia, PA, 1998.
46. Brumback B, Berg A. On effect-measure modification: relationships among changes in the relative risk, odds ratio, and risk difference. *Statistics in Medicine* 2008; **27**:3453–3465.
47. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society Series B* 2003; **65**:817–835.
48. Robins JM, Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* 2004; **91**:763–783.
49. Comté L, Vansteelandt S, Tousset E, Baxter G, Vrijens B. Linear and loglinear structural mean models to evaluate the benefits of an on-demand dosing regimen. *Clinical Trials* 2009; **6**:403–415.
50. Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B* 2003; **65**(2):331–366.
51. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic and Clinical Pharmacology and Toxicology* 2006; **98**(3):237–242.
52. Moodie E, Richardson T, Stephens D. Demystifying optimal dynamic treatment regimes. *Biometrics* 2007; **63**(2):447–455.
53. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: main content. *International Journal of Biostatistics* 2010; **6**(2). Article 8.